Exploiting Binary Floating-Point Representations for Constraint Filtering

Roberto Bagnara
BUGSENG srl and Dept. of Mathematics and Computer Science, University of Parma, Italy,
bagnara@cs.unipr.it, http://www.cs.unipr.it/~bagnara

Matthieu Carlier
INRIA Rennes Bretagne Atlantique, France

Roberta Gori
Dept. of Computer Science, University of Pisa, Italy,
gori@di.unipi.it, http://www.di.unipi.it/~gori

Arnaud Gotlieb
Certus Software V&V Center, SIMULA Research Laboratory, Norway,
arnaud@simula.no, http://simula.no/people/arnaud

Floating-point computations are quickly finding their way in the design of safety- and mission-critical systems, despite the fact that designing floating-point algorithms is significantly more difficult than designing integer algorithms. For this reason, verification and validation of floating-point computations is a hot research topic. An important verification technique, especially in some industrial sectors, is testing. However, generating test data for floating-point intensive programs proved to be a challenging problem. Existing approaches usually resort to random or search-based test data generation, but without symbolic reasoning it is almost impossible to generate test inputs that execute complex paths controlled by floating-point computations. Moreover, as constraint solvers over the reals or the rationals do not handle the rounding errors, the need arises for efficient constraint solvers over floating-point domains. In this paper, we present and fully justify improved algorithms for the filtering of arithmetic IEEE 754 binary floating-point constraints. The key point of these algorithms is a generalization of an idea by B. Marre and C. Michel that exploits a property of the representation of floating-point numbers.

Key words: software verification; testing; floating-point numbers; constraint solving

1. Introduction

During the last decade, the use of floating-point computations in the design of critical systems has become increasingly acceptable. Even in the civil and military avionics domain, which are among the most critical domains for software, floating-point numbers are now seen as a sufficiently-safe, faster and cheaper alternative to fixed-point arithmetic. To the point that, in modern avionics, floating-point is the norm rather than the exception (**?**).

Acceptance of floating-point computations in the design of critical systems took a long time. In fact, rounding errors can cause subtle bugs which are often missed by non experts (**?**), and

can lead to catastrophic failures. For instance, during the first Persian Gulf War, the failure of a Patriot missile battery in Dhahran was traced to an accumulating rounding error in the continuous execution of tracking and guidance software: this failure prevented the interception of an Iraqi Scud that hit the barracks in Dhahran, Saudi Arabia, killing 28 US soldiers (**?**). A careful analysis of this failure revealed that, even though the rounding error obtained at each step of the floating-point computation was very small, the propagation during a long loop-iterating path could lead to dramatic imprecision.

Adoption of floating-point computations in critical systems involves the use of thorough unit testing procedures that are able to exercise complex chains of floating-point operations. In particular, a popular practice among software engineers in charge of the testing of floating-point-intensive computations consists in executing carefully chosen loop-iterating paths in programs. They usually pay more attention to the paths that are most likely to expose the system to unstable numerical computations.[1] For critical systems, a complementary requirement is to demonstrate the infeasibility of selected paths, in order to convince a third-party certification authority that certain unsafe behaviors of the systems cannot be reached. As a consequence, software engineers face two difficult problems:

1. How to accurately predict the expected output of a given floating-point computation?[2]
2. How to find a test input that is able to exercise a given path, the execution of which depends on the results of floating-point computations, or to guarantee that such a path is infeasible?

The first problem has been well addressed in the literature (**?**) through several techniques. **?** report on a technique known as the *data diversity* approach, which uses multiple related program executions of a program to check their results. *Metamorphic testing* (**?**) generalizes this technique by using known numerical relations of the function implemented by a program to check the results of two or more executions. **?** proposes using the *abstract interpretation* framework (**?**) to estimate the deviation of the floating-point results with respect to an interpretation over the reals. **?** propose using a probabilistic approach to estimate round-off error propagation. More recently, **?** propose to exploit perturbation techniques to evaluate the stability of a numerical program. In addition to these approaches, it is of course possible to use a (partial) specification, a prototype or an old implementation in order to predict the results for a new implementation.

In contrast, the second problem received only little attention. Beyond the seminal work of **?**, who proposed to guide the search of floating-point inputs to execute a selected path, few approaches

[1] A computation can be called *numerically stable* if it can be proven not to magnify approximation errors. It can be called *(potentially) unstable* otherwise.

[2] This is the the well-known *oracle problem* (see **?**).

try to exactly reason over floating-point computations. The work of **?** paved the way to the development of *search-based test data generation* techniques, which consist in searching test inputs by minimizing a cost function, evaluating the distance between the currently executed path and a targeted selected path (**???**). Although these techniques enable quick and efficient coverage of testing criteria such as "all decisions," they are unfortunately sensitive to the rounding errors incurred in the computation of the branch distance (**?**). Moreover, search-based test data generation cannot be used to study path feasibility, i.e., to decide whether a possible execution path involving floating-point computations is feasible or not in the program. In addition, these techniques can be stuck in local minima without being able to provide a meaningful result (**?**). An approach to tackle these problems combines program execution and symbolic reasoning (**?**). This kind of reasoning requires solving constraints over floating-point numbers in order to generate test inputs that exercise a selected behavior of the program under test. However, solving floating-point constraints is hard and requires dedicated filtering algorithms (**??**). According to our knowledge, this approach is currently implemented in four solvers only: ECLAIR[3], FPCS (**?**), FPSE[4] (**?**), and Gatel, a test data generator for Lustre programs (**??**). It is worth noticing that existing constraint solvers dedicated to continuous domains (such as, e.g., RealPaver **?**, IBEX and Quimper **?** or ICOS **?**) handle correctly real or rational computations, but they cannot preserve the solutions of constraints over floating-point computations in all cases. Astonishing properties of floating-point computations such as absorption and cancellation **?** show that the rounding operations can severely compromise the preservation of the computation semantics between the reals and the floats. This statement is illustrated in Figure **??** and discussed in Section 5.

A promising approach to improve the filtering capabilities of constraints over floating-point variables consists in using some peculiar numerical properties of floating-point numbers. For linear constraints, this led to a relaxation technique where floating-point numbers and constraints are converted into constraints over the reals by using linear programming approaches (**?**). For interval-based consistency approaches, **?** identified a property of the representation of floating-point numbers and proposed to exploit it in filtering algorithms for addition and subtraction constraints. **?** proposed a reformulation of the Marre-Michel property in terms of filtering by maximum ULP (*Units in the Last Place*) that is generalizable to multiplication and division constraints.

? addressed the question of whether the Marre-Michel property can be useful for the automatic solution of realistic test input generation problems: they sketched (without proofs) a reformulation and correction of the filtering algorithm proposed in (**?**), along with a uniform framework that

[3] `http://bugseng.com/products/eclair`

[4] `http://www.irisa.fr/celtique/carlier/fpse.html`

generalizes the property identified by Marre and Michel to the case of multiplication and division. Most importantly, (**?**) presented the implementation of filtering by maximum ULP in FPSE and some of its critical design choices, and an experimental evaluation on constraint systems that have been extracted from programs engaging into intensive floating-point computations. These results show that the Marre-Michel property and its generalization defined in (**?**) speed up the test inputs generation process.

The present paper is, on the one hand, the theoretical counterpart of (**?**) in that all the results are thoroughly proved; on the other hand, this paper generalizes and extends (**?**) as far as the handling of subnormals and floating-point division are concerned. More precisely, the contributions of the paper are:

1. a uniform framework for filtering by maximum ULP is thoroughly defined and justified;
2. the framework is general enough to encompass all floating-point arithmetic operations and subnormals (the latter are not treated in (**?**));
3. a second indirect projection by maximum ULP for division (not present in any previous work);
4. all algorithms only use floating-point machine arithmetic operations on the same formats used by the analyzed computations.

The plan of the paper is as follows. Next section presents the IEEE 754 standard of binary floating-point numbers and introduces the notions and notations used throughout the paper. Section 3 recalls the basic principles of interval-based consistency techniques over floating-point variables and constraints. Section 4 presents our generalization of the Marre-Michel property along with a precise definition and motivation of all the required algorithms. Section 5 discusses related work. Section 6 concludes. The most technical proofs are available in the technical report version of the paper (**?**).

2. Preliminaries

In this section we recall some preliminary concepts and introduce the used notation.

2.1. IEEE 754

This section recalls the arithmetic model specified by the IEEE 754 standard for binary floating-point arithmetic (**?**). Note that, although the IEEE 754 standard also specifies formats and methods for decimal floating-point arithmetic, in this paper we only deal with binary floating-point arithmetic.

IEEE 754 binary floating-point formats are uniquely identified by quantities: $p \in \mathbb{N}$, the number of significant digits (precision); $e_{\max} \in \mathbb{N}$, the maximum exponent; $-e_{\min} \in \mathbb{N}$, the minimum exponent.[5]

[5] Note that, although the IEEE 754 formats have $e_{\min} = 1 - e_{\max}$, we never use this property and decided to keep the extra-generality, which might be useful to accommodate other formats.

The *single precision* format has $p = 24$ and $e_{\max} = 127$, the *double precision* format has $p = 53$ and $e_{\max} = 1023$ (IEEE 754 also defines extended precision formats). A finite, non-zero IEEE 754 floating-point number z has the form $(-1)^s b_1.m \times 2^e$ where s is the *sign bit*, b_1 is the *hidden bit*, m is the $(p-1)$-bit *significand* and the *exponent* e is also denoted by e_z or $\exp(z)$. Hence the number is positive when $s = 0$ and negative when $s = 1$. b_1 is termed "hidden bit" because in the *binary interchange format encodings* it is not explicitly represented, its value being encoded in the exponent (**?**).

Each format defines several classes of numbers: normal numbers, subnormal numbers, signed zeros, infinities and NaNs (*Not a Number*). The smallest positive *normal* floating-point number is $f^{\text{nor}}_{\min} = 1.0 \cdots 0 \times 2^{e_{\min}} = 2^{e_{\min}}$ and the largest is $f_{\max} = 1.1 \cdots 1 \times 2^{e_{\max}} = 2^{e_{\max}}(2 - 2^{1-p})$; normal numbers have the hidden bit $b_1 = 1$. The non-zero floating-point numbers whose absolute value is less than $2^{e_{\min}}$ are called *subnormals*: they always have exponent equal to $e_{\min}$ and fewer than p significant digits as their hidden bit is $b_1 = 0$. Every finite floating-point number is an integral multiple of the smallest subnormal $f_{\min} = 0.0 \cdots 01 \times 2^{e_{\min}} = 2^{e_{\min}+1-p}$. There are two infinities, denoted by $+\infty$ and $-\infty$, and two *signed zeros*, denoted by $+0$ and -0: they allow some algebraic properties to be maintained (**?**).[6] NaNs are used to represent the results of invalid computations such as a division of two infinities or a subtraction of infinities with the same sign: they allow the program execution to continue without being halted by an exception.

IEEE 754 defines five rounding directions: toward negative infinity (*down*), toward positive infinity (*up*), toward zero (*chop*) and toward the nearest representable value (*near*); the latter comes into two flavors that depend on different tie-break rules for numbers exactly halfway between two representable numbers: *tail-to-even* or *tail-to-away* in which values with even mantissa or values away from zero are preferred, respectively. This paper is only concerned with round-to-nearest, tail-to-even, which is, by far, the most widely used. The round-to-nearest, tail-to-even value of a real number x will be denoted by $[x]_{\text{n}}$.

The most important requirement of IEEE 754 arithmetic is the accuracy of floating-point computations: add, subtract, multiply, divide, square root, remainder, conversion and comparison operations must deliver to their destination the exact result rounded as per the rounding mode in effect and the format of the destination. It is said that these operations are "correctly rounded."

The accuracy requirement of IEEE 754 can still surprise the average programmer: for example the single precision, round-to-nearest addition of 999999995904 and 10000 (both numbers can be exactly represented) gives 999999995904, i.e., the second operand is absorbed. The maximum error committed by representing a real number with a floating-point number under some rounding mode can be expressed in terms of the function $\text{ulp}\colon \mathbb{R} \to \mathbb{R}$ (**?**). Its value on 1.0 is about 10^{-7} for the single precision format.

[6] Examples of such properties are $\sqrt{1/z} = 1/\sqrt{z}$ and $1/(1/x) = x$ for $x = \pm\infty$.

2.2. Notation

The set of real numbers is denoted by $\mathbb{R}$ while $\mathbb{F}_{p,e_{\max}}$ denotes a sub-set of binary floating-point numbers, defined from a given IEEE 754 format: this includes $-\infty, +\infty$ and zeros but neither subnormal numbers nor NaNs. Subnormals are introduced in the set $\mathbb{F}^{\text{sub}}_{p,e_{\max}} = \mathbb{F}_{p,e_{\max}} \cup \{\, (-1)^s 0.m \times 2^{e_{\min}} \mid s \in \{0,1\}, m \neq 0 \,\}$. In some cases, the exposition can be much simplified by allowing the $e_{\max}$ of $\mathbb{F}_{p,e_{\max}}$ to be ∞, i.e., by considering an idealized set of floats where the exponent is unbounded. Among the advantages is the fact that subnormals in $\mathbb{F}^{\text{sub}}_{p,e_{\max}}$ can be represented as normal floating-point numbers in $\mathbb{F}_{p,\infty}$. Given a set of floating-point numbers $\mathbb{F}$, $\mathbb{F}^+$ denotes the "non-negative" subset of $\mathbb{F}$, i.e., with $s = 0$.

For a non-zero floating-point number x, we will write $\text{even}(x)$ (resp., $\text{odd}(x)$) to signify that the least significant digit of x's mantissa is 0 (resp., 1).

When the format is clear from the context, a real decimal constant (such as 10^{12}) denotes the corresponding round-to-nearest, tail-to-even floating-point value (i.e., 999999995904 for 10^{12}).

Henceforth, for $x \in \mathbb{R}$, x^+ (resp., x^-) denotes the smallest (resp., greatest) floating-point number strictly greater (resp., smaller) than x with respect to the considered IEEE 754 format. Of course, we have $f^+_{\max} = +\infty$ and $(-f_{\max})^- = -\infty$.

Binary arithmetic operations over the floats will be denoted by $\oplus$, $\ominus$, $\otimes$ and $\oslash$, corresponding to $+$, $-$, $\cdot$ and $/$ over the reals, respectively. According to IEEE 754, they are defined, under round-to-nearest tail-to-even, by

$$x \oplus y = [x + y]_{\text{n}}, \qquad x \ominus y = [x - y]_{\text{n}},$$
$$x \otimes y = [x \cdot y]_{\text{n}}, \qquad x \oslash y = [x / y]_{\text{n}}.$$

As IEEE 754 floating-point numbers are closed under negation, we denote the negation of $x \in \mathbb{F}^{\text{sub}}_{p,e_{\max}}$ simply by $-x$. Note that negation is a bijection. The symbol $\odot$ denotes any of $\oplus$, $\ominus$, $\otimes$ or $\oslash$. A floating-point variable $\mathtt{x}$ is associated to an interval of possible floating-point values; we will write $\mathtt{x} \in [\underline{\mathtt{x}}, \overline{\mathtt{x}}]$, where $\underline{\mathtt{x}}$ and $\overline{\mathtt{x}}$ denote the smallest and greatest value of the interval, $\underline{\mathtt{x}} \leq \overline{\mathtt{x}}$ and either $\underline{\mathtt{x}} \neq +0$ or $\overline{\mathtt{x}} \neq -0$.

3. Background on Constraint Solving over Floating-Point Variables

In this section, we briefly recall the basic principles of interval-based consistency techniques over floating-point variables and constraints.

3.1. Interval-based Consistency on Arithmetic Constraints

Program analysis usually starts with the generation of an intermediate code representation in a form called *three-address code* (TAC). In this form, complex arithmetic expressions and assignments are decomposed into sequences of assignment instructions of the form

$$\mathtt{result} := \mathtt{operand}_1 \ \mathtt{operator} \ \mathtt{operand}_2.$$

$\mathtt{z} = \mathtt{x} \oplus \mathtt{y}$		$\mathtt{z} = \mathtt{x} \ominus \mathtt{y}$	
$\overline{\mathtt{z}} = \overline{\mathtt{x}} \oplus \overline{\mathtt{y}}$,	(direct)	$\overline{\mathtt{z}} = \overline{\mathtt{x}} \ominus \underline{\mathtt{y}}$,	(direct)
$\underline{\mathtt{z}} = \underline{\mathtt{x}} \oplus \underline{\mathtt{y}}$		$\underline{\mathtt{z}} = \underline{\mathtt{x}} \ominus \overline{\mathtt{y}}$	
$\overline{\mathtt{x}} = \mathrm{mid}(\overline{\mathtt{z}}, \overline{\mathtt{z}}^+) \ominus \underline{\mathtt{y}}$	(1^{st} indirect)	$\overline{\mathtt{x}} = \mathrm{mid}(\overline{\mathtt{z}}, \overline{\mathtt{z}}^+) \oplus \overline{\mathtt{y}}$	(1^{st} indirect)
$\underline{\mathtt{x}} = \mathrm{mid}(\underline{\mathtt{z}}, \underline{\mathtt{z}}^-) \ominus \overline{\mathtt{y}}$		$\underline{\mathtt{x}} = \mathrm{mid}(\underline{\mathtt{z}}, \underline{\mathtt{z}}^-) \oplus \underline{\mathtt{y}}$	
$\overline{\mathtt{y}} = \mathrm{mid}(\overline{\mathtt{z}}, \overline{\mathtt{z}}^+) \ominus \underline{\mathtt{x}}$	(2^{nd} indirect)	$\overline{\mathtt{y}} = \overline{\mathtt{x}} \ominus \mathrm{mid}(\underline{\mathtt{z}}, \underline{\mathtt{z}}^-)$	(2^{nd} indirect)
$\underline{\mathtt{y}} = \mathrm{mid}(\underline{\mathtt{z}}, \underline{\mathtt{z}}^-) \ominus \overline{\mathtt{x}}$		$\underline{\mathtt{y}} = \underline{\mathtt{x}} \ominus \mathrm{mid}(\overline{\mathtt{z}}, \overline{\mathtt{z}}^+)$	

Figure 1 Formulas for direct/indirect projections of addition/subtraction

A further refinement consists in the computation of the *static single assignment form* (SSA) whereby, labeling each assigned variable with a fresh name, assignments can be considered as if they were equality constraints. For example, the TAC form of the floating-point assignment $\mathtt{z} := \mathtt{z} * \mathtt{z} + \mathtt{z}$ is $\mathtt{t} := \mathtt{z} * \mathtt{z}$; $\mathtt{z} := \mathtt{t} + \mathtt{z}$, which in SSA form becomes $\mathtt{t}_1 := \mathtt{z}_1 * \mathtt{z}_1$; $\mathtt{z}_2 := \mathtt{t}_1 + \mathtt{z}_1$, which, in turn, can be regarded as the conjunction of the constraints $t_1 = z_1 \otimes z_1$ and $z_2 = t_1 \oplus z_1$.

In an interval-based consistency approach to constraint solving over the floats, constraints are used to iteratively narrow the intervals associated to each variable: this process is called *filtering*. A *projection* is a function that, given a constraint and the intervals associated to two of the variables occurring in it, computes a possibly refined interval for the third variable (the projection is said to be *over* the third variable). Taking $z_2 = t_1 \oplus z_1$ as an example, the projection over z_2 is called *direct projection* (it goes in the same sense of the TAC assignment it comes from), while the projections over t_1 and z_1 are called *indirect projections*.

Figure 1 gives non-optimal projections for addition and subtraction. For finite $x, y \in \mathbb{F}_{p,e_{\max}}$, $\mathrm{mid}(x, y)$ denotes the number that is exactly halfway between x and y; note that either $\mathrm{mid}(x, y) \in \mathbb{F}_{p,e_{\max}}$ or $\mathrm{mid}(x, y) \in \mathbb{F}_{p+1,e_{\max}}$. Non-optimal projections for multiplication and division can be found in (**??**). Optimal projections are known for monotonic functions over one argument (**?**), but they are generally not available for other functions. Note, however, that optimality is not required in an interval-based consistency approach to constraint solving, as filtering is just used to remove some, not necessarily all, inconsistent values.

3.2. The Marre-Michel Property

? published an idea to improve the filtering of the addition/subtraction projectors. This is based on a property of the distribution of floating-point numbers among the reals: the greater a float, the greater the distance between it and its immediate successor. More precisely, for a given float x with exponent e_x, if $x^+ - x = \Delta$, then for y of exponent $e_x + 1$ we have $y^+ - y = 2\Delta$.

PROPOSITION 1. (**?**, Proposition 1) *Let $z \in \mathbb{F}_{p,\infty}$ be such that $0 < z < +\infty$; let also*

$$z = 1.b_2 \cdots b_i \overbrace{0 \cdots 0}^{k} \times 2^{e_z}, \qquad \text{with } b_i = 1;$$

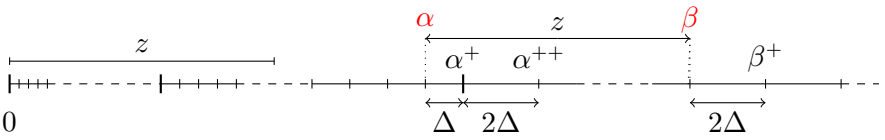

Figure 2 **An illustration of the Marre-Michel property: the segment z, if it has to represent the difference between two floats, cannot be moved past α**

$$\alpha = \overbrace{1.1\cdots1}^{p} \times 2^{e_z+k}, \qquad \textit{with } k = p - i;$$
$$\beta = \alpha \oplus z.$$

Then, for each $x, y \in \mathbb{F}_{p,\infty}$, $z = x \ominus y$ implies that $x \le \beta$ and $y \le \alpha$. Moreover, $\beta \ominus \alpha = \beta - \alpha = z$.

