IEEE754 precision-k base-β arithmetic inherited by precision-m base-β arithmetic for $k < m$

SIEGFRIED M. RUMP, Institute for Reliable Computing, Hamburg University of Technology,
and Visiting Professor at Waseda University, Faculty of Science and Engineering

Suppose an m-digit floating-point arithmetic in base $\beta \geq 2$ following the IEEE754 arithmetic standard is available. We show how a k-digit arithmetic with $k < m$ can be inherited solely using m-digit operations. This includes the rounding into k digits, the four basic operations and the square root, all for even or odd base β. In particular, we characterize the relation between k and m so that no double rounding occurs when computing in m digits and rounding the result into k digits. We discuss rounding to nearest as well as directed rounding, and our approach covers exceptional values including signed zero. For binary arithmetic, a Matlab toolbox based on binary64 including k-bit scalar, vector and matrix operations as well as k-bit interval arithmetic is part of Version 8 of INTLAB, the Matlab toolbox for reliable computing.

Categories and Subject Descriptors: G.1 [**Numerical Algorithms**]: Numerical Algorithms

General Terms: Computer arithmetic, Interval arithmetic

Additional Key Words and Phrases: Floating-point arithmetic, base-β, unit in the first place (ufp), IEEE754, double rounding, interval arithmetic, INTLAB

ACM Reference Format:
Siegfried M. Rump, 2013. IEEE754 precision-k base-β arithmetic inherited by precision-m base-β arithmetic for $k < m$. *ACM Trans. Math. Softw.* 0, 0, Article 00 (2013), 15 pages.
DOI:http://dx.doi.org/10.1145/0000000.0000000

1. PREVIOUS WORK

Suppose an arithmetic following the IEEE 754 standard in base $\beta \geq 2$ with precision m is available. Based on that, we describe in this note how to simulate an IEEE 754 precision-k arithmetic for $1 \leq k < m$. This covers in particular the rounding from precision-m into precision-k, the four basic operations and the square root. Moreover, the corresponding operations with directed rounding are covered.

Previous work includes Sipe [Lefèvre 2013], a set of C-subroutines for correctly rounded binary operations in low precision, and FLAP [Stewart 2009], a Matlab toolbox for correctly rounded decimal arithmetic. Neither library provides directed rounding. According to Pete Stewart [Stewart 2014] the latter is thoroughly tested, however, without any claim of rigor. Indeed, at least for precision larger than 8 decimals, examples of incorrect rounding to nearest can be found.

The present note aims on computing rigorously rounded results in a base $\beta \geq 2$ according to IEEE 754 including overflow, underflow and exceptional values such as ∞ and NaN, or signed zero, where rounding covers to-nearest and directed rounding.

The note is organized as follows. After introducing some notation, we describe a simple method to round a base-β precision-m number into base-β precision-k for $1 \leq k < m$

Siegfried M. Rump, Institute for Reliable Computing, Hamburg University of Technology, Schwarzenbergstraße 95, Hamburg 21071, Germany, rump@tuhh.de.

DOI:http://dx.doi.org/10.1145/0000000.0000000

using solely precision-m operations. Based on that we characterize in Section 4 the maximum value of k relative to m so that double-rounding cannot occur. This means that performing an operation with precision-k input in precision-m and round the result into precision-k produces the same result as if directly computed in precision-k. This has been done for binary arithmetic in [Figueroa 1995; Figueroa Del Cid 2000] barring overflow, underflow and exceptional values. We close the note with some remarks on implementation issues and directed rounding.

2. NOTATION

Let $1 \le k \in \mathbb{N}$ and $1 \le E \in \mathbb{N}^*$ be given where $\mathbb{N}^* := \mathbb{N} \cup \{\infty\}$. Denote by $\mathfrak{F}_{\beta,k,E}$ the set

$$\pm m_1.m_2m_3\dots m_k \cdot \beta^e \qquad \text{with } -E+1 \le e \le E \tag{1}$$

of precision-k floating-point numbers in base $\beta \ge 2$, and define

$$\mathbb{F}_{\beta,k,E} := \mathfrak{F}_{\beta,k,E} \cup \{\pm\infty\} \quad \text{and} \quad \mathbb{F}^*_{\beta,k,E} := \mathbb{F}_{\beta,k,E} \cup \{\text{NaN}\}. \tag{2}$$

Throughout this note all floating-point quantities refer to the same base β, but the precision and exponent range will vary. Thus we omit the index β for ease of notation.