This property, which can be generalized to subnormals, can intuitively be explained on Figure 2 as follows. Let $z \in \mathbb{F}_{p,\infty}$ be a strictly positive constant such that $z = x \ominus y$, where $x, y \in \mathbb{F}_{p,\infty}$ are unknown. The Marre-Michel property says that y cannot be greater than α. In fact, α is carefully positioned so that $\alpha^{++} - \alpha^{+} = 2(\alpha^{+} - \alpha)$, $e_\alpha + 1 = e_\beta$ and $z = \beta - \alpha$; if we take $y = \alpha^{+}$ we need $x > \beta$ if we want $z = x - y$; however, the smallest element of $\mathbb{F}_{p,\infty}$ that is greater than β, β^{+}, is 2Δ away from β, i.e., too much. Going further with y does not help: if we take $y \ge \alpha^{+}$, then $y - \alpha$ is an odd multiple of Δ (one Δ step from α to α^{+}, all the subsequent steps being even multiples of Δ), whereas for each $x \ge \beta$, $x - \beta$ is an even multiple of Δ. Hence, if $y > \alpha$, $\big|z - (x - y)\big| \ge \Delta = 2^{e_z+1-i}$. However, since $k \ne p - 1$, $z^{+} - z = z - z^{-} = 2^{e_z+1-p} \le \Delta$. The last inequality, which holds because $p \ge i$, implies $z \ne x \ominus y$. A similar reasoning allows one to see that x cannot be greater than β independently from the value of y. In order to improve the filtering of the addition/subtraction projectors, **?** presented an algorithm to maximize the values of α and β over an interval. That algorithm and the main ideas behind the work presented in (**?**) will be revisited, corrected and discussed in detail in Section 4.5.

4. Filtering by Maximum ULP

TODO: announce the section contents.

4.1. Motivating Example

Consider the IEEE 754 single-precision constraint $\mathtt{z} = \mathtt{x} \oplus \mathtt{y}$ with initial intervals $\mathtt{z} \in [-\infty, +\infty]$, $\mathtt{x} \in [-1.0 \times 2^{50}, 1.0 \times 2^{50}]$ and $\mathtt{y} \in [-1.0 \times 2^{30}, 1.0 \times 2^{30}]$. Forward projection gives

$$\mathtt{z} \in \big[-1.\overbrace{0\cdots0}^{19}1 \times 2^{50}, 1.\overbrace{0\cdots0}^{19}1 \times 2^{50}\big],$$

which is optimal, as both bounds are attainable. Suppose now the interval for $\mathtt{z}$ is further restricted to $\mathtt{z} \in [1.0, 2.0]$ due to, say, a constraint from an if-then-else in the program or another indirect projection.

With the classical indirect projection we obtain $\mathtt{x}, \mathtt{y} \in [-1.0 \times 2^{30}, 1.0 \times 2^{30}]$, which, however, is not optimal. For example, pick $x = 1.0 \times 2^{30}$: for $y = -1.0 \times 2^{30}$ we have $x \oplus y = 0$ and $x \oplus y^+ = 64$. By monotonicity of $\oplus$, for no $y \in [-1.0 \times 2^{30}, 1.0 \times 2^{30}]$ we can have $x \oplus y \in [1.0, 2.0]$.

With our indirect projection, fully explained later, we obtain, from $\mathtt{z} \in [1.0, 2.0]$, the much tighter intervals $\mathtt{x}, \mathtt{y} \in [-1.1\cdots1 \times 2^{24}, 1.0 \times 2^{25}]$. These are actually optimal as $-1.1\cdots1 \times 2^{24} \oplus 1.0 \times 2^{25} = 1.0 \times 2^{25} \oplus -1.1\cdots1 \times 2^{24} = 2.0$. This example shows that filtering by maximum ULP can be stronger than classical interval-consistency based filtering. However, the opposite phenomenon is also possible. Consider again $\mathtt{z} = \mathtt{x} \oplus \mathtt{y}$ with $\mathtt{z} \in [1.0, 2.0]$. Suppose now the constraints for $\mathtt{x}$ and $\mathtt{y}$ are $\mathtt{x} \in [1.0, 5.0]$ and $y \in [-f_{\max}, f_{\max}]$. As we have seen, our indirect projection gives $\mathtt{y} \in [-1.1\cdots1 \times 2^{24}, 1.0 \times 2^{25}]$; in contrast, the classical indirect projection exploits the available information on $\mathtt{x}$ to obtain $\mathtt{y} \in [-4, 1]$. Indeed, classical and maximum ULP filtering for addition and subtraction are orthogonal: both should be applied in order to obtain precise results.

For an example on multiplication, consider the IEEE 754 single-precision constraint $\mathtt{z} = \mathtt{x} \otimes \mathtt{y}$ with initial intervals $\mathtt{z} \in [1.0 \times 2^{-50}, 1.0 \times 2^{-30}]$ and $\mathtt{x}, \mathtt{y} \in [-\infty, +\infty]$. In this case, classical projections do not allow pruning the intervals. However, take $x = 1.1 \times 2^{119}$: for $y = 0$ we have $x \otimes y = 0$ and $x \otimes y^+ = 1.1 \times 2^{-30}$. By monotonicity of $\otimes$, for no $y \in [-\infty, +\infty]$ we can have $x \otimes y \in [1.0 \times 2^{-50}, 1.0 \times 2^{-30}]$.

On the same example, $\mathtt{x}, \mathtt{y} \in [-1.0\cdots0 \times 2^{119}, 1.0\cdots0 \times 2^{119}]$ are the refinements given by our indirect projection. These are clearly optimal, as $1.0 \times 2^{-30} = -1.0\cdots0 \times 2^{119} \otimes -1.0\cdots0 \times 2^{149} = 1.0\cdots0 \times 2^{119} \otimes 1.0\cdots0 \times 2^{149}$. As is the case for addition, classical indirect projection can be more precise. Consider again $\mathtt{z} = \mathtt{x} \otimes \mathtt{y}$ with $\mathtt{z} \in [1.0 \times 2^{-50}, 1.0 \times 2^{-30}]$, $\mathtt{x} \in [2.0, 4.0]$ and $\mathtt{y} \in [-f_{\max}, f_{\max}]$. Classical indirect projection infers $\mathtt{y} \in [1.0 \times 2^{-52}, 1.0 \times 2^{-31}]$. by exploiting the information on $\mathtt{x}$.

4.2. Round-To-Nearest Tail-To-Even

We now formally define the round-to-nearest, tail-to-even rounding mode. To do that, we first introduce two functions: Δ_z^+ and Δ_z^- give the distance between z^+ and z and the distance between z and z^-.

Definition 1. The partial functions $\Delta^- \colon \mathbb{F}^{\text{sub}}_{p,e_{\max}} \rightarrowtail \mathbb{R}$ and $\Delta^+ \colon \mathbb{F}^{\text{sub}}_{p,e_{\max}} \rightarrowtail \mathbb{R}$ are defined as follows, for each finite $z \in \mathbb{F}^{\text{sub}}_{p,e_{\max}}$:

$$\Delta_z^+ = \begin{cases} 2^{1-p+e_{\max}}, & \text{if } z = f_{\max}; \\ f_{\min}, & \text{if } z = +0 \text{ or } z = -0; \\ z^+ - z, & \text{otherwise;} \end{cases}$$

$$\Delta_z^- = \begin{cases} 2^{1-p+e_{\max}} & \text{if } z = -f_{\max}; \\ f_{\min}, & \text{if } z = +0 \text{ or } z = -0; \\ z - z^-, & \text{otherwise.} \end{cases}$$

Note the special cases when $z = \pm 0$: since both $+0$ and -0 represent the real number 0, the distance between $z^+ = f_{\min}$ and $z = \pm 0$ is $f_{\min}$. We can now define round-to-nearest tail-to-even.

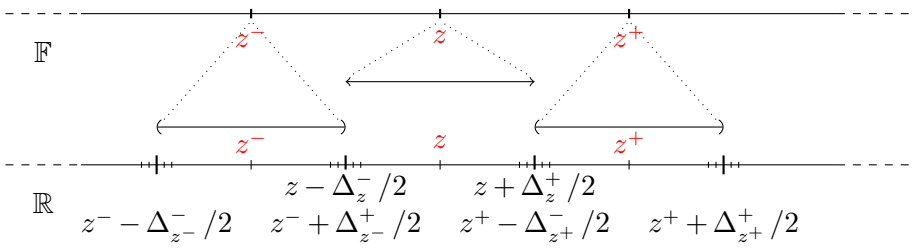

Figure 3 **Rounding of real numbers in the neighborhood of an even floating-point number z under round-to-nearest, tail-to-even**

DEFINITION 2. For $x \in \mathbb{R}$, $[x]_{\mathrm{n}}$ is defined as follows:

$$[x]_{\mathrm{n}} = \begin{cases} +0, & \text{if } 0 \le x \le \Delta_0^+ /2; \\ -0, & \text{if } -\Delta_0^- /2 \le x < 0; \\ z, & \text{if } z \in \mathbb{F}_{p,e_{\max}}^{\mathrm{sub}} \text{ and either even}(z) \text{ and} \\ & z - \Delta_z^- /2 \le x \le z + \Delta_z^+ /2 \text{ or odd}(z) \text{ and} \\ & z - \Delta_z^- /2 < x < z + \Delta_z^+ /2; \\ +\infty, & \text{if } x \ge f_{\max} + \Delta_{f_{\max}}^+ /2; \\ -\infty, & \text{if } x \le -f_{\max} - \Delta_{-f_{\max}}^- /2. \end{cases}$$

Figure 3 illustrates the round-to-nearest, tail-to-even rounding mode; if z is even, each real number between $z - \Delta_z^- /2$ and $z + \Delta_z^+ /2$, *including* extremes, is rounded to the same floating-point number z. As z is even, z^- is odd, and each real number between $z^- - \Delta_{z^-}^- /2$ and $z^- + \Delta_{z^-}^+ /2$, *excluding* extremes, is rounded to z^-. Similarly for z^+. Note that point $z - \Delta_z^- /2$ coincides with $z^- + \Delta_{z^-}^+ /2$ and $z + \Delta_z^+ /2$ coincides with $z^+ - \Delta_{z^+}^- /2$.

All rounding modes are monotonic; in particular, for each $x, y \in \mathbb{R}$, $x \le y$ implies $[x]_{\mathrm{n}} \le [y]_{\mathrm{n}}$. Moreover, the *chop* and *near* rounding modes are *symmetric*, i.e., the value after rounding does not depend on the sign: for each $x \in \mathbb{R}$, $[x]_{\mathrm{n}} = -[-x]_{\mathrm{n}}$.

TO BE REPHRASED This section reformulates the Marre-Michel property so as to generalize it to subnormals and to multiplication and division operators. The filtering algorithms that result from this generalization are collectively called *filtering by maximum ULP*.

4.3. Upper Bound

For each IEEE 754 floating-point operation $\odot \in \{\oplus, \ominus, \otimes, \oslash\}$, we will define the sets $\mathbb{F}_\odot \subseteq \mathbb{F}_{p,e_{\max}}$ and $\bar{\mathbb{F}}_\odot \subseteq \mathbb{F}_{p,\infty}$. Then we will define a function $\bar{\delta}_\odot \colon \mathbb{F}_\odot \to \bar{\mathbb{F}}_\odot$ (see Definition 3 for $\oplus$, Definition 5 for $\otimes$, Definition 6 for $\oslash$) that satisfies the following property, for each $z \in \mathbb{F}_\odot \setminus \{-0, +0, -\infty\}$:

$$\bar{\delta}_\odot(z) = \max\{v \in \bar{\mathbb{F}}_\odot \mid \exists y \in \bar{\mathbb{F}}_\odot \,.\, v \odot y = z\}. \tag{1}$$

It is worth noting that verifying that a function $\bar{\delta}_\odot$ satisfies (1) it's equivalent to prove that it satisfies the following properties, for each $z \in \mathbb{F}_\odot \setminus \{-0, +0, -\infty\}$:

$$\bar{\delta}_\odot(z) \in \{v \in \bar{\mathbb{F}}_\odot \mid \exists y \in \bar{\mathbb{F}}_\odot \,.\, v \odot y = z\}; \tag{2}$$

$$\forall z' \in \bar{\mathbb{F}}_\odot : z' > \bar{\delta}_\odot(z) \implies z' \notin \{v \in \bar{\mathbb{F}}_\odot | \exists y \in \bar{\mathbb{F}}_\odot \,.\, v \odot y = z\}. \tag{3}$$

In words, $\bar{\delta}_\odot(z)$ is the greatest float in $\bar{\mathbb{F}}_\odot$ that can be the left operand of $\odot$ to obtain z. Note that we may have $\bar{\mathbb{F}}_\odot \not\subseteq \mathbb{F}_{p,e_{\max}}$: property (1) refers to an idealized set of floating-point numbers with unbounded exponents.

Since we are interested in finding the upper bound of $\bar{\delta}_\odot(z)$ for $z \in [\underline{z}, \overline{z}]$, we need the following

PROPOSITION 2. *Let $w, v_1, \ldots, v_n \in \mathbb{F}_\odot \setminus \{-0, +0, -\infty\}$ be such that, for each $i = 1, \ldots, n$, $\bar{\delta}_\odot(w) \geq \bar{\delta}_\odot(v_i)$. Then, for each $w' \in \bar{\mathbb{F}}_\odot$ with $w' > \bar{\delta}_\odot(w)$, we have that $\forall z \in \mathbb{F}_\odot \setminus \{-0, +0, -\infty\}$, $w' \notin \{v \in \bar{\mathbb{F}}_\odot | \exists y \in \bar{\mathbb{F}}_\odot \,.\, v \odot y = z\}$.*

Proof. It follows directly from (1) .

Let $\mathtt{z} = \mathtt{x} \odot \mathtt{y}$ be a floating-point constraint where $-0, +0, -\infty \notin [\underline{z}, \overline{z}]$ and let $w \in [\underline{z}, \overline{z}]$ be such that $\bar{\delta}_\odot(w) \geq \bar{\delta}_\odot(v)$ for each $v \in [\underline{z}, \overline{z}]$: then no element of $\mathtt{x}$ that is greater than $\bar{\delta}_\odot(w)$ can participate to a solution of the constraint.

Dually, in order to refine the upper bound of $\mathtt{y}$ subject to $\mathtt{z} = \mathtt{x} \odot \mathtt{y}$, it is possible to define a function $\bar{\delta}'_\odot$ satisfying the following property, for each $z \in \mathbb{F}_\odot \setminus \{-0, +0, -\infty\}$:

$$\bar{\delta}'_\odot(z) = \max\{v \in \bar{\mathbb{F}}_\odot | \exists x \in \bar{\mathbb{F}}_\odot \,.\, x \odot v = z\}. \tag{4}$$

Based on property (4), we can state for $\bar{\delta}'_\odot$ an analogous result of the one of Proposition 2, allowing us to refine the interval for $\mathtt{y}$.

Note, though, that when $\odot$ is commutative (i.e., it is $\oplus$ or $\otimes$), $\bar{\delta}_\odot = \bar{\delta}'_\odot$.

4.4. Lower bound

For computing the lower bound, we will introduce functions $\underline{\delta}_\odot \colon \mathbb{F}_\odot \to \bar{\mathbb{F}}_\odot$ satisfying the following property, for each $z \in \mathbb{F}_\odot \setminus \{-0, +0, +\infty\}$:

$$\bar{\delta}_\odot(z) = \min\{v \in \bar{\mathbb{F}}_\odot | \exists y \in \bar{\mathbb{F}}_\odot \,.\, v \odot y = z\}. \tag{5}$$

This property entails a result similar to Proposition 2: given constraint $\mathtt{z} = \mathtt{x} \odot \mathtt{y}$ where $-0, +0, +\infty \notin [\underline{z}, \overline{z}]$ and $w \in [\underline{z}, \overline{z}]$ such that $\underline{\delta}_\odot(w) \leq \underline{\delta}_\odot(v)$ for each $v \in [\underline{z}, \overline{z}]$, the float $\underline{\delta}_\odot(w)$ is a possibly refined lower bound for $\mathtt{x}$.

In a dual way, in order to refine the lower bound of $\mathtt{y}$ subject to $\mathtt{z} = \mathtt{x} \odot \mathtt{y}$, we will define functions $\underline{\delta}'_\odot$ satisfying, for each $z \in \mathbb{F}_\odot \setminus \{-0, +0, +\infty\}$:

$$\bar{\delta}'_\odot(z) = \min\{v \in \bar{\mathbb{F}}_\odot | \exists x \in \bar{\mathbb{F}}_\odot \,.\, x \odot v = z\}. \tag{6}$$

Property (6) ensures that, under $\mathtt{z} = \mathtt{x} \odot \mathtt{y}$ where $-0, +0, +\infty \notin [\underline{\mathtt{z}}, \overline{\mathtt{z}}]$, if $w \in [\underline{\mathtt{z}}, \overline{\mathtt{z}}]$ is such that $\underline{\delta}'_{\odot}(w) \leq \underline{\delta}'_{\odot}(v)$ for each $v \in [\underline{\mathtt{z}}, \overline{\mathtt{z}}]$, then the float $\underline{\delta}'_{\odot}(w)$ is a possibly refined lower bound for $\mathtt{y}$.

Again, when $\odot$ is commutative $\underline{\delta}_{\odot} = \underline{\delta}'_{\odot}$.

4.5. Filtering by Maximum ULP on Addition/Subtraction

In this section we introduce the functions $\overline{\delta}_{\oplus}$, $\underline{\delta}_{\oplus}$, $\overline{\delta}_{\ominus}$, $\overline{\delta}'_{\ominus}$, $\underline{\delta}_{\ominus}$ and $\underline{\delta}'_{\ominus}$. Note that, since $\oplus$ is commutative, we have $\overline{\delta}'_{\oplus} = \overline{\delta}_{\oplus}$ and $\underline{\delta}'_{\oplus} = \underline{\delta}_{\oplus}$.

The first step consists in extending Proposition 1 in order to explicitly handle subnormal numbers. Such extension was already sketched by **?**: here we fully describe it and prove its correctness. Subnormals, which in $\mathbb{F}^{\mathrm{sub}}_{p,e_{\max}}$ are represented by numbers having the hidden bit $b_1 = 0$ and exponent $e_{\min}$, can be represented in $\mathbb{F}_{p,\infty}$ by numbers with $b_1 = 1$ and exponent strictly smaller than $e_{\min}$. Namely, the element of $\mathbb{F}^{\mathrm{sub}}_{p,e_{\max}}$

$$0.0 \cdots 01 b_{j+1} \cdots b_p \times 2^{e_{\min}}$$

can be represented in $\mathbb{F}_{p,\infty}$ by the (normal) float

$$1.b_{j+1} \cdots b_p \overbrace{0 \cdots 0}^{j-1} \times 2^{e_{\min}-(j-1)}.$$

Based on this observation we can state the following

PROPOSITION 3. *Let $z \in \mathbb{F}^{\mathrm{sub}}_{p,e_{\min}}$ be such that $0 < z < f^{\mathrm{nor}}_{\min}$; define also*

$$\begin{aligned} &z = 0.0 \cdots 01 b_{j+1} \cdots b_i \overbrace{0 \cdots 0}^{k} \times 2^{e_{\min}}, && \text{with } b_i = 1; \\ &\alpha = \overbrace{1.1 \cdots 1}^{p} \times 2^{e_{\min}+k}, && \text{with } k = p - i; \\ &\beta = \alpha \oplus z. \end{aligned}$$

Then, for each $x, y \in \mathbb{F}^{\mathrm{sub}}_{p,e_{\max}}$, $z = x \ominus y$ implies that $x \leq \beta$ and $y \leq \alpha$. Moreover, $\beta \ominus \alpha = \beta - \alpha = z$.

Proof. The subnormal z is represented in $\mathbb{F}_{p,\infty}$ by the normal float

$$\hat{z} = 1.b_{j+1} \cdots b_i \overbrace{0 \cdots 0}^{k} \overbrace{0 \cdots 0}^{j-1} \times 2^{e_{\min}-(j-1)} = 1.b_{j+1} \cdots b_i \overbrace{0 \cdots 0}^{k+j-1} \times 2^{e_{\min}-(j-1)}.$$

We can apply Proposition 1 to $\hat{z}$ and obtain $\alpha = 1.1 \cdots 1 \times 2^{e_{\min}-(j-1)+k+j-1} = 1.1 \cdots 1 \times 2^{e_{\min}+k}$. Moreover, Proposition 1 assures that

$$\beta = \alpha \oplus 1.b_{j+1} \cdots b_i \overbrace{0 \cdots 0}^{k+j-1} \times 2^{e_{\min}-(j-1)}$$

is such that, for each $x, y \in \mathbb{F}_{p,\infty}$, $z = x \ominus y$ implies $x \leq \beta$ and $y \leq \alpha$ and $\beta \ominus \alpha = \beta - \alpha = z$. Since each number in $\mathbb{F}^{\text{sub}}_{p,e_{\max}}$ has an equivalent representation in $\mathbb{F}_{p,\infty}$, we only need to prove that $\beta = \alpha \oplus z$, which clearly holds, since

$$\begin{aligned}\beta &= \alpha \oplus 1.b_{j+1} \cdots b_i \overbrace{0 \cdots 0}^{k+j-1} \times 2^{e_{\min}-(j-1)} \\ &= \alpha \oplus 0.0 \cdots 01 b_{j+1} \cdots b_i \underbrace{0 \cdots 0}_{k} \times 2^{e_{\min}} \\ &= \alpha \oplus z.\end{aligned}$$

□

Using Propositions 1 and 3, we formally define the function $\bar{\delta}_{\oplus}$ as follows.

Definition 3. Let $\mathbb{F}_{\oplus} = \mathbb{F}^{\text{sub}}_{p,e_{\max}}$, $\bar{\mathbb{F}}_{\oplus} = \mathbb{F}^{+}_{p,\infty}$, and $z \in \mathbb{F}_{\oplus}$ be such that $|z| = b_1.b_2 \cdots b_i 0 \cdots 0 \times 2^{e_z}$, with $b_i = 1$. Similarly to Propositions 1 and 3, let $k = p - i$, $\alpha = 1.1 \cdots 1 \times 2^{e_z + k}$ and $\beta = \alpha \oplus |z|$. Then $\bar{\delta}_{\oplus} \colon \mathbb{F}_{\oplus} \to \bar{\mathbb{F}}_{\oplus}$ is defined, for each $z \in \mathbb{F}_{\oplus}$, by

$$\bar{\delta}_{\oplus}(z) = \begin{cases} +\infty, & \text{if } z = -\infty \text{ or } z = +\infty; \\ \alpha, & \text{if } -\infty < z < 0; \\ +0, & \text{if } z = -0 \text{ or } z = +0; \\ \beta, & \text{if } 0 < z < +\infty. \end{cases}$$

Theorem 1. *$\bar{\delta}_{\oplus}$ is well-defined and satisfies* (2) *and* (3).

Proof. We first show that $\bar{\delta}_{\oplus}(z)$ is well-defined, i.e., that it is a total function from $\mathbb{F}^{\text{sub}}_{p,e_{\max}}$ to $\mathbb{F}^{+}_{p,\infty}$. To this aim note that α and β are always non-negative normal floating-point numbers belonging to $\mathbb{F}_{p,\infty}$, and that $\bar{\delta}_{\oplus}(z)$ is defined for each $z \in \mathbb{F}^{\text{sub}}_{p,e_{\max}}$. Secondly, let us consider the following cases:

$z = +\infty$: for each $y \neq -\infty$ we have $+\infty \oplus y = +\infty$; thus, as $\bar{\delta}_{\oplus}(z) = +\infty$, (2) holds and (3) vacuously holds.

$f^{\text{nor}}_{\min} \leq z < +\infty$: we can apply Proposition 1 to obtain $z = \beta \ominus \alpha$. Then note that $\beta \ominus \alpha = [\beta - \alpha]_{\text{n}} = [\beta + -\alpha]_{\text{n}} = \beta \oplus -\alpha$. Hence, $\beta \oplus -\alpha = z$. Thus, $\bar{\delta}_{\oplus}(z) \oplus -\alpha = \beta \oplus -\alpha = z$ and (2) is satisfied with $y = -\alpha$. For proving (3), first note that $\beta > -\alpha$ since $\beta > 0$ and $\alpha > 0$. Moreover, by Proposition 1, we know that there does not exist an $x \in \mathbb{F}_{p,\infty}$ with $x > \beta$ such that there exists $y \in \mathbb{F}_{p,\infty}$ that satisfies $x \ominus y = z$. Since $x \ominus y = x \oplus -y$ we can conclude that, for each $z' > \beta = \bar{\delta}_{\oplus}(z)$, it does not exist $y' \in \mathbb{F}_{p,\infty}$ such that $z' \oplus y' = z$. Hence also (3) holds.

$0 < z < f^{\text{nor}}_{\min}$: by applying Proposition 3 instead of Proposition 1 we can reason exactly as in the previous case.

$-\infty < z \leq -f^{\text{nor}}_{\min}$: since $0 < -z < +\infty$ we can apply Proposition 1 to $-z$ and obtain $\beta \ominus \alpha = -z$ and thus $-(\beta \ominus \alpha) = z$. As $[\cdot]_{\text{n}}$ is a symmetric rounding mode, we have $-(\beta \ominus \alpha) = -[\beta - \alpha]_{\text{n}} = [\alpha - \beta]_{\text{n}} = \alpha \oplus -\beta = z$. Thus, $\bar{\delta}_{\oplus}(z) \oplus -\beta = \alpha \oplus -\beta = z$ and (2) is satisfied with $y = -\beta$. For

proving (3), first note that $\alpha > -\beta$ since $\alpha > 0$ and $\beta > 0$. Moreover, by Proposition 1, we know that there does not exist an $y \in \mathbb{F}_{p,\infty}$ with $y > \alpha$ such that there exists $x \in \mathbb{F}_{p,\infty}$ that satisfies $x \ominus y = -z$. Since $x \ominus y = -z$ is equivalent to $y \oplus -x = z$, we can conclude that, for each $z' > \alpha = \bar{\delta}_{\oplus}(z)$, it does not exist $y' \in \mathbb{F}_{p,\infty}$ such that $z' \oplus y' = z$. Therefore, also in this case, (3) holds.

$-f^{\mathrm{nor}}_{\mathrm{min}} < z < 0$: by applying Proposition 3 instead of Proposition 1 we can reason exactly as in the previous case. □

As we have already observed, since $\oplus$ is commutative we have $\bar{\delta}'_{\oplus} = \bar{\delta}_{\oplus}$, that is, the same function $\bar{\delta}_{\oplus}$ is used to filter both x and y in the constraint $z = x \oplus y$.

The function $\underline{\delta}_{\oplus} \colon \mathbb{F}_{\oplus} \to \overline{\mathbb{F}}_{\oplus}$ is defined dually: for each $z \in \mathbb{F}_{\oplus} \setminus \{-0, +0, +\infty\}$, $\underline{\delta}_{\oplus}(z) = -\bar{\delta}_{\oplus}(-z)$. It is easy to see that properties (2) and (3) of $\bar{\delta}_{\oplus}$ entail property (5) of $\underline{\delta}_{\oplus}$. Again, since $\oplus$ is commutative, $\underline{\delta}'_{\oplus} = \underline{\delta}_{\oplus}$.

We now need algorithms to maximize $\bar{\delta}_{\oplus}$ and minimize $\underline{\delta}_{\oplus}$ over an interval of floating-point values. Since the two problems are dual to each other, we will focus on the maximization of $\bar{\delta}_{\oplus}$. As $\bar{\delta}_{\oplus}$ is not monotonic, a nontrivial analysis of its range over an interval is required. When the interval contains only finite, nonzero and positive (resp., negative) values, the range of $\bar{\delta}_{\oplus}$ has a simple shape. We are thus brought to consider an interval $[\underline{\mathtt{z}}, \overline{\mathtt{z}}]$ such that $\underline{\mathtt{z}} \notin \{-\infty, -0, +0\}$ and $\overline{\mathtt{z}} \notin \{-0, +0, +\infty\}$ and where $\underline{\mathtt{z}}$ and $\overline{\mathtt{z}}$ have the same sign. We will now revisit, correct and extend to subnormal floating-point numbers the algorithm originally proposed by **?** to maximize $\bar{\delta}_{\oplus}$ over $[\underline{\mathtt{z}}, \overline{\mathtt{z}}]$.

The idea presented in (**?**) is the following. When dealing with an interval $[\underline{\mathtt{z}}, \overline{\mathtt{z}}]$ with $\underline{\mathtt{z}} > 0$, α (and thus β and, therefore, our $\bar{\delta}_{\oplus}$) grows (i) with the exponent and (ii) with the number of successive 0 bits to the right of the mantissa, i.e., k in Propositions 1 and 3 and in Definition 3. Thus, maximizing these two criteria allows one to maximize α over the interval.

Definition 4. Let $\mathtt{z}$ be a variable over $\mathbb{F}^{\mathrm{sub}}_{p,e_{\max}}$. If we have $0 < \underline{\mathtt{z}} < \overline{\mathtt{z}} < +\infty$, then $\mu_{\oplus}(\mathtt{z}) \in [\underline{\mathtt{z}}, \overline{\mathtt{z}}]$ is given by:

1. $\mu_{\oplus}(\mathtt{z}) = 1.0 \cdots 0 \times 2^{e_{\overline{\mathtt{z}}}}$, if $e_{\underline{\mathtt{z}}} \neq e_{\overline{\mathtt{z}}}$;
2. $\mu_{\oplus}(\mathtt{z}) = b_1.b_2 \cdots b_{i-1} a 0 \cdots 0 \times 2^{e_{\overline{\mathtt{z}}}}$, if $e_{\underline{\mathtt{z}}} = e_{\overline{\mathtt{z}}}$, where, for some $b_i \neq b'_i$:

$$\underline{\mathtt{z}} = b_1.b_2 \cdots b_{i-1} b_i \cdots \times 2^{e_{\overline{\mathtt{z}}}};$$

$$\overline{\mathtt{z}} = b_1.b_2 \cdots b_{i-1} b'_i \cdots \times 2^{e_{\overline{\mathtt{z}}}};$$

$$a = \begin{cases} 0, & \text{if } b_1.b_2 \cdots b_{i-1} 0 \cdots 0 \times 2^{e_{\overline{\mathtt{z}}}} = \underline{\mathtt{z}}; \\ 1, & \text{otherwise.} \end{cases}$$

If $0 < \underline{\mathtt{z}} = \overline{\mathtt{z}} < +\infty$, then $\mu_{\oplus}(\mathtt{z}) = \underline{\mathtt{z}}$. If $-\infty < \underline{\mathtt{z}} \leq \overline{\mathtt{z}} < 0$, then $\mu_{\oplus}(\mathtt{z}) \in [\underline{\mathtt{z}}, \overline{\mathtt{z}}]$ is simply defined by $\mu_{\oplus}(\mathtt{z}) = -\mu_{\oplus}(\mathtt{w})$ where $\mathtt{w} \in [-\overline{\mathtt{z}}, -\underline{\mathtt{z}}]$. We leave $\mu_{\oplus}(\mathtt{z})$ undefined otherwise.

THEOREM 2. *Let* $\mathtt{z}$ *be over* $\mathbb{F}^{\mathrm{sub}}_{p,e_{\max}}$ *with* $\underline{\mathtt{z}} \notin \{-\infty,-0,+0\}$ *and* $\overline{\mathtt{z}} \notin \{-0,+0,+\infty\}$ *having the same sign. Then, for each* $z \in [\underline{\mathtt{z}},\overline{\mathtt{z}}]$, $\bar{\delta}_{\oplus}(z) \leq \bar{\delta}_{\oplus}\big(\mu_{\oplus}(\mathtt{z})\big)$.

Proof. Without loss of generality, assume $\underline{\mathtt{z}} > 0$. As the result is trivial if $\underline{\mathtt{z}} = \overline{\mathtt{z}}$, let us also assume $\underline{\mathtt{z}} < \overline{\mathtt{z}}$. We start proving that α and β of Definition 3 computed over $\mu_{\oplus}(\mathtt{z})$ are greater than or equal to the α's and β's computed over any other value in $[\underline{\mathtt{z}},\overline{\mathtt{z}}]$.

We first prove that $\mu_{\oplus}(\mathtt{z})$ maximizes α. For $z \in [\underline{\mathtt{z}},\overline{\mathtt{z}}]$ we have

$$\alpha = 1.1\cdots1 \times 2^{e_z+k},$$

where k is the number of successive 0's to the right of the mantissa of z. Let us consider the maximum exponent of the values in $\mathtt{z}$, which is $e_{\overline{\mathtt{z}}}$. Among the values in $[\underline{\mathtt{z}},\overline{\mathtt{z}}]$ with such an exponent, we want to select the one with the highest number of successive zeros to the right of the mantissa. Since $\underline{\mathtt{z}} > 0$, the maximum value for α would be attained by the float $1.0\cdots0 \times 2^{e_{\overline{\mathtt{z}}}}$, if this belongs to $[\underline{\mathtt{z}},\overline{\mathtt{z}}]$. This happens in three cases:

1. $e_{\underline{\mathtt{z}}} \neq e_{\overline{\mathtt{z}}}$ and $\mu_{\oplus}(\mathtt{z}) = 1.0\cdots0 \times 2^{e_{\overline{\mathtt{z}}}}$, by the first case of Definition 4.

2. $e_{\underline{\mathtt{z}}} = e_{\overline{\mathtt{z}}}$ and $\underline{\mathtt{z}} = 1.0\cdots0 \times 2^{e_{\overline{\mathtt{z}}}}$; in this case we have, again, $\mu_{\oplus}(\mathtt{z}) = 1.0\cdots0 \times 2^{e_{\overline{\mathtt{z}}}}$, so defined by the second case of Definition 4; in fact, for some $i \in \{2,\ldots,p-1\}$ that depends on $\overline{\mathtt{z}}$, we have

$$\overline{\mathtt{z}} = 1.b_2\cdots b_{i-1}10\cdots0 \times 2^{e_{\overline{\mathtt{z}}}},$$
$$\underline{\mathtt{z}} = 1.b_2\cdots b_{i-1}00\cdots0 \times 2^{e_{\overline{\mathtt{z}}}}$$

with $b_2 = \cdots = b_{i-1} = 0$, and the algorithm gives $1.b_2\cdots b_{i-1}a0\cdots0 \times 2^{e_{\overline{\mathtt{z}}}}$ with $a = 0$, i.e., $1.0\cdots0 \times 2^{e_{\overline{\mathtt{z}}}}$.

3. $e_{\underline{\mathtt{z}}} = e_{\overline{\mathtt{z}}}$, $\underline{\mathtt{z}} = 0.b_2\cdots b_p \times 2^{e_{\min}}$ and $\overline{\mathtt{z}} = 1.b'_2\cdots b'_p \times 2^{e_{\min}}$; thus we have, $\mu_{\oplus}(\mathtt{z}) = 1.0\cdots0 \times 2^{e_{\min}}$, once again by the second case of Definition 4 where $i = 1$, hence $\mu_{\oplus}(\mathtt{z}) = a.0\cdots0 \times 2^{e_{\min}}$. Moreover, since $\underline{\mathtt{z}} > 0$, necessarily $\underline{\mathtt{z}} \neq 0.0\cdots0 \times 2^{e_{\min}}$ and we must have $a = 1$.

We are now left with the case when $1.0\cdots0 \times 2^{e_{\overline{\mathtt{z}}}} \notin [\underline{\mathtt{z}},\overline{\mathtt{z}}]$. This occurs when $e_{\underline{\mathtt{z}}} = e_{\overline{\mathtt{z}}}$ but either $\underline{\mathtt{z}} > 1.0\cdots0 \times 2^{e_{\overline{\mathtt{z}}}}$ or $\overline{\mathtt{z}} < 1.0\cdots0 \times 2^{e_{\overline{\mathtt{z}}}}$. In both cases, all the floats in $[\underline{\mathtt{z}},\overline{\mathtt{z}}]$ have the same exponent and the same most significant bit (b_1). Therefore, in order to maximize α, we need to choose among them the one with the greatest number of successive zeros to the right of the mantissa. The first step is to find the index of the most significant mantissa bit where $\underline{\mathtt{z}}$ and $\overline{\mathtt{z}}$ differ: since $\underline{\mathtt{z}} < \overline{\mathtt{z}}$, such an index must exist. Let then

$$\underline{\mathtt{z}} = b_1.b_2\cdots b_{i-1}b_i\cdots \times 2^{e_{\overline{\mathtt{z}}}},$$
$$\overline{\mathtt{z}} = b_1.b_2\cdots b_{i-1}b'_i\cdots \times 2^{e_{\overline{\mathtt{z}}}},$$

where $b_i = 0$ and $b'_i = 1$ for some $i > 1$. The mantissa maximizing α is clearly $b_1.b_2 \cdots b_{i-1}0 \cdots 0$. Indeed, any float having a mantissa with a larger number of consecutive zeros to the right does not belong to $[\underline{\mathtt{z}}, \overline{\mathtt{z}}]$. However, it is not always the case that $b_1.b_2 \cdots b_{i-1}0 \cdots 0 \times 2^{e_{\overline{\mathtt{z}}}}$ belongs to $[\underline{\mathtt{z}}, \overline{\mathtt{z}}]$: we must have

$$\underline{\mathtt{z}} = b_1.b_2 \cdots b_{i-1}b_i0 \cdots 0 \times 2^{e_{\overline{\mathtt{z}}}}. \tag{7}$$

If (7) is true, then the second case of Definition 4 gives

$$\mu_{\oplus}(\mathtt{z}) = b_1.b_2 \cdots b_{i-1}a0 \cdots \times 2^{e_{\overline{\mathtt{z}}}}, \qquad \text{with } a = 0,$$

which is indeed equal to $\underline{\mathtt{z}}$. On the other hand, if (7) is false, then no float with mantissa $b_1.b_2 \cdots b_{i-1}00 \cdots 0$ belongs to $[\underline{\mathtt{z}}, \overline{\mathtt{z}}]$, hence the mantissa maximizing α is necessarily the one with one less zero to the right, i.e., $b_1.b_2 \cdots b_{i-1}10 \cdots 0$, which is guaranteed to belong to $[\underline{\mathtt{z}}, \overline{\mathtt{z}}]$. This is consistent with the second case of Definition 4, which gives

$$\mu_{\oplus}(\mathtt{z}) = b_1.b_2 \cdots b_{i-1}a0 \cdots 0 \times 2^{e_{\overline{\mathtt{z}}}}, \qquad \text{with } a = 1.$$

We have proved that Definition 4 gives a float $\mu_{\oplus}(\mathtt{z})$ that maximizes the value α. We now prove that $\mu_{\oplus}(\mathtt{z})$ also maximizes the value of β. By Propositions 1 and 3 and Definition 3, $\beta = \alpha \oplus z$. Note that $\mu_{\oplus}(\mathtt{z})$ maximizes α; however, since β also depends on z, we have to prove that no $z \in [\underline{\mathtt{z}}, \overline{\mathtt{z}}]$ such that $z > \mu_{\oplus}(\mathtt{z})$ results into a greater β. Observe first that, by construction, $\mu_{\oplus}(\mathtt{z})$ has the maximum exponent in $[\underline{\mathtt{z}}, \overline{\mathtt{z}}]$. Therefore any $z > \mu_{\oplus}(\mathtt{z})$ in $[\underline{\mathtt{z}}, \overline{\mathtt{z}}]$ must have a larger mantissa. Assume that $\mu_{\oplus}(\mathtt{z}) = b_1.b_2 \cdots b_j0 \cdots 0 \times 2^{e_{\overline{\mathtt{z}}}}$ with $b_j = 1$ for some $j \in \{1, \ldots, p\}$. The exponent of the corresponding α is $e_{\overline{\mathtt{z}}} + p - j$. Suppose now there exists $z > \mu_{\oplus}(\mathtt{z})$ in $[\underline{\mathtt{z}}, \overline{\mathtt{z}}]$ with a larger mantissa: this must have the form $b_1.b_2 \cdots b_\ell 0 \cdots 0 \times 2^{e_{\overline{\mathtt{z}}}}$ with $b_\ell = 1$ and $j < \ell \leq p$. The exponent of the corresponding α is $e_{\overline{\mathtt{z}}} + p - \ell$, which is smaller than the α computed for $\mu_{\oplus}(\mathtt{z})$ by at least one unit. Hence, we can conclude that $b_1.b_2 \cdots b_j0 \cdots 0 \times 2^{e_{\overline{\mathtt{z}}}} + 1.1 \cdots 1 \times 2^{e_{\overline{\mathtt{z}}}+p-j} > b_1.b_2 \cdots b_l0 \cdots 0 \times 2^{e_{\overline{\mathtt{z}}}} + 1.1 \cdots 1 \times 2^{e_{\overline{\mathtt{z}}}+p-\ell}$, since $\ell > j$. This shows that the float $\mu_{\oplus}(\mathtt{z})$ also maximizes the value of β. We have proved that Definition 4 gives a float $\mu_{\oplus}(\mathtt{z})$ that maximizes the value of both α and β over $\mathtt{z}$. Since Definition 3 defines $\bar{\delta}_{\oplus}(z) = \alpha$ for $-\infty < z < 0$ and $\bar{\delta}_{\oplus}(z) = \beta$ for $0 < z < +\infty$, we can conclude that, for each $z \in [\underline{\mathtt{z}}, \overline{\mathtt{z}}]$, $\bar{\delta}_{\oplus}(z) \leq \bar{\delta}_{\oplus}(\mu_{\oplus}(\mathtt{z}))$. □

As we have already pointed out, the algorithm of Definition 4, if restricted to normal numbers, is similar to the algorithm presented in (**?**). There is an important difference, though, in the case when $\underline{\mathtt{z}} = b_1.b_2 \cdots b_{i-1}b_i0 \cdots 0 \times 2^{e_{\overline{\mathtt{z}}}}$, $\overline{\mathtt{z}} = b_1.b_2 \cdots b_{i-1}b'_i \cdots \times 2^{e_{\overline{\mathtt{z}}}}$ and $\underline{\mathtt{z}} > 0$. In this case the algorithm of **?** returns $b_1.b_2 \cdots b_{i-1}10 \cdots 0 \times 2^{e_{\overline{\mathtt{z}}}}$. Note, however, that the value that maximizes α is $\underline{\mathtt{z}}$, which is different from $b_1.b_2 \cdots b_{i-1}10 \cdots 0 \times 2^{e_{\overline{\mathtt{z}}}}$.

Definition 4 cannot be extended to intervals containing zeros or infinities. Note that, for example, if $z = +0$ then no interesting bounds can be derived for x and y, since any value for x in the interval

$[-f_{\max}, +f_{\max}]$ would satisfy the constraint. Hence, when $\mathtt{z}$'s interval contains zeros or infinities, only the classical filtering (**??**) is applied.