The arithmetic operations shall follow the IEEE754 floating-point arithmetic standard [IEEE 2008]. The set of nonzero floating-point numbers with $m_1 = 0$ and $e = -E+1$ represents the underflow range. Exceptional values like $\pm\infty$ and NaN are defined and treated as in IEEE754. Note that always $\mathbb{F}_{k,E} = -\mathbb{F}_{k,E}$. A very thorough and readable introduction to all aspects of floating-point is [Muller et al. 2009].

IEEE754 binary32 (single precision) and binary64 (double precision) with $\beta = 2$ are characterized by $k = 24$, $E = 127$ and $k = 53$, $E = 1023$, respectively. Note that k refers to the precision in bits, not the stored number of bits.[1]

For example, in the exceptional case $k = 1$ the set $\mathbb{F}_{1,E}$ for $\beta = 2$ consists only of powers of 2, namely $\mathbb{F}_{1,E} = \{\pm 2^e : -E+1 \le e \le E\} \cup \{\pm 0, \pm\infty\}$, and there is no underflow range.

For base β and precision-k,

$$\mathbf{u}_k := \frac{1}{2}\beta^{1-k} \quad \text{denotes the relative rounding error unit.}$$

Note that $\mathbf{u}_k \in \mathbb{F}_{k,E}$ if and only if β is even. For $E \ne \infty$, the largest normalized, smallest normalized and smallest denormalized positive floating-point numbers in $\mathbb{F}_{k,E}$ are

$$\text{realmax}_{k,E} = \beta^{E+1}(1-\beta^{-k}),\ \text{realmin}_{k,E} = \beta^{-E+1},\ \text{subrealmin}_{k,E} = \beta^{-E+2-k}, \tag{3}$$

respectively. Hence

$$0 \ne f \in \mathfrak{F}_{k,E} \quad \text{normalized} \quad \Leftrightarrow \quad |f| \ge \text{realmin}_{k,E}.$$

Moreover,

$$f \in \mathfrak{F}_{k,E},\ \beta^e \le f \le \beta^{e+1},\ -E+1 \le e \le E \quad \Rightarrow \quad f = \beta^e(1 + n\beta^{1-k}) \ \text{with } n \in \mathbb{N}. \tag{4}$$

For the analysis of floating-point properties of algorithms, I introduced the concept “unit in the first place” (ufp) defined by

$$0 \ne x \in \mathbb{R}: \quad \text{ufp}_\beta(x) := \beta^{\lfloor \log_\beta |x| \rfloor} \quad \text{and} \quad \text{ufp}_\beta(0) := 0.$$

For nonzero x it is the weight of the leading digit in the base-β representation of x. We use ufp$(\cdot)$ because the base β is always clear from the context. This concept proved

[1] For normalized binary numbers the “implicit-one” [or “implicit-integer”] bit $m_1 = 1$ need not to be stored.

useful in formalizing proofs of certain properties related to floating-point operations, see [Rump 2009] and papers cited there. For base β it follows

$$0 \neq x \in \mathbb{R} \quad \Rightarrow \quad \mathbf{ufp}(x) \leq |x| < \beta\mathbf{ufp}(x) \tag{5}$$

$$f \in \mathbb{F}_{k,E} \quad \Rightarrow \quad \mathbf{ufp}(f) \leq |f| \leq \beta(1-\beta^{-k})\mathbf{ufp}(f) \tag{6}$$

$$f \in \mathbb{F}_{k,E} \quad \Rightarrow \quad f \in 2\mathbf{u}_k\mathbf{ufp}(f)\mathbb{Z} = \beta^{1-k}\mathbf{ufp}(f)\mathbb{Z} \tag{7}$$

and

$$x \in \beta^{\ell-k}\mathbb{Z},\ \mathbf{realmin}_{k,E} \leq |x| \leq \beta^{\ell},\ |x| < \beta^{E+1} \quad \Rightarrow \quad x \in \mathfrak{F}_{k,E}. \tag{8}$$