For efficiency reasons, filtering by maximum ULP might be applied only when $\bar{\delta}_{\oplus}\big(\mu_{\oplus}(\mathtt{z})\big) \leq f_{\max}$ so as to avoid the use of wider floating-point formats.

In order to define $\bar{\delta}_{\ominus}$, $\bar{\delta}'_{\ominus}$, $\underline{\delta}_{\ominus}$ and $\underline{\delta}'_{\ominus}$, we can use the following observation. Since $x \ominus y = [x-y]_{\mathrm{n}} = [x+-y]_{\mathrm{n}} = x \oplus -y$, the constraints $\mathtt{z} = \mathtt{x} \ominus \mathtt{y}$ and $\mathtt{z} = \mathtt{x} \oplus -\mathtt{y}$ are equivalent. Thus we have $\bar{\delta}_{\ominus} = \bar{\delta}_{\oplus}$ and $\underline{\delta}_{\ominus} = \underline{\delta}_{\oplus}$, while $\bar{\delta}'_{\ominus} = -\underline{\delta}_{\oplus}$ and $\underline{\delta}'_{\ominus} = -\bar{\delta}_{\oplus}$ since, if $-y \in [\underline{\delta}_{\oplus}(z), \bar{\delta}_{\oplus}(z)]$, then $y \in [-\bar{\delta}_{\oplus}(z), -\underline{\delta}_{\oplus}(z)]$. Moreover, since $\mu_{\oplus}(\mathtt{z})$ maximizes $\bar{\delta}_{\oplus}$ and minimizes $\underline{\delta}_{\oplus}$ over an interval of floating-point values $\mathtt{z}$, $\mu_{\oplus}(\mathtt{z})$ can be used as well to maximize $\bar{\delta}'_{\ominus}$ and minimize $\underline{\delta}'_{\ominus}$ on $\mathtt{z}$.

4.6. Filtering by Maximum ULP on Multiplication

In order to be able to filter on multiplication, we need to determine the maximum and the minumum x satisfying $z = x \otimes y$ (see (1) and (5)). It is worth noting that, when dealing with multiplication (and similarly for division) we cannot use the same maximum ULP property upon which the treatment of addition and subtraction rests. This is because the ULP property of z is only loosely related to the ULP property of x and y when they are being multiplied. Indeed, in the idealized set of floating point numbers $\mathbb{F}_{p,\infty}$, the ULP distance of x is not strictly related to the ULP distance of z only, since the multiplication of x by y may increase or decrease the ULP distance of x. Since there isn't any minimum absolute value on $\mathbb{F}_{p,\infty}$, no maximum x satisfying $z = x \otimes y$ can be found. However, when dealing with the finite set of floating point numbers $\mathbb{F}_{p,e_{\max}}$, we can use the minimum absolute value of the set of floating point numbers different from zeros in order to determine the maximum x satisfying $z = x \otimes y$.

Consider a strictly positive constant $z \in \mathbb{F}_{p,e_{\max}}$ and two unknowns $x, y \in \mathbb{F}^{\mathrm{sub}}_{p,e_{\max}}$ such that $z = x \otimes y$. If $z \leq f_{\max}/f_{\min}$, there exists a greatest float $x_{\mathrm{m}} \in \mathbb{F}^{\mathrm{sub}}_{p,e_{\max}}$ such that there exists $y \in \mathbb{F}^{\mathrm{sub}}_{p,e_{\max}}$ satisfying $z = x_{\mathrm{m}} \otimes y$. More precisely, x_{m} must clearly satisfy $z = x_{\mathrm{m}} \otimes f_{\min}$ and it turns out that we can take $x_{\mathrm{m}} = z \oslash f_{\min}$. Since, for $z \leq f_{\max}/f_{\min}$, division of z by $f_{\min} = 2^{e_{\min}+1-p}$ amounts to an exponent shifting, we have that $\mathbb{F}^{\mathrm{sub}}_{p,e_{\max}} \ni x_{\mathrm{m}} = z/f_{\min}$. Moreover, we have that $x_{\mathrm{m}} = z/f_{\min}$ is the greatest float such that $z = x_{\mathrm{m}} \otimes f_{\min}$.[7]

On the other hand, there is no other float $y < f_{\min}$ such that $z = x \otimes y$, since y must be greater than $+0$, for otherwise $x \otimes y$ would not be strictly positive. However, for no $y \in \mathbb{F}^{\mathrm{sub}}_{p,e_{\max}}$ we have $+0 < y < f_{\min}$. Therefore, the greatest value x_{m} such that $z = x_{\mathrm{m}} \otimes f_{\min}$ is the greatest value for x that can satisfy $z = x \otimes y$ for some $y \in \mathbb{F}^{\mathrm{sub}}_{p,e_{\max}}$.

When dealing with subnormal floating-point numbers a similar argument applies. In fact, also in this case there exists a greatest float $x_{\mathrm{m}} \in \mathbb{F}^{\mathrm{sub}}_{p,e_{\max}}$ satisfying $z = x_{\mathrm{m}} \otimes y$ for some $y \in \mathbb{F}^{\mathrm{sub}}_{p,e_{\max}}$. As

[7] See the proof of forthcoming Theorem 3 available in the technical report version of the paper (**?**).

before, such x_{m} must satisfy $z = x_{\mathrm{m}} \otimes f_{\min}$. However, it turns out that, when z is subnormal, there may exist values for x_{m} greater than $z/f_{\min}$ that still satisfy $z = x_{\mathrm{m}} \otimes f_{\min}$. This is because the distance between subnormal numbers, being fixed to $f_{\min}$, does not depend on z.

Based on the previous reasoning, we can define $\bar{\delta}_{\otimes}$ and $\underline{\delta}_{\otimes}$.

Definition 5. Let $\mathbb{F}_{\otimes} = \left\{ z \in \mathbb{F}^{\mathrm{sub}}_{p,e_{\max}} \;\middle|\; |z|/f_{\min} \leq f_{\max} \right\}$ and $\bar{\mathbb{F}}_{\otimes} = \mathbb{F}_{p,e_{\max}}$. Then $\bar{\delta}_{\otimes} \colon \mathbb{F}_{\otimes} \to \bar{\mathbb{F}}_{\otimes}$ is defined, for each $z \in \mathbb{F}_{\otimes}$, by

$$\bar{\delta}_{\otimes}(z) = \begin{cases} |z| \oslash f_{\min}, & \text{if } |z| \geq f^{\mathrm{nor}}_{\min}; \\ \big(|z| \oslash f_{\min}\big) \oplus 2^{-1}, & \text{if } 0 < |z| < f^{\mathrm{nor}}_{\min} \text{ and } \mathrm{even}(z); \\ \Big(\big(|z| \oslash f_{\min}\big) \oplus 2^{-1}\Big)^{-}, & \text{if } 0 < |z| < f^{\mathrm{nor}}_{\min} \text{ and } \mathrm{odd}(z). \end{cases}$$

Theorem 3. *Function $\bar{\delta}_{\otimes}$ is well-defined and satisfies* (2) *and* (3).

Proof. Given in (**?**).

A monotonicity property of $\bar{\delta}_{\otimes}$ makes it easy to identify an element of the interval $\mathtt{z}$ that maximizes the value of $\bar{\delta}_{\otimes}$ over $\mathtt{z}$.

Proposition 4. *Let $z \in \mathbb{F}_{\otimes}$ be nonzero. If $z > 0$ then $\bar{\delta}_{\otimes}(z^{+}) \geq \bar{\delta}_{\otimes}(z)$; on the other hand, if $z < 0$ then $\bar{\delta}_{\otimes}(z^{-}) \geq \bar{\delta}_{\otimes}(z)$.*

Proof. Given in (**?**).

Since $\otimes$ is commutative, $\bar{\delta}'_{\otimes} = \bar{\delta}_{\otimes}$, and the same bounds can be used to filter both x and y in the constraint $z = x \otimes y$.

The function $\underline{\delta}_{\otimes} \colon \mathbb{F}_{\otimes} \to \bar{\mathbb{F}}_{\otimes}$ is defined dually: for each $z \in \mathbb{F}_{\otimes} \setminus \{-0, +0\}$, $\underline{\delta}_{\otimes}(z) = -\bar{\delta}_{\otimes}(z)$. It is easy to see that properties (2) and (3) of $\bar{\delta}_{\otimes}$ entail property (5) of $\underline{\delta}_{\otimes}$. Again, since $\otimes$ is commutative we have $\underline{\delta}'_{\otimes} = \underline{\delta}_{\otimes}$.

Thanks to Proposition 4 we know that the value $M \in [\underline{\mathtt{z}}, \overline{\mathtt{z}}]$ that maximizes $\bar{\delta}_{\otimes}$ is the one with the greatest absolute value, i.e., $M = \max\{|\underline{\mathtt{z}}|, |\overline{\mathtt{z}}|\}$. Since $\underline{\delta}_{\otimes}$ is defined as $-\bar{\delta}_{\otimes}(z)$, the value that minimizes $\underline{\delta}_{\otimes}$ is again M. Hence, if $[\underline{\mathtt{z}}, \overline{\mathtt{z}}]$ does not contain zeros, $\bar{\delta}_{\otimes}(M)$ (resp., $\underline{\delta}_{\otimes}(M)$) is an upper bound (resp., a lower bound) of $\mathtt{x}$ with respect to the constraint $\mathtt{z} = \mathtt{x} \otimes \mathtt{y}$.

The restriction to intervals $\mathtt{z}$ not containing zeros is justified by the fact that, e.g., if $z = 0$ then $z = x \otimes y$ holds with $x = f_{\max}$ and $y = 0$, hence, in this case, no useful filtering can be applied to x. The same thing of course happens when $\max\{|\underline{\mathtt{z}}|, |\overline{\mathtt{z}}|\}/f_{\min} > f_{\max}$. Moreover, whenever the interval of $\mathtt{y}$ does not contain zeros, filtering by maximum ULP for multiplication, in order to refine $\mathtt{x}$, is subsumed by the standard indirect projection. In contrast, when the interval of $\mathtt{y}$ does contain zeros our filter is able to derive bounds that cannot be obtained with the standard indirect projection, which, in this case, does not allow any refinement of the interval. Thus, for multiplication (and, as

we will see, for division as well), the standard indirect projection and filtering by maximum ULP are mutually exclusive: one applies when the other cannot derive anything useful.

Commenting a previous version of the present paper, Claude Michel observed that one could modify the standard indirect projections with interval splitting so that indirect projections are always applied to source intervals not containing zeros. This idea rests on the observation that, for $\mathtt{z} = \mathtt{x} \odot \mathtt{y}$ with $\odot \in \{\otimes, \oslash\}$, when the interval of $\mathtt{z}$ is a subset of the finite non zero floats neither $\mathtt{x}$ nor $\mathtt{y}$ do have any support for ± 0 and $\pm\infty$. For multiplication, ordinary standard indirect projection would be modified as follows, assuming that $\mathtt{z}$ is positive and we want to apply the standard indirect projection to $\mathtt{z}$ and $\mathtt{y}$ in order to refine $\mathtt{x}$ (the other cases being similar):

- we apply the ordinary standard indirect projection to $\mathtt{z}$ and $\mathtt{y} \cap [-f_{\max}, -f_{\min}]$, intersecting the resulting interval with $[-f_{\max}, -f_{\min}]$;
- we apply the ordinary standard indirect projection to $\mathtt{z}$ and $\mathtt{y} \cap [f_{\min}, f_{\max}]$, intersecting the resulting interval with $[f_{\min}, f_{\max}]$;
- finally, we use the convex union of the two intervals so computed to refine $\mathtt{x}$.

It can be shown that, when the applied ordinary (i.e., non-splitting) standard indirect projection is as precise as the one specified by **?**, the refining interval computed for $\mathtt{x}$ by that procedure either coincides with the result of the ordinary standard indirect projection (when filtering by maximum ULP is not applicable) or it coincides with the result of filtering by maximum ULP (when the ordinary standard indirect projection would not help). This approach has the advantage to be applicable to any rounding mode. On the other hand the standard indirect projections specified in (**?**) require working on rationals or on larger floating-point formats, whereas one of our aims is to always work with machine floating-point numbers of the same size of those used in the analyzed computation.

Example 1. Consider the IEEE 754 single-precision constraint $\mathtt{z} = \mathtt{x} \otimes \mathtt{y}$ with $\mathtt{z}$ subnormal: $\mathtt{z} \in [-0.00000000000000010001001 \times 2^{-126}, -0.00000000000010000000000 \times 2^{-126}]$ and $\mathtt{x}, \mathtt{y} \in [-\infty, +\infty]$. Our indirect projection gives the possible refinement $\mathtt{x}, \mathtt{y} \in [-1.00000000001 \times 2^{10}, 1.00000000001 \times 2^{10}]$, while classical inverse projections do not allow pruning the intervals for x and y.

4.7. Filtering by Maximum ULP on Division

We now define filtering by maximum ULP for floating-point constraints of the form $\mathtt{z} = \mathtt{x} \oslash \mathtt{y}$. We begin defining the first indirect projection. We will then tackle the problem of defining the second indirect projection, which, as we will see, is significantly more involved than the first one: the solution we propose is new to this paper.

4.7.1. The First Indirect Projection A role similar to the one of $f_{\min}$ in the definition of filtering by maximum ULP on multiplication is played by $f_{\max}$ in the definition of the first indirect projection for division.

DEFINITION 6. Let us define the sets $\mathbb{F}'_{\oslash} = \{\, z \in \mathbb{F}^{\mathrm{sub}}_{p,e_{\max}} \mid |z| \otimes f_{\max} \le f_{\max} \,\}$ and $\bar{\mathbb{F}}'_{\oslash} = \mathbb{F}_{p,e_{\max}}$. Let also $q = 1 - p + e_{\min} + e_{\max}$.[8] Then $\bar{\delta}_{\oslash} \colon \mathbb{F}'_{\oslash} \to \bar{\mathbb{F}}'_{\oslash}$ is defined, for each $z \in \mathbb{F}'_{\oslash}$, by

$$\bar{\delta}_{\oslash}(z) = \begin{cases} |z| \otimes f_{\max}, & \text{if } f^{\mathrm{nor}}_{\min} \le |z| \le 1; \\ \big(|z| \otimes f_{\max}\big) \oplus 2^q, & \text{if } 0 \le |z| < f^{\mathrm{nor}}_{\min} \\ & \quad \wedge \big(|z| \ne 1 \times 2^{e_z} \vee e_z = e_{\min} - 1\big); \\ \Big(\big(|z| \otimes f_{\max}\big) \oplus 2^q\Big)^-, & \text{otherwise.} \end{cases}$$

Observe that we have $|z| \otimes f_{\max} \le f_{\max}$ if and only if $|z| \le 1$. In fact, for $z = 1^+ = 1 + 2^{1-p}$, we obtain

$$\begin{aligned} |z| \otimes f_{\max} &= (1 + 2^{1-p}) \otimes f_{\max} \\ &= \big[(1 + 2^{1-p}) f_{\max}\big]_{\mathrm{n}} \\ &= \big[f_{\max} + (2 - 2^{1-p}) 2^{e_{\max}+1-p}\big]_{\mathrm{n}} \\ &= +\infty, \end{aligned} \tag{8}$$

where (8) holds by Definition 2, since $(2 - 2^{1-p}) 2^{e_{\max}+1-p} > \Delta^+_{f_{\max}}/2 = 2^{e_{\max}-p}$. By monotonicity of $\otimes$ we can conclude that $z \in \mathbb{F}'_{\oslash}$ if and only if $|z| \le 1$.

THEOREM 4. *$\bar{\delta}_{\oslash}$ is well-defined and satisfies* (2) *and* (3).

Proof. Given in (**?**).

The function $\underline{\delta}_{\oslash}$ is defined, for each $z \in \mathbb{F}'_{\oslash}$, by $\underline{\delta}_{\oslash} = -\bar{\delta}_{\oslash}(z)$.

A monotonicity property of $\bar{\delta}_{\oslash}$ makes it trivial to identify the value of $\mathtt{z}$ that maximizes the function.

PROPOSITION 5. *Let $z \in \mathbb{F}_{\oslash}$ be nonzero. If $z > 0$ then $\bar{\delta}_{\oslash}(z^+) \ge \bar{\delta}_{\oslash}(z)$; on the other hand, if $z < 0$ then $\bar{\delta}_{\oslash}(z^-) \ge \bar{\delta}_{\oslash}(z)$.*

Proof. Given in (**?**).

By monotonicity, the value $M \in [\underline{\mathtt{z}}, \overline{\mathtt{z}}]$ that maximizes $\bar{\delta}_{\oslash}$ is the one that has the greatest absolute value, i.e., $M = \max\{|\underline{\mathtt{z}}|, |\overline{\mathtt{z}}|\}$. Since $\underline{\delta}_{\oslash}$ is defined as $-\bar{\delta}_{\oslash}(z)$, M is also the value that minimizes $\underline{\delta}_{\oslash}$. Hence, if $[\underline{\mathtt{z}}, \overline{\mathtt{z}}]$ does not contain zeros, $\bar{\delta}_{\oslash}(M)$ (resp., $\underline{\delta}_{\oslash}(M)$) is an upper bound (resp. a lower bound) of $\mathtt{x}$ with respect to the constraint $\mathtt{z} = \mathtt{x} \oslash \mathtt{y}$. The restriction to intervals not containing zeros

[8] In the very common case where $e_{\min} = 1 - e_{\max}$ we have $q = 2 - p$.

is justified by the fact that, e.g., if $z = 0$ then $z = x \oslash y$ holds with $x = f_{\max}$ and $y = \infty$; hence, in this case, no useful filtering can be applied to x. The same thing happens when $\max\{|\underline{\mathtt{z}}|, |\overline{\mathtt{z}}|\} \otimes f_{\max} > f_{\max}$. In addition, whenever the interval of the variable $\mathtt{y}$ does not contain infinities, filtering by maximum ULP for division in order to refine $\mathtt{x}$ is subsumed by the standard indirect projection. On the other hand, when the interval of $\mathtt{y}$ does contain infinities, the standard indirect projection gives nothing whereas filtering by maximum ULP provides nontrivial bounds. Thus, as is the case for multiplication, the standard indirect projection and filtering by maximum ULP for division are mutually exclusive: one applies when the other cannot derive anything useful. And, just as for multiplication, if using rationals or extended floating-point formats is an option, then a pruning variant (one that cut off infinities) of the indirect projection specified in (**?**) will be equally precise.

Example 2. Consider the IEEE 754 single-precision constraint $\mathtt{z} = \mathtt{x} \oslash \mathtt{y}$ with initial intervals $\mathtt{z} \in [-1.0 \times 2^{-110}, -1.0 \times 2^{-121}]$ and $\mathtt{x}, \mathtt{y} \in [-\infty, +\infty]$. We have

$$\begin{aligned}
\overline{\delta}_{\oslash}(1.0 \times 2^{-110}) &= 1.0 \times 2^{-110} \cdot 1.1 \cdots 1 \times 2^{127} \\
&= 1.1 \cdots 1 \times 2^{17}, \\
\underline{\delta}_{\oslash}(1.0 \times 2^{-110}) &= -1.0 \times 2^{-110} \cdot 1.1 \cdots 1 \times 2^{127} \\
&= -1.1 \cdots 1 \times 2^{17}.
\end{aligned}$$

Filtering by maximum ULP improves upon classical filtering, which would not restrict any interval, with $\mathtt{x} \in [-1.1 \ldots 1 \times 2^{17}, 1.1 \ldots 1 \times 2^{17}]$.

For an example involving subnormals, consider $\mathtt{z} = \mathtt{x} \oslash \mathtt{y}$ with initial interval for $\mathtt{z}$ equal to $[0.00000000000000000000001 \times 2^{-126}, 0.01 \times 2^{-126}]$: our algorithm produces the possible refinement $\mathtt{x} \in [-1.00000000000000000000001 \times 2^{-46}, 1.00000000000000000000001 \times 2^{-46}]$. As before, if $\mathtt{y} \in [-\infty, +\infty]$ then classical filtering would not restrict the interval for x.

4.7.2. The Second Indirect Projection The discussion in Section 4.7.1 shows that, for $|z| \leq 1$, we have $\overline{\delta}'_{\oslash}(z) = f_{\max}$. We thus need to study $\overline{\delta}'_{\oslash}(z)$ for $|z| > 1$. It turns out that, due to rounding, the restriction of $\overline{\delta}'_{\oslash}$ over that subdomain is not a simple function. Given $z \in \mathbb{F}^{\mathrm{sub}}_{p,e_{\max}}$, $\overline{\delta}'_{\oslash}(z)$ is the maximum y such that $x \oslash y = z$. Note that, in order to maximize y, x must be maximized as well. A qualitative reasoning on the reals tells us that, since $f_{\max}/(f_{\max}/z) = z$, y should be roughly equal to $f_{\max}/|z|$. Indeed, it can be proved that, for $|z| > 1$, $f_{\max} \oslash \big(f_{\max} \oslash |z|\big)$ is equal to z, z^- or z^+ depending on the value of z. This allows the determination of a quite tight upper bound to the values that $\mathtt{z}$ may take, which is ultimately our goal for filtering $\mathtt{y}$ values. To this aim we define the function $\tilde{\delta}'_{\oslash}$.

DEFINITION 7. The function $\tilde{\delta}'_{\oslash}\colon \mathbb{F}^{\mathrm{sub}}_{p,e_{\max}} \to \mathbb{F}^{\mathrm{sub}+}_{p,e_{\max}}$ is defined, for each $z \in \mathbb{F}^{\mathrm{sub}}_{p,e_{\max}}$, as follows:

$$\tilde{\delta}'_{\oslash}(z) = \begin{cases} f_{\max} \oslash |z|^{--}, & \text{if } 1 < |z| \leq f_{\max}; \\ f_{\max}, & \text{otherwise.} \end{cases}$$

It turns out that $\tilde{\delta}'_{\oslash}(z)$ is, in general, an upper bound rather than the minumum as required by (4).

THEOREM 5. *Let* $\mathbb{F}''_{\oslash} = \mathbb{F}^{\mathrm{sub}}_{p,e_{\max}}$ *and* $\bar{\mathbb{F}}''_{\oslash} = \mathbb{F}_{p,e_{\max}}$. *Let* $\bar{\delta}'_{\oslash}\colon \mathbb{F}''_{\oslash} \to \bar{\mathbb{F}}''_{\oslash}$ *be a function satisfying* (4). *Then, for* $0 < |z| \leq 1$ *or* $z = +\infty$, *we have* $\bar{\delta}'_{\oslash}(z) \leq \tilde{\delta}'_{\oslash}(z)$; *moreover, for* $1 < |z| \leq f_{\max}$, $\bar{\delta}'_{\oslash}(z) < \tilde{\delta}'_{\oslash}(z)$.

Proof. Given in (**?**).

Dually, a lower bound for the function $\underline{\delta}'_{\oslash}$ can be obtained by means of the function $\underline{\tilde{\delta}}'_{\oslash}$, defined by $\underline{\tilde{\delta}}'_{\oslash}(z) = -\tilde{\delta}'_{\oslash}(z)$.

The value $N \in [\underline{\mathtt{z}}, \overline{\mathtt{z}}]$ that maximizes $\tilde{\delta}'_{\oslash}$ is the one that has the smallest absolute value, i.e., $N = \min\{|\underline{\mathtt{z}}|, |\overline{\mathtt{z}}|\}$. Since $\underline{\tilde{\delta}}'_{\oslash}$ is defined as $-\tilde{\delta}'_{\oslash}(z)$, N is also the value that minimizes $\underline{\tilde{\delta}}'_{\oslash}$. Thus, if $[\underline{\mathtt{z}}, \overline{\mathtt{z}}]$ does not contain zeros, $\tilde{\delta}'_{\oslash}(N)$ (resp., $\underline{\tilde{\delta}}'_{\oslash}(N)$) is an upper bound (resp. a lower bound) for $\mathtt{x}$ with respect to the constraint $\mathtt{z} = \mathtt{x} \oslash \mathtt{y}$. The restriction to intervals not containing zeros is justified by the fact that if, e.g., $z = 0$, then the equality $z = x \oslash y$ holds with $y = \infty$ for each x such that $0 \leq x \leq f_{\max}$. Hence, as in the case of the first projection, no useful filtering can be applied to $\mathtt{y}$. Analogously to the case of the filter for the first projection, this filter is useful whenever the interval of $\mathtt{x}$ contains infinities. In this case, in fact, it is able to derive useful bounds for $\mathtt{y}$ where the standard indirect projection does not allow any refinement of the interval. Just as is the case for multiplication and the first indirect projection of division, the standard indirect projection and filtering by maximum ULP are mutually exclusive: one applies when the other cannot derive anything useful.

Note that, only for this projection, we have chosen to compute a (very tight) upper bound that, in general, is not the least upper bound. We did so in order to trade precision for efficiency: this way we have an algorithm that only uses floating-point machine arithmetic operations on the same format used by the analyzed constraint $\mathtt{z} = \mathtt{x} \oslash \mathtt{y}$. When using rationals or larger floating-point formats is an option, a pruning variant (as in the previous case, one that cut off infinities) of a second indirect projection satisfying the precision constraints set forth in (**?**) may result in extra precision at a comparatively higher computational cost.

EXAMPLE 3. Consider the IEEE 754 single-precision division constraint $\mathtt{z} = \mathtt{x} \oslash \mathtt{y}$ with initial intervals $\mathtt{z} \in [1.0\cdots010 \times 2^{110}, 1.0 \times 2^{121}]$ and $\mathtt{x}, \mathtt{y} \in [-\infty, +\infty]$. We have

$$\begin{aligned} \tilde{\delta}'_{\oslash}(1.0\cdots01 \times 2^{110}) &= 1.1\cdots1 \times 2^{127} \oslash \left((1.0\cdots01 \times 2^{110})^{-}\right)^{-} \\ &= 1.1\cdots1 \times 2^{127} \oslash 1.1\cdots1 \times 2^{109} \end{aligned}$$

$$
\begin{aligned}
&= 1.0 \times 2^{18},\\
\underline{\tilde{\delta}}'_{\oslash}(1.0\cdots01 \times 2^{110}) &= -1.1\cdots1 \times 2^{127} \oslash \left((1.0\cdots01 \times 2^{110})^{-}\right)^{-}\\
&= -1.0 \times 2^{18}.
\end{aligned}
$$

Filtering by maximum ULP improves upon classical filtering, which gives nothing, with the refinement $\mathtt{y} \in [-1.0 \times 2^{18}, 1.0 \times 2^{18}]$.

4.8. Synthesis

Table 1 provides a compact presentation of filtering by maximum ULP.

Table 1 Filtering by maximum ULP synopsis

Constraint	$\mathtt{x} \subseteq \cdot$	$\mathtt{y} \subseteq \cdot$	Condition(s)
$\mathtt{z} = \mathtt{x} \oplus \mathtt{y},\ 0 < \mathtt{z} \le f_{\max}$	$[\ \underline{\delta}_{\oplus}(\zeta\ ),\ \ \overline{\delta}_{\oplus}(\zeta\ )]$	$[\ \underline{\delta}_{\oplus}(\zeta\ ),\ \ \overline{\delta}_{\oplus}(\zeta\ )]$	$\zeta = \mu_{\oplus}(\mathtt{z}),\ \ -f_{\max} \le \underline{\delta}_{\oplus}(\zeta),\ \ \overline{\delta}_{\oplus}(\zeta) \le f_{\max}$
$\mathtt{z} = \mathtt{x} \oplus \mathtt{y},\ -f_{\max} \le \mathtt{z} < 0$	$[-\overline{\delta}_{\oplus}(\zeta'), -\underline{\delta}_{\oplus}(\zeta')]$	$[-\overline{\delta}_{\oplus}(\zeta'), -\underline{\delta}_{\oplus}(\zeta')]$	$\zeta' = \mu_{\oplus}(-\mathtt{z}),\ \ -f_{\max} \le \underline{\delta}_{\oplus}(\zeta'),\ \ \overline{\delta}_{\oplus}(\zeta') \le f_{\max}$
$\mathtt{z} = \mathtt{x} \ominus \mathtt{y},\ 0 < \mathtt{z} \le f_{\max}$	$[\ \underline{\delta}_{\oplus}(\zeta\ ),\ \ \overline{\delta}_{\oplus}(\zeta\ )]$	$[-\overline{\delta}_{\oplus}(\zeta\ ), -\underline{\delta}_{\oplus}(\zeta\ )]$	$\zeta = \mu_{\oplus}(\mathtt{z}),\ \ -f_{\max} \le \underline{\delta}_{\oplus}(\zeta),\ \ \overline{\delta}_{\oplus}(\zeta) \le f_{\max}$
$\mathtt{z} = \mathtt{x} \ominus \mathtt{y},\ -f_{\max} \le \mathtt{z} < 0$	$[-\overline{\delta}_{\oplus}(\zeta'), -\underline{\delta}_{\oplus}(\zeta')]$	$[\ \underline{\delta}_{\oplus}(\zeta'),\ \ \overline{\delta}_{\oplus}(\zeta')]$	$\zeta' = \mu_{\oplus}(-\mathtt{z}),\ \ -f_{\max} \le \underline{\delta}_{\oplus}(\zeta'),\ \ \overline{\delta}_{\oplus}(\zeta') \le f_{\max}$
$\mathtt{z} = \mathtt{x} \otimes \mathtt{y},\ \lvert z\rvert \le 2^{2-p}(2-2^{1-p})$	$[\ \underline{\delta}_{\otimes}(m),\ \ \overline{\delta}_{\otimes}(m)]$	$[\ \underline{\delta}_{\otimes}(m),\ \ \overline{\delta}_{\otimes}(m)]$	$m = \max\{\lvert\underline{\mathtt{z}}\rvert, \lvert\overline{\mathtt{z}}\rvert\}$
$\mathtt{z} = \mathtt{x} \oslash \mathtt{y},\ 0 < \lvert\mathtt{z}\rvert \le 1$	$[\ \underline{\delta}_{\oslash}(m),\ \ \overline{\delta}_{\oslash}(m)]$		$m = \max\{\lvert\underline{\mathtt{z}}\rvert, \lvert\overline{\mathtt{z}}\rvert\}$
$\mathtt{z} = \mathtt{x} \oslash \mathtt{y},\ 1 < \lvert\mathtt{z}\rvert \le f_{\max}$		$[\ \underline{\tilde{\delta}}'_{\oslash}(n),\ \ \tilde{\delta}'_{\oslash}(n)]$	$n = \min\{\lvert\underline{\mathtt{z}}\rvert, \lvert\overline{\mathtt{z}}\rvert\}$

$$\overline{\delta}_{\oplus}(z) = \begin{cases} \beta, & \text{if } 0 < z < +\infty, \\ \alpha, & \text{if } -\infty < z < 0; \end{cases} \qquad \underline{\delta}_{\oplus}(z) = -\overline{\delta}_{\oplus}(-z);$$

$$\overline{\delta}_{\otimes}(z) = \begin{cases} \lvert z\rvert \oslash f_{\min}, & \text{if } z \ge f^{\mathrm{nor}}_{\min}; \\ \big(\lvert z\rvert \oslash f_{\min}\big) \oplus 2^{-1}, & \text{if } 0 < z < f^{\mathrm{nor}}_{\min} \text{ and even}(z); \\ \Big(\big(\lvert z\rvert \oslash f_{\min}\big) \oplus 2^{-1}\Big)^{-}, & \text{if } 0 < z < f^{\mathrm{nor}}_{\min} \text{ and odd}(z); \end{cases} \qquad \underline{\delta}_{\otimes}(z) = -\overline{\delta}_{\otimes}(z);$$

$$\overline{\delta}_{\oslash}(z) = \begin{cases} \lvert z\rvert \otimes f_{\max}, & \text{if } f^{\mathrm{nor}}_{\min} \le \lvert z\rvert \le 1; \\ \big(\lvert z\rvert \otimes f_{\max}\big) \oplus 2^{q}, {}^{(*)} & \text{if } 0 \le \lvert z\rvert < f^{\mathrm{nor}}_{\min} \wedge \big(\lvert z\rvert \ne 1 \times 2^{e_z} \vee e_z = e_{\min} - 1\big); \\ \Big(\big(\lvert z\rvert \otimes f_{\max}\big) \oplus 2^{q}\Big)^{-}, & \text{otherwise}; \end{cases} \qquad \underline{\delta}_{\oslash}(z) = -\overline{\delta}_{\oslash}(z);$$

$$\tilde{\delta}'_{\oslash}(z) = \begin{cases} f_{\max} \oslash \lvert z\rvert^{--}, & \text{if } 1 < \lvert z\rvert \le f_{\max}; \\ f_{\max}, & \text{otherwise}; \end{cases} \qquad \underline{\tilde{\delta}}'_{\oslash}(z) = -\tilde{\delta}'_{\oslash}(z);$$

$(*)$ $q = 1 - p + e_{\min} + e_{\max}$.

5. Discussion

This work is part of a long-term research effort concerning the correct, precise and efficient handling of floating-point constraints (????????) for software verification purposes.

Restricting the attention to test data generation other authors have considered using search-based techniques with a specific notion of distance in their fitness function (??). For instance, search-based tools like AUSTIN and FloPSy can generate a test input for a specific path by evaluating the path covered by some current input with respect to a targeted path in the program. However, they cannot *solve* the constraints of path conditions, since: 1) they cannot determine unsatisfiability when the path is infeasible, and 2) they can fail to find a test input while the set of constraints is satisfiable (?).

Recently, ? combined a search-based test data generation engine with the RealPaver ? interval constraint solver, which is well-known in the Constraint Programming community. Even though constraint solvers over continuous domains (e.g., RealPaver ?, Quimper ? or ICOS ?) and the work described in the present paper are based on similar principles, the treatment of intervals is completely different. While our approach preserves all the solutions over the floats, it is not at all concerned with solutions over the reals. In contrast, RealPaver preserves solutions over the reals by making the appropriate choices in the rounding modes used for computing the interval bounds, but RealPaver can lose solutions over the floats. For instance, a constraint like $(\mathtt{x} > 0.0 \wedge \mathtt{x} \oplus 10000.0 \leq 10000.0)$ is shown to be unsatisfiable on the reals by RealPaver, while it is satisfied by many IEEE 754 floating-point values of single or double precision format for $\mathtt{x}$ (?). Note that RealPaver has recently been used to tackle test input generation in presence of transcendental functions (?), but this approach, as mentioned by the authors, is neither correct nor complete due to the error rounding of floating-point computations.

6. Conclusion

This paper concerns constraint solving over binary floating-point numbers. Interval-based consistency techniques are very effective for the solution of such numerical constraints, provided precise and efficient filtering algorithms are available. We reformulated and corrected the filtering algorithm proposed by ? for addition and subtraction. We proposed a uniform framework that generalizes the property identified by Marre and Michel to the case of multiplication and division. We also revised, corrected and extended our initial ideas, sketched in ?, to subnormals and to the effective treatment of floating-point division.

An important objective of this work has been to allow maximum efficiency by defining all algorithms in terms of IEEE 754 elementary operations on the same formats as the ones of the filtered constraints. Indeed, the computational cost of filtering by maximum ULP as defined in the present

paper and properly implemented is negligible. As shown in (**?**), the improvement of filtering procedures with these techniques brings speedups of the overall constraint solving process that can be substantial (we have observed up to an order of magnitude); in the cases where filtering by maximum ULP does not allow significant extra pruning, the slowdowns are always of very modest entity (up to a few percent on the overall solution time). In comparison, the choice of different heuristics concerning the selection of constraints and variables to subject to filtering and the labeling strategy has a much more dramatic effect on solution time, even though the positive or negative effects of such heuristics change wildly from one analyzed program to the other. Filtering by maximum ULP contributes to reducing this variability. To understand this, consider the elementary constraint $\mathtt{z} = \mathtt{x} \odot \mathtt{y}$: if `x` and `y` are subject to labeling before `z`, then filtering with maximum ULP will not help. However, `z` might be labeled before `x` or `y`: this can happen under *any* labeling heuristic and constitutes a performance bottleneck.

In the latter case, filtering by maximum ULP may contribute to a much improved pruning of the domains of `x` and `y` and remove the bottleneck.

Future work includes coupling filtering by maximum ULP with sophisticated implementations of classical filtering based on multi-intervals and with dynamic linear relaxation algorithms (**?**) using linear relaxation formulas such as the ones proposed by **?**. Another extension, by far more ambitious, concerns the handling of transcendental functions (i.e., sin, cos, exp, ...): as IEEE 754 does not impose formal correctness requirements upon those functions, solutions will be dependent on the particular implementation and/or be imprecise; in other words, generated test inputs will not be applicable to other implementations and/or may fail to exercise the program paths they were supposed to traverse.

Acknowledgments

We are grateful to Abramo Bagnara (BUGSENG srl, Italy) for the many fruitful discussions we had on the subject of this paper, and to Paul Zimmermann (INRIA Lorraine, France) for the help he gave us proving a crucial result. We are also indebted to Claude Michel for several constructive remarks that allowed us to improve the paper.

Note

The following appendix, taken verbatim from **?**, is included here for the convenience of the reviewers only.

Appendix. Technical Proofs

THEOREM 3. *Function $\bar{\delta}_{\otimes}$ is well-defined and satisfies* (2) *and* (3).

Proof. First note that $\mathbb{F}_{\otimes}$ is the set of all $z \in \mathbb{F}^{\text{sub}}_{p,e_{\max}}$ such that

$$|z| \leq f_{\max} \cdot f_{\min} = (2 - 2^{1-p})2^{e_{\max}+e_{\min}+1-p}$$

and that the range of $\bar{\delta}_{\otimes}$ is the positive subset of $\mathbb{F}_{p,e_{\max}}$. This is because its domain is $\mathbb{F}_{\otimes}$ and multiplication by $2^{-(e_{\min}+1-p)}$, for $z \in \mathbb{F}_{\otimes}$, boils down to summing exponents. Moreover, $\big(|z|/f_{\min}\big) \oplus 2^{-1} = |z|/f_{\min} + 2^{-1}$. In fact, let $|z| = m2^{e_z}$ for some $1 \leq m < 2$. We have

$$m < 2 - 2^{e_{\min}-e_z}2^{1-p}, \tag{9}$$

since z is subnormal and m is a normalized mantissa. Hence,

$$\begin{aligned}
\big(|z|/f_{\min}\big) \oplus 2^{-1} &= [m2^{e_z}/f_{\min} + 2^{-1}]_{\text{n}} \\
&= [m2^{e_z-e_{\min}-1+p} + 2^{-1}]_{\text{n}} \\
&= \big[(m + 2^{e_{\min}-e_z-1}2^{1-p})2^{e_z-e_{\min}-1+p}\big]_{\text{n}} \\
&= (m + 2^{e_{\min}-e_z-1}2^{1-p})2^{e_z-e_{\min}-1+p} \qquad (10) \\
&= |z|/f_{\min} + 2^{-1},
\end{aligned}$$

where (10) holds because of (9).

Consider now the following cases:

$f^{\text{nor}}_{\min} \leq z \leq (2 - 2^{1-p})2^{e_{\max}+e_{\min}+1-p}$: We have $\bar{\delta}_{\otimes}(z) = |z|2^{-(e_{\min}+1-p)}$, hence $y = f_{\min} = 2^{e_{\min}+1-p}$ satisfies (2):

$$\begin{aligned}
\bar{\delta}_{\otimes}(z) \otimes y &= \big(|z|2^{-(e_{\min}+1-p)}\big) \otimes 2^{e_{\min}+1-p} \\
&= \big[|z|2^{-(e_{\min}+1-p)}2^{e_{\min}+1-p}\big]_{\text{n}} \qquad (11) \\
&= |z| \\
&= z.
\end{aligned}$$

Eq. (11) holds because, since z is normal, we have $z2^{-(e_{\min}+1-p)} \leq f_{\max}$. In order to prove (3), we have to show that, for each $z' > \bar{\delta}_{\otimes}(z)$ there does not exist $y \in \mathbb{F}^{\text{sub}}_{p,e_{\max}}$ such that $z' \otimes y = z$. By monotonicity of $\otimes$, a y satisfying $z' \otimes y = z$ should be smaller than or equal to $f_{\min}$ and greater than $+0$. However, the smallest float in $\mathbb{F}^{\text{sub}}_{p,e_{\max}}$ that is greater than $+0$ is $f_{\min}$. Hence we are left to prove that $\forall z' > \bar{\delta}_{\otimes}(z) : z' \otimes f_{\min} > z$. Since $z' \geq \bar{\delta}_{\otimes}(z)^{+}$, we have two cases:

$\bar{\delta}_{\otimes}(z)^{+} = +\infty$: In this case, $z' \otimes f_{\min} = +\infty > z$.

$\bar{\delta}_\otimes(z)^+ \neq +\infty$: Letting $z = m \times 2^{e_z}$ we have

$$\begin{aligned}
\bar{\delta}_\otimes(z)^+ &= (m \times 2^{e_z - e_{\min} - 1 + p})^+ \\
&= (m + 2^{1-p}) 2^{e_z - e_{\min} - 1 + p} \\
&= m 2^{e_z - e_{\min} - 1 + p} + 2^{e_z - e_{\min}} \\
&= \bar{\delta}_\otimes(z) + 2^{e_z - e_{\min}},
\end{aligned}$$

hence

$$\begin{aligned}
z' \otimes f_{\min} &= [z' f_{\min}]_{\mathrm{n}} \\
&\geq \left[(\bar{\delta}_\otimes(z) + 2^{e_z - e_{\min}}) f_{\min}\right]_{\mathrm{n}} \\
&= \left[(z f_{\min}^{-1} + 2^{e_z - e_{\min}}) f_{\min}\right]_{\mathrm{n}} \\
&= [z + 2^{e_z - e_{\min}} f_{\min}]_{\mathrm{n}} \\
&= [z + 2^{e_z - e_{\min}} 2^{e_{\min} + 1 - p}]_{\mathrm{n}} \\
&= [z + 2^{e_z + 1 - p}]_{\mathrm{n}} \\
&= z^+ \qquad (12) \\
&> z,
\end{aligned}$$

where (12) holds because $z \geq f_{\min}^{\mathrm{nor}}$. In any case, (3) holds.