These properties are easily verified. The successor and predecessor of finite $x \in \mathbb{R}$ with $|x| \leq \mathbf{realmax}_{k,E}$ are defined by

$$\mathbf{succ}_{k,E}(x) := \inf\{f \in \mathbb{F}_{k,E} : x < f\} \quad \text{and} \quad \mathbf{pred}_{k,E}(x) := \sup\{f \in \mathbb{F}_{k,E} : f < x\},$$

respectively. For example, $\mathbf{succ}(\mathbf{realmax}_{k,E}) = \infty$ and

$$\mathbf{succ}(\beta^e) = (1+2\mathbf{u}_k)\beta^e = (1+\beta^{1-k})\beta^e \quad \text{and} \quad \mathbf{pred}(\beta^e) = (1-\beta^{-k})\beta^e \tag{9}$$

for $-E+2 \leq e \leq E$. The rounding to nearest $\mathbf{fl}_{k,E} : \mathbb{R} \to \mathbb{F}_{k,E}$ according to IEEE 754 is characterized by

$$x \in \mathbb{R},\ |x| \leq \mathbf{realmax}_{k,E} : \quad |\mathbf{fl}_{k,E}(x) - x| = \min\{|f-x| : f \in \mathbb{F}_{k,E}\}$$

with rounding ties to even in case x is equal to a "switching point", i.e. the midpoint of two adjacent elements in $\mathbb{F}_{k,E}$. For $|x| > \mathbf{realmax}_{k,E}$, the switching point is given by

$$|\mathbf{fl}_{k,E}(x)| = \infty \qquad \Leftrightarrow \qquad |x| \geq \beta^{E+1}(1 - \frac{1}{2}\beta^{-k}). \tag{10}$$

For convenience we use $\mathbb{F}_k$, $\mathbf{fl}_k$, $\mathbf{succ}_k$, etc. if $E = \infty$. In that case $\mathbf{fl}_k(x) = 0 \ \Leftrightarrow\ x = 0$. For infinite exponent range the set of switching points $\mathbb{S}_k$ for the rounding $\mathbf{fl}_k(\cdot)$ is characterized by

$$x \in \mathbb{S}_k \ :\Leftrightarrow\ |x| = f + \mathbf{u}_k\mathbf{ufp}(f) \quad \text{for some} \quad 0 < f \in \mathfrak{F}_k. \tag{11}$$

It is $\mathbb{S}_k \subseteq \mathbb{F}_{k+1}$ for even β, and $\mathbb{S}_k \cap \mathbb{F}_m = \emptyset$ for any m in case of odd β. For odd β, the rounded-to-nearest result is unique.

For $k = 1$ and $x \in \mathbb{S}_k$ we have $x = \pm(m_1 + \frac{1}{2})\beta^e$, and we interpret "rounding ties to even" as $\mathbf{fl}_k(x) = \pm m_1\beta^e$ for even m_1 and $\mathbf{fl}_k(x) = \pm(m_1+1)\beta^e$ otherwise. In any case

$$0 \leq x \in \mathbb{R},\ 1 \leq k \in \mathbb{N} \quad \Rightarrow \quad \mathbf{ufp}(x) \leq \mathbf{ufp}(\mathbf{fl}_k(x)). \tag{12}$$

The best error estimate for rounding to nearest into normalized numbers in $\mathbb{F}_{k,E}$ is

$$x \in \mathbb{R},\ f = \mathbf{fl}_{k,E}(x),\ \mathbf{realmin}_{k,E} \leq |f| < \infty \ \Rightarrow\ |f-x| \leq \mathbf{u}_k\mathbf{ufp}(x) \leq \mathbf{u}_k\mathbf{ufp}(f). \tag{13}$$

In the literature, e.g. [Higham 2002], often $|f-x| \leq \mathbf{u}_k|x|$ is used[2]. However, this is too weak for our purposes as the bound can be worse than (13) by almost a factor β depending on whether the mantissa is close to 1 or close to β.