$0 < z < f_{\min}^{\mathrm{nor}}$ *and* $\mathrm{even}(z)$: We have $\bar{\delta}_\otimes(z) = |z| 2^{-(e_{\min} + 1 - p)} + 2^{-1}$, hence $y = f_{\min} = 2^{e_{\min} + 1 - p}$ satisfies (2):

$$\begin{aligned}
\bar{\delta}_\otimes(z) \otimes f_{\min} &= \left[\left((z / f_{\min}) + 2^{-1}\right) f_{\min}\right]_{\mathrm{n}} \\
&= [z + 2^{-1} 2^{e_{\min} + 1 - p}]_{\mathrm{n}} \\
&= [z + 2^{e_{\min} - p}]_{\mathrm{n}} \\
&= [z + \Delta_z^+ / 2]_{\mathrm{n}} \\
&= z. \qquad (13)
\end{aligned}$$

Note that, as we have $\mathrm{even}(z)$, (13) holds by Definition 2

In order to prove (3), we have to show that, for each $z' > \bar{\delta}_\otimes(z)$, $z' \otimes f_{\min} > z$. Of course, as observed in the previous case, y cannot be smaller than $f_{\min}$. However, for each $z' \geq \left(\bar{\delta}_\otimes(z)\right)^+$, we have

$$\begin{aligned}
z' \otimes f_{\min} &\geq \left(\bar{\delta}_\otimes(z)\right)^+ \otimes f_{\min} \qquad (14) \\
&> \left[\left((z / f_{\min}) + 2^{-1} + 2^{1-p} 2^{e_z - e_{\min} - 1 + p}\right) f_{\min}\right]_{\mathrm{n}} \qquad (15) \\
&= [z + 2^{e_{\min} - p} + 2^{1 - p + e_z}]_{\mathrm{n}} \\
&> [z + \Delta_z^+ / 2]_{\mathrm{n}} \\
&\geq z^+, \qquad (16)
\end{aligned}$$

where (14) holds by monotonicity of $\otimes$, (15) holds because $\exp\left(\bar{\delta}_\otimes(z)\right) = \exp(z / f_{\min} + 2^{-1}) \geq e_z - e_{\min} - 1 + p$, and (16) holds by Definition 2.

$0 < z < f^{\text{nor}}_{\text{min}}$ *and* odd(z): We have $\bar{\delta}_{\otimes}(z) = (|z|2^{-(e_{\min}+1-p)} + 2^{-1})^{-}$ and we prove that (2) is satisfied with $y = f_{\min} = 2^{e_{\min}+1-p}$. To this aim we show that $\bar{\delta}_{\otimes}(z) \otimes f_{\min} = \left[\bar{\delta}_{\otimes}(z) f_{\min}\right]_{\text{n}} = z$. In order to prove the latter equality, by Definition 2, we need to show that $z - 2^{e_{\min}-p} \le \bar{\delta}_{\otimes}(z) f_{\min} \le z + 2^{e_{\min}-p}$. In fact, on the one hand we have

$$\begin{aligned}
\bar{\delta}_{\otimes}(z) f_{\min} &\le (z/f_{\min} + 2^{-1} - 2^{1-p}2^{e_z - e_{\min} - 1 + p}) f_{\min} && (17)\\
&= z + 2^{-1}2^{e_{\min}+1-p} - 2^{1-p+e_z} \\
&= z + 2^{e_{\min}-p} - 2^{1-p+e_z} \\
&< z + 2^{e_{\min}-p},
\end{aligned}$$

where (17) holds because $\exp\left(\bar{\delta}_{\otimes}(z) + 2^{-1}\right) \le \exp(z f_{\min}) = e_z - e_{\min} - 1 + p$. On the other hand, we can prove that $\bar{\delta}_{\otimes}(z) f_{\min} \ge z - 2^{e_{\min}-p}$:

$$\begin{aligned}
\bar{\delta}_{\otimes}(z) f_{\min} &\ge (z/f_{\min} + 2^{-1} - 2^{1-p}2^{e_z - e_{\min} + p}) f_{\min} && (18)\\
&= z + 2^{-1}2^{e_{\min}+1-p} - 2^{-p+e_z} \\
&= z + 2^{e_{\min}-p} - 2^{-p+e_z} \\
&> z - 2^{e_{\min}-p}, && (19)
\end{aligned}$$

where (18) holds because $\exp(\bar{\delta}_{\otimes}(z) + 2^{-1}) \ge \exp(z f_{\min} + 1) = e_z - e_{\min} + p$ and, since z is subnormal, (19) holds because $2^{-p+e_z} < 2^{e_{\min}-p}$. By Definition 2, we can conclude that $\bar{\delta}_{\otimes}(z) \otimes f_{\min} = \left[\bar{\delta}_{\otimes}(z) f_{\min}\right]_{\text{n}} = z$, as we have odd$(z)$.

In order to prove (3), we have to show that, for each $z' > \bar{\delta}_{\otimes}(z)$, $z' \otimes f_{\min} > z$. Again, y cannot be smaller than $f_{\min}$ and for $z' \ge \left(\bar{\delta}_{\otimes}(z)\right)^{+}$ we have:

$$\begin{aligned}
z' \otimes f_{\min} &\ge \left(\bar{\delta}_{\otimes}(z)\right)^{+} \otimes f_{\min} \\
&= \left(\left([z/f_{\min} + 2^{-1}]_{\text{n}}\right)^{-}\right)^{+} \\
&= [z/f_{\min} + 2^{-1}]_{\text{n}} \\
&= [z + 2^{-1}2^{e_{\min}+1-p}]_{\text{n}} \\
&= [z + 2^{e_{\min}-p}]_{\text{n}} \\
&= [z + \Delta_z^{+}/2]_{\text{n}} \\
&= z^{+}. && (20)
\end{aligned}$$

Note that (20) holds by Definition 2, since we have odd(z).

$-(2 - 2^{1-p})2^{e_{\max}+e_{\min}+1-p} \le z < 0$: Choosing $y = -f_{\min}$ we can reason, depending on the value of $|z|$, as in the previous cases. □

PROPOSITION 5. *Let $z \in \mathbb{F}_{\otimes}$ be nonzero. If $z > 0$ then $\bar{\delta}_{\otimes}(z^{+}) \ge \bar{\delta}_{\otimes}(z)$; if $z < 0$ then $\bar{\delta}_{\otimes}(z^{-}) \ge \bar{\delta}_{\otimes}(z)$.*

Proof. Assume $z > 0$, the other case being symmetric. For $z \ge f^{\text{nor}}_{\text{min}}$ the property holds by monotonicity of division on the dividend. The following cases remain:

$0 < z < (f^{\mathrm{nor}}_{\mathrm{min}})^-$ *and* $\mathrm{even}(z)$: We need to show that $\bar{\delta}_{\otimes}(z^+) \geq \bar{\delta}_{\otimes}(z)$. Since z is subnormal, by Definition 5 and the observation that all the floating-point operations that occur in it are exact, we have

$$\begin{aligned}
\bar{\delta}_{\otimes}(z^+) &= \big((z + 2^{1-p+e_{\mathrm{min}}})/f_{\mathrm{min}} + 2^{-1}\big)^- \\
&\geq (z + 2^{1-p+e_{\mathrm{min}}})/f_{\mathrm{min}} + 2^{-1} - 2^{e_z-1+p-e_{\mathrm{min}}} && (21)\\
&= z/f_{\mathrm{min}} + 1 + 2^{-1} - 2^{e_z-1+p-e_{\mathrm{min}}} \\
&\geq z/f_{\mathrm{min}} + 2^{-1} && (22)\\
&= \bar{\delta}_{\otimes}(z),
\end{aligned}$$

where (21) holds because $\exp\big(\bar{\delta}_{\otimes}(z) + 2^{-1}\big) \geq \exp(z f_{\mathrm{min}} + 1) = e_z - e_{\mathrm{min}} + p$, whereas (22) holds because $2^{e_z-1+p-e_{\mathrm{min}}} \leq 1$.

$0 < z < (f^{\mathrm{nor}}_{\mathrm{min}})^-$ *and* $\mathrm{odd}(z)$: This case is trivial since

$$\begin{aligned}
\bar{\delta}_{\otimes}(z) &= (z/f_{\mathrm{min}} + 2^{-1})^- \\
&< z/f_{\mathrm{min}} + 2^{-1} \\
&< (z^+)/f_{\mathrm{min}} + 2^{-1} \\
&= \bar{\delta}_{\otimes}(z^+).
\end{aligned}$$

$z = (f^{\mathrm{nor}}_{\mathrm{min}})^-$*:* Note that in this case we have $\mathrm{odd}(z)$, hence,

$$\begin{aligned}
\bar{\delta}_{\otimes}(z) &= (z/f_{\mathrm{min}} + 2^{-1})^- \\
&< z/f_{\mathrm{min}} + 2^{-1} \\
&< z/f_{\mathrm{min}} + 1 \\
&= (z + 2^{1-p+e_{\mathrm{min}}})/f_{\mathrm{min}} \\
&= \bar{\delta}_{\otimes}(z^+) \\
&= \bar{\delta}_{\otimes}(f^{\mathrm{nor}}_{\mathrm{min}}).
\end{aligned}$$

□

LEMMA 1. *If* $z \in \mathbb{F}'_{\oslash}$*, then* $(z \otimes f_{\mathrm{max}}) \oslash f_{\mathrm{max}} = z$.

Proof.[9] As $[\cdot]_{\mathrm{n}}$ is a symmetric rounding mode we can focus on the cases where $+0 \leq z \leq 1$: the cases where $-1 \leq z \leq -0$ are symmetric. We thus consider the following cases:
$z = 1$: We have $z \otimes f_{\mathrm{max}} = [z f_{\mathrm{max}}]_{\mathrm{n}} = f_{\mathrm{max}}$, hence,

$$\begin{aligned}
(z \otimes f_{\mathrm{max}}) \oslash f_{\mathrm{max}} &= \big[(z \otimes f_{\mathrm{max}})/f_{\mathrm{max}}\big]_{\mathrm{n}} \\
&= [f_{\mathrm{max}}/f_{\mathrm{max}}]_{\mathrm{n}} \\
&= 1 \\
&= z.
\end{aligned}$$

[9] The main idea of this proof is due to Paul Zimmermann, INRIA, France.

$z = 1/2$: As $z \otimes f_{\max} = [2^{-1} f_{\max}]_{\mathrm{n}} = \big[(2 - 2^{1-p})2^{e_{\max}-1}\big]_{\mathrm{n}} = (2 - 2^{1-p})2^{e_{\max}-1}$, we have

$$\begin{aligned}(z \otimes f_{\max}) \oslash f_{\max} &= \big[(z \otimes f_{\max})/f_{\max}\big]_{\mathrm{n}} \\ &= \left[\frac{(2 - 2^{1-p})2^{e_{\max}-1}}{(2 - 2^{1-p})2^{e_{\max}}}\right]_{\mathrm{n}} \\ &= 1/2 \\ &= z.\end{aligned}$$

$1/2 < z < 1$: In this case we have

$$\begin{aligned}z \otimes f_{\max} &= [z f_{\max}]_{\mathrm{n}} && (23)\\ &= \big[z(2 - 2^{1-p})2^{e_{\max}}\big]_{\mathrm{n}} && (24)\\ &= \big[z(1 - 2^{-p})2^{e_{\max}+1}\big]_{\mathrm{n}} && (25)\\ &= \big[z(1 - 2^{-p})\big]_{\mathrm{n}} 2^{e_{\max}+1} && (26)\\ &= [z - z2^{-p}]_{\mathrm{n}} 2^{e_{\max}+1} && (27)\\ &= z^{-} \cdot 2^{e_{\max}+1}. && (28)\end{aligned}$$

Note that equality (26) holds because the multiplication by $2^{e_{\max}+1}$ can give rise neither to an overflow, since $z f_{\max} < f_{\max}$, nor to an underflow, since $z(1 - 2^{-p}) > 2^{-1}(1 - 2^{-p}) \gg f_{\min}$. To see why equality (28) holds, recall Definition 2 and consider that $\Delta_z^- = \Delta_z^+ = 2^{-p}$ for $1/2 < z < 1$; we thus have $z^- - \Delta_{z^-}^- /2 = (z - 2^{-p}) - 2^{-p-1} < z - z2^{-p} < z - 2^{-p-1} = z^- + \Delta_{z^-}^+ /2$. Now we can write

$$\begin{aligned}(z \otimes f_{\max})/f_{\max} &= (z^- \cdot 2^{e_{\max}+1})/f_{\max} \\ &= \frac{(z - 2^{-p})2^{e_{\max}+1}}{(1 - 2^{-p})2^{e_{\max}+1}} \\ &= (z - 2^{-p})/(1 - 2^{-p}) \\ &< z,\end{aligned}$$

and, since $z \geq 1/2 + 2^{-p}$, whence $1 - z \leq 1/2 - 2^{-p}$,

$$\begin{aligned}z - \big((z \otimes f_{\max})/f_{\max}\big) &= z - \big((z - 2^{-p})/(1 - 2^{-p})\big) \\ &= (z - z2^{-p} - z + 2^{-p})/(1 - 2^{-p}) \\ &= \big(2^{-p}(1 - z)\big)/(1 - 2^{-p}) \\ &\leq \big(2^{-p}(1/2 - 2^{-p})\big)/(1 - 2^{-p}) \\ &= 2^{-p}\big((1/2 - 2^{-p})/(1 - 2^{-p})\big) \\ &< 2^{-p} \cdot 1/2 \\ &= \Delta_z^- /2.\end{aligned}$$

As $0 < z - \big((z \otimes f_{\max})/f_{\max}\big) < \Delta_z^- /2$, we have $z - \Delta_z^- /2 < (z \otimes f_{\max}) \oslash f_{\max} < z$. Hence, by Definition 2, we can conclude that $\big[(z \otimes f_{\max})/f_{\max}\big]_{\mathrm{n}} = z$.

$f^{\text{nor}}_{\min} \le z < 1/2$: In this case z is such that $2^{-\ell} \le z < 2^{-\ell+1}$ with $-e_{\min} \le \ell \le 2$, and we can apply the same reasoning of the last two cases above by substituting the exponent -1 with the exponent $-\ell$; this is because $z \otimes f_{\max}$ does never generate an overflow (a fortiori, since z is now smaller) nor an underflow, because $z(1-2^{-p}) \ge 2^{e}_{\min}(1-2^{-p}) > f_{\min}$.

$2^{e_{\min}-1} < z < f^{\text{nor}}_{\min}$: In this case we have

$$\begin{aligned} z \otimes f_{\max} &= [z f_{\max}]_{\text{n}} && (29)\\ &= \left[z(2-2^{1-p})2^{e_{\max}}\right]_{\text{n}} && (30)\\ &= \left[(2z - z2^{1-p})2^{e_{\max}}\right]_{\text{n}} && (31)\\ &= \left[(z - z2^{-p})2^{e_{\max}+1}\right]_{\text{n}} && (32)\\ &= (z2^{e_{\max}+1})^{-}. && (33) \end{aligned}$$

To see why (33) holds, note that we can express z as $m \times 2^{e_z}$ with $1 < m < 2$ and $e_z = e_{\min} - 1$. Then $z2^{e_{\max}+1} = m2^{e_z+e_{\max}+1}$. Since $m > 1$,

$$\begin{aligned} \Delta^{-}_{z2^{e_{\max}+1}} &= z2^{e_{\max}+1} - (z2^{e_{\max}+1})^{-}\\ &= m2^{e_z+e_{\max}+1} - (m-2^{1-p})2^{e_z+e_{\max}+1}\\ &= 2^{1-p}2^{e_z+e_{\max}+1}. && (34) \end{aligned}$$

Similarly,

$$\begin{aligned} \Delta^{+}_{(z2^{e_{\max}+1})^{-}} &= \left((z2^{e_{\max}+1})^{-}\right)^{+} - (z2^{e_{\max}+1})^{-}\\ &= z2^{e_{\max}+1} - (z2^{e_{\max}+1})^{-}\\ &= 2^{1-p}2^{e_z+e_{\max}+1}. && (35) \end{aligned}$$

Finally, exploiting once again the fact that $m > 1$,

$$\begin{aligned} \Delta^{-}_{(z2^{e_{\max}+1})^{-}} &= (z2^{e_{\max}+1})^{-} - \left((z2^{e_{\max}+1})^{-}\right)^{-}\\ &\le (m-2^{1-p})2^{e_z+e_{\max}+1} - (m-2^{2-p})2^{e_z+e_{\max}+1} && (36)\\ &= 2^{1-p}2^{e_z+e_{\max}+1}. && (37) \end{aligned}$$

For (36), note that $m > 1$ implies that $(z2^{e_{\max}+1})^{-} = (m-2^{1-p})2^{e_z+e_{\max}+1}$. Applying the same reasoning to $\left((z2^{e_{\max}+1})^{-}\right)^{-} = \left((m-2^{1-p})2^{e_z+e_{\max}+1}\right)^{-}$, we have two cases:

$(m-2^{1-p}) > 1$: then, as before, we have $\Delta^{-}_{(z2^{e_{\max}+1})^{-}} = 2^{1-p}2^{e_z+e_{\max}+1}$ and thus

$$\begin{aligned} \left((z2^{e_{\max}+1})^{-}\right)^{-} &= (m-2^{1-p})2^{e_z+e_{\max}+1} - 2^{1-p}2^{e_z+e_{\max}+1}\\ &= (m-2^{2-p})2^{e_z+e_{\max}+1}; \end{aligned}$$

as a consequence, (36) holds with the equality;

$(m - 2^{1-p}) = 1$: in this case $\Delta^-_{(z2^{e_{\max}+1})^-} = 2^{1-p}2^{e_z+e_{\max}}$, hence

$$\begin{aligned}\left((z2^{e_{\max}+1})^-\right)^- &= (m - 2^{1-p})2^{e_z+e_{\max}+1} - 2^{1-p}2^{e_z+e_{\max}} \\ &= (m - 2^{1-p} - 2^{-p})2^{e_z+e_{\max}+1};\end{aligned}$$

as a consequence, (36) holds with the inequality.

In order to prove (33), by Definition 2, we have to show

$$(z2^{e_{\max}+1})^- - \frac{\Delta^-_{(z2^{e_{\max}+1})^-}}{2} < (z - z2^{-p})2^{e_{\max}+1} \tag{38}$$

$$< (z2^{e_{\max}+1})^- + \frac{\Delta^+_{(z2^{e_{\max}+1})^-}}{2}. \tag{39}$$

To prove (38) observe that, by (34),

$$(z2^{e_{\max}+1})^- = z2^{e_{\max}+1} - \Delta^-_{z2^{e_{\max}+1}} = z2^{e_{\max}+1} - 2^{1-p+e_z+e_{\max}+1}. \tag{40}$$

Hence, by (37), we have

$$\begin{aligned}(z2^{e_{\max}+1})^- - \frac{\Delta^-_{(z2^{e_{\max}+1})^-}}{2} &\leq (z2^{e_{\max}+1}) - 2^{1-p+e_z+e_{\max}+1} - 2^{-p+e_z+e_{\max}+1} \\ &< (z - 2^{1-p+e_z})2^{e_{\max}+1} \\ &< (z - m2^{-p+e_z})2^{e_{\max}+1} \qquad (41) \\ &= (z - z2^{-p})2^{e_{\max}+1},\end{aligned}$$

where (41) holds because $1 < m < 2$. We are left to prove (39). To this aim, we write the following sequence of inequalities, which are all equivalent:

$$(z - z2^{-p})2^{e_{\max}+1} < (z2^{e_{\max}+1})^- + \Delta^+_{(z2^{e_{\max}+1})^-}/2 \tag{42}$$

$$(z - z2^{-p})2^{e_{\max}+1} < (z2^{e_{\max}+1} - 2^{1-p+e_z+e_{\max}+1}) + 2^{-p+e_z+e_{\max}+1} \tag{43}$$

$$\begin{aligned} z - z2^{-p} &< (z - 2^{1-p+e_z}) + 2^{-p+e_z} \\ -z2^{-p} &< -2^{-p+e_z} \\ 2^{-p+e_z} &< z2^{-p} \\ 2^{-p+e_z} &< (m2^{e_z})2^{-p} \\ 1 &< m \end{aligned}$$

where (42) is equivalent to (43) because of (40) and (35). Moreover, since we have decomposed z so that $1 < m < 2$, the last inequality holds and we can conclude that $z \otimes f_{\max} = (z2^{e_{\max}+1})^-$. Now we can write

$$\begin{aligned}(z \otimes f_{\max})/f_{\max} &= (z2^{e_{\max}+1})^-/f_{\max} \\ &= \frac{(z - 2^{1-p+e_z})2^{e_{\max}+1}}{(1 - 2^{-p})2^{e_{\max}+1}} \\ &= \frac{z - 2^{1-p+e_z}}{1 - 2^{-p}}.\end{aligned}$$

As in the previous case, we want to show that $z - (z \otimes f_{\max})/f_{\max} < \Delta_z^- /2$, since this will guarantee that $(z \otimes f_{\max}) \oslash f_{\max} = z$. In fact,

$$\begin{aligned}
z - (z \otimes f_{\max})/f_{\max} &= \frac{z - (z - 2^{1-p+e_z})}{1 - 2^{-p}} \\
&= \frac{z - z2^{-p} - z + 2^{1-p+e_z}}{1 - 2^{-p}} \\
&= \frac{-z2^{-p} + 2^{1-p+e_z}}{1 - 2^{-p}} \\
&= \frac{2^{e_{\min}-p} - z2^{-p}}{1 - 2^{-p}} && (44) \\
&< \frac{2^{e_{\min}-p} - 2^{e_{\min}-p-1}}{1 - 2^{-p}} && (45) \\
&= \frac{2^{e_{\min}-p-1}}{1 - 2^{-p}} \\
&< 2^{e_{\min}-p} \\
&= \Delta_z^- /2, && (46)
\end{aligned}$$

where Eq. (44) holds as $e_z = e_{\min} - 1$; moreover, (45) holds as $2^{e_{\min}-1} < z < f_{\min}^{\text{nor}}$; and (46) holds because, since z is subnormal, $\Delta_z^- = f_{\min}$. From $0 < z - (z \otimes f_{\max})/f_{\max} < \Delta_z^- /2$ we get $z - \Delta_z^- /2 < (z \otimes f_{\max})/f_{\max} < z$. Thus, by Definition 2, we can conclude $(z \otimes f_{\max}) \oslash f_{\max} = \big[(z \otimes f_{\max})/f_{\max}\big]_{\text{n}} = z$.

$z = 2^{e_{\min}-1}$: We have

$$\begin{aligned}
z \otimes f_{\max} &= \big[2^{e_{\min}-1} 2^{e_{\max}} (2 - 2^{1-p})\big]_{\text{n}} \\
&= \big[(2 - 2^{1-p}) 2^{e_{\max}+e_{\min}-1}\big]_{\text{n}} \\
&= (2 - 2^{1-p}) 2^{e_{\max}+e_{\min}-1},
\end{aligned}$$

hence

$$\begin{aligned}
\big[(z \otimes f_{\max})/f_{\max}\big]_{\text{n}} &= \left[\frac{(2 - 2^{1-p}) 2^{e_{\max}+e_{\min}-1}}{(2 - 2^{1-p}) 2^{e_{\max}}}\right]_{\text{n}} \\
&= 2^{e_{\min}-1} \\
&= z.
\end{aligned}$$

$f_{\min} \le z < 2^{e_{\min}-1}$: In this case z is such that $2^{-\ell} \le z < 2^{-\ell+1}$ provided that $-(e_{\min} - p + 1) \le \ell \le -e_{\min} + 2$, hence, we can apply the same reasoning of the last two cases above by substituting the exponent $e_{\min} - 1$ with ℓ.

$z = 0$: Note that, for $z = +0$, we have $(z \otimes f_{\max}) \oslash f_{\max} = +0 \oslash f_{\max} = +0$ while, for $z = -0$, we have $(z \otimes f_{\max}) \oslash f_{\max} = -0 \oslash f_{\max} = -0$.

□

LEMMA 2. *The restriction of $\bar{\delta}_{\oslash}$ to $\mathbb{F}'_{\oslash} \cap \mathbb{F}_{p,e_{\max}}$ is well-defined and satisfies* (2) *and* (3).

Proof. Note that the range of $\bar{\delta}_{\oslash}$ is constituted by non negative elements of $\mathbb{F}_{p,e_{\max}}$.

Consider first the case where $z > 0$. By definition, $\bar{\delta}_{\oslash}(z) = z \otimes f_{\max}$; hence, choosing $y = f_{\max}$ and applying Lemma 1, we get $\bar{\delta}_{\oslash}(z) \oslash y = (z \otimes f_{\max}) \oslash f_{\max} = z$, so that (2) holds. In order to prove (3), we have to

show that, for each $z' \in \mathbb{F}^{\text{sub}}_{p,e_{\max}}$ with $z' > \bar{\delta}_{\oslash}(z)$, there is no $y \in \mathbb{F}^{\text{sub}}_{p,e_{\max}}$ such that $z' \oslash y = z$. We first prove that $z' \oslash f_{\max} > z$. Let $\hat{z}$ be the smallest floating-point number strictly greater than $\bar{\delta}_{\oslash}(z) = z \otimes f_{\max}$, i.e., $\hat{z} = z \otimes f_{\max} + 2^{1-p+\exp(z\otimes f_{\max})}$. We have two cases:

$\exp(z \otimes f_{\max}) = e_z + e_{\max} + 1$: Then

$$\hat{z}/f_{\max} = \frac{(z \otimes f_{\max}) + 2^{1-p+e_z+e_{\max}+1}}{f_{\max}}$$

and, following the steps (23)–(28) of the proof of Lemma 1, we obtain

$$\begin{aligned}
\hat{z}/f_{\max} &= \frac{(z \otimes f_{\max}) + 2^{1-p}2^{e_z+e_{\max}+1}}{f_{\max}} \\
&= \frac{(z - 2^{1-p+e_z})2^{e_{\max}+1} + 2^{2-p+e_z+e_{\max}}}{f_{\max}} \\
&= \frac{(2z - 2^{2-p+e_z})2^{e_{\max}} + 2^{2-p+e_z+e_{\max}}}{(2-2^{1-p})2^{e_{\max}}} \\
&= \frac{2z}{2-2^{1-p}} \\
&= \frac{z}{1-2^{-p}}.
\end{aligned}$$

We now want to show that $\hat{z} \oslash f_{\max} = [\hat{z}/f_{\max}]_{\text{n}} \geq z^+$. Hence, by Definition 2, we need to prove that $z/(1-2^{-p}) > z + \Delta^+_z/2 = z^+ - \Delta^-_{z^+} = z + 2^{-p+e_z}$. To this aim we write the following sequence of inequalities, which are all equivalent:

$$\begin{aligned}
\frac{z}{1-2^{-p}} &> z + 2^{-p+e_z} \\
z &> z + 2^{-p+e_z} - z2^{-p} - 2^{-2p+e_z} \\
0 &> 2^{-p+e_z} - z2^{-p} - 2^{-2p+e_z} \\
0 &> 2^{-p+e_z} - m2^{-p+e_z} - 2^{-2p+e_z} \\
0 &> (1-m)2^{-p+e_z} - 2^{-2p+e_z}.
\end{aligned}$$

Since $z \in \mathbb{F}_{p,e_{\max}}$, $z = m \times 2^{e_z}$ with $1 \leq m < 2$. Hence, the last inequality holds and, by Definition 2, round-to-nearest gives $[\hat{z}/f_{\max}]_{\text{n}} = \hat{z} \oslash f_{\max} \geq z^+ > z$.

$\exp(z \otimes f_{\max}) = e_z + e_{\max}$: This implies that $z = 1.0\ldots0 \times 2^{-\ell}$ for some ℓ such that $-e_{\min} \leq \ell \leq 0$. In fact, $z \geq f^{\text{nor}}_{\min}$ as $z \in \mathbb{F}'_{\oslash} \cap \mathbb{F}_{p,e_{\max}}$. We thus have that $z \otimes f_{\max} = (2-2^{1-p})2^{e_{\max}-\ell}$ and

$$\begin{aligned}
\hat{z}/f_{\max} &= \frac{(z \otimes f_{\max}) + 2^{1-p-\ell+e_{\max}}}{f_{\max}} \\
&= \frac{(2-2^{1-p})2^{e_{\max}-\ell} + 2^{1-p-\ell+e_{\max}}}{f_{\max}} \\
&= \frac{2^{1+e_{\max}-\ell}}{(2-2^{1-p})2^{e_{\max}}} \\
&= \frac{2^{1-\ell}}{2-2^{1-p}} \\
&= \frac{2^{-\ell}}{1-2^{-p}}.
\end{aligned}$$

As in the previous case, we want to show that $\hat{z} \oslash f_{\max} = [\hat{z}/f_{\max}]_{\mathrm{n}} \geq z^+$. Hence, by Definition 2, we need to prove that $2^{-\ell}/(1-2^{-p}) > z + \Delta_z^+/2 = z^+ - \Delta_{z^+}^- = 2^{-\ell} + 2^{-\ell-p}$. To this aim we write the following sequence of inequalities, which are all equivalent:

$$\begin{aligned} \frac{2^{-\ell}}{1-2^{-p}} &> 2^{-\ell} + 2^{-\ell-p} \\ 2^{-\ell} &> 2^{-\ell} + 2^{-\ell-p} - 2^{-\ell-p} - 2^{-2p-\ell} \\ 0 &> -2^{-2p-\ell}. \end{aligned}$$

Since the last inequality holds, we can conclude that round-to-nearest gives $[\hat{z}/f_{\max}]_{\mathrm{n}} = \hat{z} \oslash f_{\max} \geq z^+ > z$. In both cases an $y \in \mathbb{F}^+_{p,e_{\max}}$ satisfying $z' \oslash y = z$ should be greater than $f_{\max}$ and less than $+\infty$: as such y does not exist, (3) holds.

For the case where $z < 0$ we can reason as before choosing $y = -f_{\max}$. □

LEMMA 3. *The restriction of $\bar{\delta}_\oslash$ to $\mathbb{F}'_\oslash \setminus \mathbb{F}_{p,e_{\max}}$ is well-defined and satisfies* (2) *and* (3).

Proof. As already observed, the range of $\bar{\delta}_\oslash$ is constituted by non negative elements of $\mathbb{F}_{p,e_{\max}}$.

Consider first the case where $z > 0$. Choosing $y = f_{\max}$ and applying Lemma 1, we obtain $(z \otimes f_{\max}) \oslash y = (z \otimes f_{\max}) \oslash f_{\max} = z$, but this is not enough. In order to prove that (2) holds, we have to show that $\bar{\delta}_\oslash(z) \oslash f_{\max} = z$. We first show that

$$z \otimes f_{\max} = (z 2^{e_{\max}+1})^-. \tag{47}$$

We have two cases on the value of z:

$z = 1 \times 2^{e_z}$ *with* $e_{\min} - p + 1 \leq e_z \leq e_{\min} - 1$: In this case

$$\begin{aligned} z \otimes f_{\max} &= \left[(2 - 2^{1-p}) 2^{e_z + e_{\max}}\right]_{\mathrm{n}} \\ &= [1 \times 2^{e_z + e_{\max} + 1} - 2^{1-p+e_z+e_{\max}}]_{\mathrm{n}} \\ &= [z 2^{e_{\max}+1} - \Delta^-_{z 2^{e_{\max}+1}}]_{\mathrm{n}} \\ &= (z 2^{e_{\max}+1})^-. \end{aligned}$$

$z = m \times 2^{e_z}$ *with* $m > 1$: Following exactly the same steps (29)–(32) of the proof of Lemma 1, we obtain $z \otimes f_{\max} = (z 2^{e_{\max}+1})^-$.

In order to prove $\bar{\delta}_\oslash(z) \oslash f_{\max} = z$, observe that $(z \otimes f_{\max}) \oslash f_{\max} \leq \bar{\delta}_\oslash(z) \oslash f_{\max}$, since $\oslash$ is monotonically non-decreasing in its first argument. By Lemma 1, we have $(z \otimes f_{\max}) \oslash f_{\max} = z$, therefore $z \leq \bar{\delta}_\oslash(z) \oslash f_{\max}$. Hence, by Definition 2, we are left to prove $\bar{\delta}_\oslash(z)/f_{\max} < z + \Delta_z^+/2$. We now distinguish three cases on z:

$z \neq 1 \times 2^{e_z}$: Recall that $q = 1 - p + e_{\min} + e_{\max}$. We begin by proving that we have $\bar{\delta}_\oslash(z) = \big(z \otimes f_{\max}\big) \oplus 2^q = \big(z \otimes f_{\max}\big) + 2^q = (z 2^{e_{\max}+1})^- + 2^q$. Let $z = m 2^{e_z}$, for some m with $1 \leq m < 2$. It is worth to observe that, for $z = m 2^{e_z}$,

$$m < 2 - 2^{e_{\min} - e_z} 2^{1-p}, \tag{48}$$

since the normalized mantissa m was obtained from a denormalized mantissa $m' = 0.0 \cdots 0 b_{e_{\min}-e_z+1} \cdots b_p$ with $b_{e_{\min}-e_z+1} = 1$. Then we can write

$$
\begin{aligned}
(z \otimes f_{\max}) \oplus 2^q &= \big[\big(z \otimes f_{\max}\big) + 2^q\big]_{\mathrm{n}} \\
&= \big[(z2^{e_{\max}+1})^- + 2^q\big]_{\mathrm{n}} && (49)\\
&= \big[(m2^{e_z}2^{e_{\max}+1})^- + 2^q\big]_{\mathrm{n}} \\
&= \big[(m - 2^{1-p})2^{e_{\max}+1+e_z} + 2^q\big]_{\mathrm{n}} \\
&= \big[\big((m - 2^{1-p}) + 2^{1-p}2^{e_{\min}-e_z-1}\big)2^{e_{\max}+1+e_z}\big]_{\mathrm{n}} \\
&= \big((m - 2^{1-p}) + 2^{1-p}2^{e_{\min}-e_z-1}\big)2^{e_{\max}+1+e_z} && (50)\\
&= (m2^{e_z}2^{e_{\max}+1})^- + 2^q \\
&= (z2^{e_{\max}+1})^- + 2^q && (51)\\
&= \big(z \otimes f_{\max}\big) + 2^q,
\end{aligned}
$$

where (49) holds because of Eq.(47). For (50) observe that, by (48), we have $(m - 2^{1-p}) + 2^{1-p}2^{e_{\min}-e_z-1} < 2 - 2^{1-p}$, hence the left-hand side of the latter inequality can be expressed by a normalized mantissa without resorting to a greater exponent.

Now in order to prove that (2) holds, note that the following inequalities are all equivalent:

$$
\begin{aligned}
\frac{(z \otimes f_{\max}) \oplus 2^q}{f_{\max}} &< z + \frac{\Delta_z^+}{2} && (52)\\
\frac{(z2^{e_{\max}+1})^- + 2^q}{f_{\max}} &< z + \frac{\Delta_z^+}{2} && (53)\\
\frac{z2^{e_{\max}+1} - 2^{1-p+e_z+e_{\max}+1} + 2^q}{(1-2^{-p})2^{e_{\max}+1}} &< z + 2^{e_{\min}-p} && (54)\\
\frac{z - 2^{1-p+e_z} + 2^{-p+e_{\min}}}{1-2^{-p}} &< z + 2^{e_{\min}-p} \\
z - 2^{1-p+e_z} + 2^{-p+e_{\min}} &< (z + 2^{e_{\min}-p})(1-2^{-p}) \\
z - 2^{1-p+e_z} + 2^{-p+e_{\min}} &< z + 2^{e_{\min}-p} - z2^{-p} - 2^{e_{\min}-2p} \\
-2^{1-p+e_z} &< -z2^{-p} - 2^{e_{\min}-2p} \\
2^{e_{\min}-p} &< 2^{1+e_z} - z \\
2^{e_{\min}-p} &< (2-m)2^{e_z} \\
2^{e_{\min}-p} &< (2 - (2 - (2^{e_{\min}-e_z}2^{1-p})))2^{e_z} \\
2^{e_{\min}-p} &< (2^{e_{\min}-e_z}2^{1-p}))2^{e_z} \\
2^{e_{\min}-p} &< 2^{e_{\min}+1-p},
\end{aligned}
$$

where (52) is equivalent to (53) because of (51) and because $\Delta_z^+ = f_{\min}$, since z is subnormal. Moreover, (53) is equivalent to (54), since $\Delta^-_{(z \otimes f_{\max}) \oplus 2^q} = 2^{1-p+e_z+e_{\max}+1}$.

In order to prove (3) we need to prove that $\bar{\delta}_{\oslash}(z)^+ \oslash f_{\max} = \left[\bar{\delta}_{\oslash}(z)^+ / f_{\max}\right]_{\mathrm{n}} > z$. By Definition 2 it suffices to prove that $\bar{\delta}_{\oslash}(z)^+ / f_{\max} > z + \Delta_z^+ /2$. We have that

$$\begin{aligned}
\frac{\bar{\delta}_{\oslash}(z)^+}{f_{\max}} &= \frac{(z \otimes f_{\max}) + 2^q + 2^{1-p+\exp(z\otimes f_{\max})}}{f_{\max}} \\
&= \frac{z2^{e_{\max}+1} - 2^{1-p+\exp(z\otimes f_{\max})} + 2^{1-p+\exp(z\otimes f_{\max})} + 2^q}{f_{\max}} \qquad (55)\\
&= \frac{z + 2^{-p+e_{\min}}}{1-2^{-p}} \\
&> z + 2^{e_{\min}-p} \\
&= z + \Delta_z^+ /2,
\end{aligned}$$

where (55) holds because of (47). Hence (3) is proved.

$z = 1 \times 2^{e_{\min}-1}$: We first prove that, in this case, we have

$$\bar{\delta}_{\oslash}(z) = \left(z \otimes f_{\max}\right) \oplus 2^q = z2^{e_{\max}+1}. \qquad (56)$$

By (47) we have that

$$\begin{aligned}
\left(z \otimes f_{\max}\right) \oplus 2^q &= (z2^{e_{\max}+1})^- \oplus 2^q \\
&= (2 - 2^{1-p})2^{e_{\max}+1+e_{\min}-2} \oplus 2^q \\
&= \left[(2 - 2^{1-p})2^{e_{\max}+e_{\min}-1} + 2^q\right]_{\mathrm{n}} \\
&= \left[(2 - 2^{1-p})2^{e_{\max}+e_{\min}-1} + 2^{1-p}2^{e_{\min}+e_{\max}}\right]_{\mathrm{n}} \\
&= \left[\left((2 - 2^{1-p}) + 2^{1-p} + 2^{1-p}\right)2^{e_{\min}+e_{\max}-1}\right]_{\mathrm{n}} \\
&= \left[(1 + 2^{-p})2^{e_{\min}+e_{\max}}\right]_{\mathrm{n}} \\
&= \left[2^{e_{\min}+e_{\max}} + \Delta^+_{2^{e_{\min}+e_{\max}}} /2\right]_{\mathrm{n}} \qquad (57)\\
&= 2^{e_{\min}+e_{\max}} \\
&= z2^{e_{\max}+1},
\end{aligned}$$

where (57) holds by Definition 2 as we have even(z), and so is $z2^{e_{\max}+1} = 2^{e_{\min}+e_{\max}}$.

Then, in order to prove (2), note that the following inequalities are all equivalent:

$$\begin{aligned}
\frac{\bar{\delta}_{\oslash}(z)}{f_{\max}} &< z + \frac{\Delta_z^+}{2} && (58)\\
\frac{z2^{e_{\max}+1}}{(1-2^{-p})2^{e_{\max}+1}} &< z + 2^{e_{\min}-p} && (59)\\
\frac{z}{1-2^{-p}} &< z + 2^{e_{\min}-p} \\
z &< (z + 2^{e_{\min}-p})(1-2^{-p}) \\
z &< z + 2^{e_{\min}-p} - z2^{-p} - 2^{e_{\min}-2p} \\
0 &< 2^{e_{\min}-p} - z2^{-p} - 2^{e_{\min}-2p} \\
0 &< 2^{e_{\min}-p} - 2^{e_{\min}-p-1} - 2^{e_{\min}-2p} \\
0 &< 2^{e_{\min}-p-1} - 2^{e_{\min}-2p} \\
0 &< 2^{-1} - 2^{-p},
\end{aligned}$$

where (58) is equivalent to (59) because of (56). Moreover, assuming $p > 1$, the last inequality holds.

In order to prove (3) , we need to prove that $\bar{\delta}_{\oslash}(z)^+ \oslash f_{\max} > z$. By Definition 2 it suffices to prove that $\bar{\delta}_{\oslash}(z)^+ / f_{\max} > z + \Delta_z^+ /2$. Indeed,

$$\begin{aligned}
\frac{\bar{\delta}_{\oslash}(z)^+}{f_{\max}} &= \frac{z2^{e_{\max}+1} + 2^q}{f_{\max}} \\
&= \frac{z2^{e_{\max}+1} + 2^{1-p+e_{\min}+e_{\max}}}{2^{e_{\max}}(2 - 2^{1-p})} \\
&= \frac{z + 2^{e_{\min}-p}}{1 - 2^{-p}} \\
&> z + 2^{e_{\min}-p} \qquad (60) \\
&= z + \Delta_z^+ /2,
\end{aligned}$$

where (60) holds because of (56). Hence $\bar{\delta}_{\oslash}(z)^+ \oslash f_{\max} = \left[\bar{\delta}_{\oslash}(z)^+ / f_{\max}\right]_{\mathrm{n}} \geq z^+$, which proves (3).

$z = 1 \times 2^{e_z}$ *with* $e_z < e_{\min} - 1$: We first prove that, in this case,

$$\big(z \otimes f_{\max}\big) \oplus 2^q = z2^{e_{\max}+1} + 2^q. \tag{61}$$

Applying (47) we have that

$$\begin{aligned}
\big(z \otimes f_{\max}\big) \oplus 2^q &= (z2^{e_{\max}+1})^- \oplus 2^q \\
&= (2 - 2^{1-p})2^{e_{\max}+1+e_z-1} \oplus 2^q \\
&= \left[(2 - 2^{1-p})2^{e_{\max}+e_z} + 2^q\right]_{\mathrm{n}} \\
&= \left[(2 - 2^{1-p})2^{e_{\max}+e_z} + 2^{1-p}2^{e_{\min}-e_z}2^{e_{\max}+e_z}\right]_{\mathrm{n}} \\
&= \left[\big((2 - 2^{1-p}) + 2^{1-p} + 2^{1-p} + (2^{e_{\min}-e_z} - 2)2^{1-p}\big)2^{e_{\max}+e_z}\right]_{\mathrm{n}} \\
&= \left[(1 + 2^{-p} + (2^{e_{\min}-e_z-1} - 1)2^{1-p})2^{e_{\max}+e_z+1}\right]_{\mathrm{n}} \\
&= \left[(1 + (2^{e_{\min}-e_z-1} - 1)2^{1-p})2^{e_{\max}+e_z+1} + 2^{-p}2^{e_{\max}+e_z+1}\right]_{\mathrm{n}} \\
&= (1 + 2^{1-p} + (2^{e_{\min}-e_z-1} - 1)2^{1-p})2^{e_{\max}+e_z+1} \qquad (62) \\
&= z2^{e_{\max}+1} + 2^{e_{\min}-e_z-1}2^{1-p}2^{e_{\max}+e_z+1} \\
&= z2^{e_{\max}+1} + 2^{e_{\min}+1-p+e_{\max}}.
\end{aligned}$$

In order to appreciate why (62) holds, note first that, as $e_z \geq e_{\min} - p + 1$, we have $1 + (2^{e_{\min}-e_z-1} - 1)2^{1-p} < 1 + 2^{-1}$. This ensures that the floating-point number $\big(1 + (2^{e_{\min}-e_z-1} - 1)2^{1-p}\big)2^{e_{\max}+e_z+1}$ is represented by a normalized mantissa of the form $1.0b_2 \cdots b_p$. Moreover, observe that $1 + (2^{e_{\min}-e_z-1} - 1)2^{1-p}$ — and, consequently, $\big(1 + (2^{e_{\min}-e_z-1} - 1)2^{1-p}\big)2^{e_{\max}+e_z+1}$ — is necessarily represented by an odd mantissa, since the number that multiplies 2^{1-p} is odd. Finally, note that

$$\frac{\Delta^+_{(1+(2^{e_{\min}-e_z-1}-1)2^{1-p})2^{e_{\max}+e_z+1}}}{2} = 2^{-p}2^{e_{\max}+e_z+1}$$

and thus, by Definition 2, since $\mathrm{odd}\big(\big(1 + (2^{e_{\min}-e_z-1} - 1)2^{1-p}\big)2^{e_{\max}+e_z+1}\big)$, we can conclude that (62) holds.

Consider now the following sequence of equivalent inequalities:

$$\frac{\bar{\delta}_{\oslash}(z)}{f_{\max}} < z + \frac{\Delta_z^+}{2} \tag{63}$$

$$\frac{(z2^{e_{\max}+1} + 2^{e_{\min}+1-p+e_{\max}})^-}{(1-2^{-p})2^{e_{\max}+1}} < z + 2^{e_{\min}-p} \tag{64}$$

$$\frac{z2^{e_{\max}+1} + 2^{e_{\min}+1-p+e_{\max}} - 2^{1-p+e_{\max}+e_z+1}}{(1-2^{-p})2^{e_{\max}+1}} < z + 2^{e_{\min}-p} \tag{65}$$

$$\begin{aligned}
\frac{z + 2^{e_{\min}-p} - 2^{1-p+e_z}}{1-2^{-p}} &< z + 2^{e_{\min}-p} \\
z + 2^{e_{\min}-p} - 2^{1-p+e_z} &< (z + 2^{e_{\min}-p})(1-2^{-p}) \\
z + 2^{e_{\min}-p} - 2^{1-p+e_z} &< z + 2^{e_{\min}-p} - z2^{-p} - 2^{e_{\min}-2p} \\
-2^{1-p+e_z} &< -z2^{-p} - 2^{e_{\min}-2p} \\
-2^{1-p+e_z} &< -2^{e_z-p} - 2^{e_{\min}-2p} \\
2^{e_{\min}-2p} &< 2^{-p+e_z} \qquad (66) \\
2^{e_{\min}-2p} &< 2^{e_{\min}-2p+1}, \qquad (67)
\end{aligned}$$

where (63) is equivalent to (64) because of (61), and (66) is equivalent to (67) because $e_z \geq e_{\min} - p + 1$. As for the equivalence between (64) and (65), note that the exponent of $z2^{e_{\max}+1} + 2^{e_{\min}+1-p+e_{\max}}$ is $e_{\max} + e_z + 1$, hence $\Delta^-_{z2^{e_{\max}+1}+2^{e_{\min}+1-p+e_{\max}}} = 2^{1-p}2^{e_{\max}+e_z+1}$. Finally, assuming $p > 1$, the last inequality holds.

In order to prove (3), we need to prove that $\bar{\delta}_{\oslash}(z)^+ \oslash f_{\max} > z$. By Definition 2, it suffices to prove that $\bar{\delta}_{\oslash}(z)^+ / f_{\max} > z + \Delta_z^+ / 2$. In this case we have that

$$\begin{aligned}
\frac{\bar{\delta}_{\oslash}(z)^+}{f_{\max}} &= \frac{z2^{e_{\max}+1} + 2^{e_{\min}+1-p+e_{\max}}}{f_{\max}} \\
&= \frac{z + 2^{-p+e_{\min}}}{1-2^{-p}} \\
&> z + 2^{e_{\min}-p} \\
&= z + \Delta_z^+ / 2,
\end{aligned} \tag{68}$$

where (68) holds because of (61). Hence $\bar{\delta}_{\oslash}(z)^+ \oslash f_{\max} = \left[\bar{\delta}_{\oslash}(z)^+ / f_{\max}\right]_{\mathrm{n}} \geq z^+$, which proves (3).

For $z < 0$ we can reason as before choosing $y = -f_{\max}$. □

THEOREM 4. *$\bar{\delta}_{\oslash}$ is well-defined and satisfies* (2) *and* (3).

Proof. Immediate from Lemma 2 and Lemma 3. □

PROPOSITION 6. *Let $z \in \mathbb{F}_{\oslash}$ be nonzero. If $z > 0$ then $\bar{\delta}_{\oslash}(z^+) \geq \bar{\delta}_{\oslash}(z)$; if $z < 0$ then $\bar{\delta}_{\oslash}(z^-) \geq \bar{\delta}_{\oslash}(z)$.*

Proof. Assume for simplicity that $z > 0$. We need to investigate the following critical cases on z: $0 < z < (f^{\mathrm{nor}}_{\min})^-$ *and* $z = 1 \times 2^{e_z}$ *with* $e_z < e_{\min} - 1$: This case is trivial since

$$\begin{aligned}
\bar{\delta}_{\oslash}(z^+) &= \left((z + 2^{1-p+e_{\min}}) \otimes f_{\max}\right) \oplus 2^q \\
&\geq (z \otimes f_{\max}) \oplus 2^q \\
&\geq \left((z \otimes f_{\max}) \oplus 2^q\right)^- \\
&= \bar{\delta}_{\oslash}(z).
\end{aligned}$$

$z = 1.1\ldots1 \times 2^{e_z}$ *with* $e_z < e_{\min} - 2$: We need to show that $\bar{\delta}_{\oslash}(z^+) \geq \bar{\delta}_{\oslash}(z)$. Note that, by Definition 6, we have

$$\begin{aligned}
\bar{\delta}_{\oslash}(z^+) &= \big(\big((z^+ \otimes f_{\max}\big) \oplus 2^q\big)^- && (69)\\
&= \big((z^+ 2^{e_{\max}+1}) + 2^q\big)^- && (70)\\
&= (z^+ 2^{e_{\max}+1}) + 2^q - 2^{1-p+e_{\max}+e_z+2} && (71)\\
&= (z + 2^{1-p+e_{\min}})2^{e_{\max}+1} + 2^q - 2^{1-p+e_{\max}+e_z+2} && (72)\\
&= (z2^{e_{\max}+1} + 2^{1-p+e_{\max}+e_{\min}}) + 2^q - 2^{1-p+e_{\max}+e_z+2})\\
&> z2^{e_{\max}+1} + 2^q && (73)\\
&> (z2^{e_{\max}+1})^- + 2^q\\
&= \bar{\delta}_{\oslash}(z), && (74)
\end{aligned}$$

where (69) holds by Definition 6 and (70) holds by (61). In order to show that (71) holds, note that the exponent of $z^+ 2^{e_{\max}+1} + 2^{e_{\min}+1-p+e_{\max}}$ is $e_{\max} + e_z + 2$; hence $\Delta^-_{z+2^{e_{\max}+1}+2^{e_{\min}+1-p+e_{\max}}} = 2^{1-p}2^{e_{\max}+e_z+2}$. Eq. (72) holds because z is subnormal, hence $\Delta^+_z = f_{\min}$, whereas (73) holds because we have assumed $e_z < e_{\min} - 2$. Finally, (74) holds because of (51).

$z = (f^{\mathrm{nor}}_{\min})^-$: Namely, in this case, $z = (2 - 2^{2-p})2^{e_{\min}-1}$ and $z^+ = 2^{e_{\min}}$. We can thus write

$$\begin{aligned}
\bar{\delta}_{\oslash}(z) &= (z \otimes f_{\max}) + 2^q\\
&= (z2^{e_{\max}+1})^- + 2^q && (75)\\
&= (2 - 2^{2-p})2^{e_z+e_{\max}+1} - 2^{1-p+e_z+e_{\max}+1} + 2^q\\
&= (2 - 2^{2-p})2^{e_{\min}+e_{\max}} - 2^q + 2^q\\
&= (2 - 2^{2-p})2^{e_{\min}+e_{\max}}\\
&= (2 - 2^{2-p})2^{e_{\max}}2^{e_{\min}}\\
&= (2 - 2^{2-p})2^{e_{\max}} \otimes z^+\\
&< (2 - 2^{1-p})2^{e_{\max}} \otimes z^+\\
&= \bar{\delta}_{\oslash}(z^+),
\end{aligned}$$

where (75) is justified by (47).

Hence, taking into account the monotonicity of $\otimes$ and $\oplus$, we can conclude that $\bar{\delta}_{\oslash}$ is monotone. □

In order to prove Theorem 5 we need the following intermediate result.

LEMMA 4. *Let* $z \in \mathbb{F}^{\mathrm{sub}}_{p,e_{\max}}$ *be such that* $1 < |z| \leq f_{\max}$. *Then* $f_{\max} \oslash \tilde{\delta}'_{\oslash}(z) < |z|$.

Proof. By Definition 7, we have to prove that $f_{\max} \oslash \big(f_{\max} \oslash |z|^{--}\big) < |z|$ for $1 < |z| \leq f_{\max}$. Assume by simplicity that $z > 0$. The case $z < 0$ can be obtained by considering the absolute value of z.

We have the following cases on z:

$z = 1.0\cdots01 \times 2^{e_z}$: We have $z^{--} = (2-2^{1-p})2^{e_z-1}$ and thus

$$\begin{aligned} f_{\max} \oslash z^{--} &= \left[\frac{(2-2^{1-p})2^{e_{\max}}}{(2-2^{1-p})2^{e_z-1}}\right]_{\mathrm{n}} \\ &= \left[2^{e_{\max}-e_z+1}\right]_{\mathrm{n}} \\ &= 2^{e_{\max}-e_z+1}, \end{aligned}$$

therefore

$$\begin{aligned} f_{\max} \oslash \left(f_{\max} \oslash |z|^{--}\right) &= \left[\frac{(2-2^{1-p})2^{e_{\max}}}{2^{e_{\max}-e_z+1}}\right]_{\mathrm{n}} \\ &= \left[(2-2^{1-p})2^{e_z-1}\right]_{\mathrm{n}} \\ &= (2-2^{1-p})2^{e_z-1} \\ &< 1.0\cdots01 \times 2^{e_z} \\ &= z. \end{aligned}$$

$z = 1.0\cdots00 \times 2^{e_z}$: We have $z^{--} = (2-2^{2-p})2^{e_z-1}$ and thus

$$\begin{aligned} f_{\max} \oslash z^{--} &= \left[\frac{(2-2^{1-p})2^{e_{\max}}}{(2-2^{2-p})2^{e_z-1}}\right]_{\mathrm{n}} \\ &= \left[\frac{2-2^{2-p}+2^{1-p}}{2-2^{2-p}}\right]_{\mathrm{n}} 2^{e_{\max}-e_z+1} \qquad (76) \\ &= \left[1+\frac{2^{1-p}}{2-2^{2-p}}\right]_{\mathrm{n}} 2^{e_{\max}-e_z+1} \\ &= (1+2^{1-p})2^{e_{\max}-e_z+1}. \qquad (77) \end{aligned}$$

Eq. (76) holds because the multiplication by $2^{e_{\max}-e_z+1}$ can give rise neither to an overflow — because $z \geq 2$ and thus $f_{\max} \oslash z^{--} < f_{\max}$ — nor to an underflow — because $z \leq 2^{e_{\max}}$ and thus $f_{\max} \oslash z^{--} \gg f_{\min}$. Moreover, Eq. (77) holds because

$$1+\frac{2^{1-p}}{2-2^{2-p}} < 1+2^{-p}+2^{1-p} = 1^+ + \Delta^+_{1^+}/2$$

and

$$1+\frac{2^{1-p}}{2-2^{2-p}} > 1+\frac{2^{1-p}}{2} = 1+2^{-p} = 1+\Delta^+_1/2 = 1^+ - \Delta^-_{1^+}/2.$$

Hence, by Definition 2, $\left[1+\frac{2^{1-p}}{2-2^{2-p}}\right]_{\mathrm{n}} = 1^+ = 1+2^{1-p}$. We can thus write

$$\begin{aligned} f_{\max} \oslash \left(f_{\max} \oslash |z|^{--}\right) &= \left[\frac{(2-2^{1-p})2^{e_{\max}}}{(1+2^{1-p})2^{e_{\max}-e_z+1}}\right]_{\mathrm{n}} \\ &\leq \left[(2-2^{1-p})2^{e_z-1}\right]_{\mathrm{n}} \\ &= (2-2^{1-p})2^{e_z-1} \\ &< 1.0\cdots00 \times 2^{e_z} \\ &= z. \end{aligned}$$

$z \neq 1.0\cdots0 \times 2^{e_z}$ *and* $z \neq 1.0\cdots01 \times 2^{e_z}$: In this case We have $z = m \times 2^{e_z}$ with $1 + 2^{2-p} \leq m \leq (2 - 2^{1-p})$ and thus

$$\begin{aligned} f_{\max} \oslash \left(f_{\max} \oslash |z|^{--}\right) &= \left[\frac{(2-2^{1-p})2^{e_{\max}}}{\left[\frac{(2-2^{1-p})2^{e_{\max}}}{(m-2^{2-p})2^{e_z}}\right]_{\mathrm{n}}}\right]_{\mathrm{n}} \\ &= \left[\frac{(2-2^{1-p})2^{e_{\max}}}{\left[\frac{2-2^{1-p}}{m-2^{2-p}}\right]_{\mathrm{n}} 2^{e_{\max}-e_z}}\right]_{\mathrm{n}} \qquad (78) \\ &= \left[\frac{2-2^{1-p}}{\left[\frac{2-2^{1-p}}{m-2^{2-p}}\right]_{\mathrm{n}}}\right]_{\mathrm{n}} 2^{e_z}, \qquad (79) \end{aligned}$$

where (78) and (79) hold because the multiplications by $2^{e_{\max}-e_z}$ and by 2^{e_z}, respectively, can give rise neither to an overflow nor to an underflow, since $m \geq 1 + 2^{2-p}$. We are thus left to prove that

$$\left[\frac{2-2^{1-p}}{\left[\frac{2-2^{1-p}}{m-2^{2-p}}\right]_{\mathrm{n}}}\right]_{\mathrm{n}} < m \qquad (80)$$

subject to $1 + 2^{2-p} \leq m \leq 2 - 2^{1-p}$. We distinguish two cases on the value of $\left[\frac{2-2^{1-p}}{m-2^{2-p}}\right]_{\mathrm{n}}$:

$\left[\frac{2-2^{1-p}}{m-2^{2-p}}\right]_{\mathrm{n}} \geq \frac{2-2^{1-p}}{m-2^{2-p}}$: Thus

$$\begin{aligned} \left[\frac{2-2^{1-p}}{\left[\frac{2-2^{1-p}}{m-2^{2-p}}\right]_{\mathrm{n}}}\right]_{\mathrm{n}} &\leq \left[\frac{2-2^{1-p}}{\frac{2-2^{1-p}}{m-2^{2-p}}}\right]_{\mathrm{n}} \\ &= [m-2^{2-p}]_{\mathrm{n}} \\ &= m-2^{2-p} \\ &< m, \end{aligned}$$

and (80) holds.

$[\frac{2-2^{1-p}}{m-2^{2-p}}]_{\mathrm{n}} < \frac{2-2^{1-p}}{m-2^{2-p}}$: By Definition 2 we know that

$$\left[\frac{2-2^{1-p}}{m-2^{2-p}}\right]_{\mathrm{n}} + \frac{\Delta^{+}_{\frac{2-2^{1-p}}{m-2^{2-p}}}}{2} > \frac{2-2^{1-p}}{m-2^{2-p}}. \qquad (81)$$

Since $\Delta^{+}_{\frac{2-2^{1-p}}{m-2^{2-p}}} = 2^{1-p}$, from (81) we obtain

$$[\frac{2-2^{1-p}}{m-2^{2-p}}]_{\mathrm{n}} \geq \frac{2-2^{1-p}}{m-2^{2-p}} - 2^{-p}. \qquad (82)$$

Hence, applying (82), we have:

$$\begin{aligned} \left[\frac{2-2^{1-p}}{\left[\frac{2-2^{1-p}}{m-2^{2-p}}\right]_{\mathrm{n}}}\right]_{\mathrm{n}} &\leq \left[\frac{2-2^{1-p}}{\frac{2-2^{1-p}}{m-2^{2-p}} - 2^{-p}}\right]_{\mathrm{n}} \\ &= \left[\frac{(2-2^{1-p})(m-2^{2-p})}{2-2^{1-p}-2^{-p}(m-2^{2-p})}\right]_{\mathrm{n}} \\ &\leq \left[\frac{(2-2^{1-p})(m-2^{2-p})}{2-2^{1-p}-2^{-p}(2)}\right]_{\mathrm{n}} \qquad (83) \end{aligned}$$

$$
\begin{aligned}
&= \left[\frac{(2-2^{1-p})m - 2^{3-p} + 2^{3-2p}}{2-2^{1-p}-2^{1-p}}\right]_{\mathrm{n}} \\
&= \left[\frac{(2-2^{1-p})m - 2^{3-p} + 2^{3-2p}}{2-2^{2-p}}\right]_{\mathrm{n}} \\
&= \left[\frac{(2-2^{2-p}+2^{1-p})m - 2^{3-p} + 2^{3-2p}}{2-2^{2-p}}\right]_{\mathrm{n}} \\
&= \left[m + \frac{2^{1-p}m}{2-2^{2-p}} - \frac{2^{3-p}}{2-2^{2-p}} + \frac{2^{3-2p}}{2-2^{2-p}}\right]_{\mathrm{n}} \\
&\leq \left[m + 2^{1-p}m - 2^{3-p} + 2^{3-2p}\right]_{\mathrm{n}} && (84) \\
&\leq \left[m + 2^{2-p} - 2^{3-p} + 2^{3-2p}\right]_{\mathrm{n}} && (85) \\
&= \left[m + 2^{2-p}(1-2) + 2^{3-2p}\right]_{\mathrm{n}} \\
&= \left[m - 2^{2-p} + 2^{3-2p}\right]_{\mathrm{n}} \\
&\leq \left[m - 2^{1-p}\right]_{\mathrm{n}} \\
&= m^{-} \\
&< m.
\end{aligned}
$$

Note that (83) and (85) hold because $m \leq (2-2^{1-p}) < 2$, whereas (84) holds because $(2-2^{1-p}) > 1$.

In any case (80) holds and this concludes the proof.

□

THEOREM 5. *Let* $\mathbb{F}''_{\oslash} = \mathbb{F}^{\mathrm{sub}}_{p,e_{\max}}$ *and* $\bar{\mathbb{F}}''_{\oslash} = \mathbb{F}^{\mathrm{sub}+}_{p,e_{\max}}$. *Let* $\bar{\delta}'_{\oslash} \colon \mathbb{F}''_{\oslash} \to \bar{\mathbb{F}}''_{\oslash}$ *be a function satisfying* (4). *Then, for* $0 < |z| \leq 1$ *or* $z = +\infty$, $\bar{\delta}'_{\oslash}(z) \leq \tilde{\delta}'_{\oslash}(z)$; *moreover, for* $1 < |z| \leq f_{\max}$, $\bar{\delta}'_{\oslash}(z) < \tilde{\delta}'_{\oslash}(z)$.

Proof. Recall that, by definition, $\bar{\delta}'_{\oslash}$ satisfies (4) and, thus, for each $z \in \mathbb{F}''_{\oslash} \setminus \{-0, +0, -\infty\}$ there exists $x \in \bar{\mathbb{F}}''_{\oslash}$ such that $x \oslash \bar{\delta}'_{\oslash}(z) = z$. There are two cases on z:

$z = +\infty$ *or* $0 < |z| \leq 1$: As we have $\tilde{\delta}'_{\oslash}(z) = f_{\max}$, we just have to show that $\bar{\delta}'_{\oslash}(z) \neq +\infty$. Indeed, if $\bar{\delta}'_{\oslash}(z) = +\infty$, then $x \oslash \bar{\delta}'_{\oslash}(z)$ can only give ± 0 (if $-f_{\max} \leq x \leq f_{\max}$) or NaN (if $x = \pm\infty$), so that (4) cannot be satisfied.

$1 < |z| \leq f_{\max}$: Assume, towards a contradiction, that $\tilde{\delta}'_{\oslash}(z) \leq \bar{\delta}'_{\oslash}(z)$ for some z such that $1 < |z| \leq f_{\max}$. Hence, as $\oslash$ is antitone in its second argument, $f_{\max} \oslash \bar{\delta}'_{\oslash}(z) \leq f_{\max} \oslash \tilde{\delta}'_{\oslash}(z)$. By Lemma 4, $f_{\max} \oslash \tilde{\delta}'_{\oslash}(z) < z$, hence we also have $f_{\max} \oslash \bar{\delta}'_{\oslash}(z) < z$. This contradicts the hypothesis that $\bar{\delta}'_{\oslash}$ satisfies (4). In fact, as $\oslash$ is monotone in its first argument, $x \oslash \bar{\delta}'_{\oslash}(z) = z$ would require $x > f_{\max}$ or, equivalently $x = +\infty$. But $+\infty \oslash \bar{\delta}'_{\oslash}(z)$ is either equal to $\pm\infty$, if $\bar{\delta}'_{\oslash}(z) \leq f_{\max}$, or NaN, if $\bar{\delta}'_{\oslash}(z) > f_{\max}$. This concludes the proof.

□