Moreover,

$$x \in \mathbb{R}, f \in \mathbb{F}_k \quad \text{and} \quad |f-x| < \mathbf{u}_k\mathbf{ufp}(x) \quad \Rightarrow \quad f = \mathbf{fl}_k(x) \tag{14}$$

for infinite exponent range. Note that $\mathbf{ufp}(x)$ cannot be replaced by $\mathbf{ufp}(f)$ in (14) when f is a power of β.

[2] This can be improved into $|f-x| \leq \frac{\mathbf{u}_k}{1+\mathbf{u}_k}|x|$ as noted in [Knuth 1998], see also [Jeannerod and Rump 2014].

Directed rounding of a real number $x \in \mathbb{R}$ into $\mathbb{F}_{k,E}$ is depicted by $\mathbf{fl}^{\nabla}_{k,E}, \mathbf{fl}^{\Delta}_{k,E}, \mathbf{fl}^{><}_{k,E}$ and $\mathbf{fl}^{<>}_{k,E}$ for rounding downwards, rounding upwards, rounding towards and away from zero, respectively. Those are defined by

$$\mathbf{fl}^{\nabla}_{k,E}(x) := \max\{f \in \mathbb{F}_{k,E} : f \le x\},\ \mathbf{fl}^{\Delta}_{k,E}(x) := \min\{f \in \mathbb{F}_{k,E} : x \le f\} \quad \text{and}$$

$$\mathbf{fl}^{><}_{k,E}(x) := \begin{cases} \mathbf{fl}^{\nabla}_{k,E}(x) \text{ if } x \ge 0, \\ \mathbf{fl}^{\Delta}_{k,E}(x) \text{ otherwise }, \end{cases} \qquad \mathbf{fl}^{<>}_{k,E}(x) := \begin{cases} \mathbf{fl}^{\Delta}_{k,E}(x) \text{ if } x \ge 0, \\ \mathbf{fl}^{\nabla}_{k,E}(x) \text{ otherwise }, \end{cases}$$

respectively. It follows, for example, $\mathrm{succ}_{k,E}(f) = \mathbf{fl}^{\Delta}_{k,E}(f + \mathrm{subrealmin}_{k,E})$ for $f \in \mathfrak{F}_{k,E}$.

3. ROUNDING INTO K-BIT FORMAT

The rounded image $\mathbf{fl}_{\beta,k,E}(x)$ of $x \in \mathbb{R}$ is clear if there is direct access to the β-adic representation of x. For an m-digit working precision and a k-digit destination precision with $k < m$ we show how to compute $\mathbf{fl}_k(\cdot)$ solely using operations in working precision.

The problem is that switching points are rounded downwards or upwards, depending on how the tie is resolved. If the direction of rounding is always the same as for $\mathbf{fl}^{\nabla}, \mathbf{fl}^{\Delta}$ and $\mathbf{fl}^{><}$, then double rounding cannot occur, i.e. always $\mathbf{fl}_k(\mathbf{fl}_m(x)) = \mathbf{fl}_k(x)$.

In order to compute $\mathbf{fl}_k(x)$ using only operations in $\mathbb{F}_m$ we add a larger constant C basically cutting off all but the leading k digits of x, and then subtract C again. We present in Lemma 3.1 and Theorem 3.2 a general method for base $\beta \ge 2$, which is the basis for calculating $\mathbf{fl}_k(\cdot)$ with Algorithm 3. A suitable constant C is computed by Algorithms 1 and 2 for even and for odd base β, respectively. We first treat input with image in the normalized range.

LEMMA 3.1. *For base $\beta \ge 2$, let $k, m \in \mathbb{N}$ and $E, E' \in \mathbb{N}^*$ be given with $1 \le k < m$, $1 \le E$ and $E' \ge E + m - k$. For $x \in \mathbb{F}_{m,E'}$ define*

$$\underline{C} := \beta^{m-k}\mathrm{sign}(x)\mathrm{ufp}(x) \quad \textit{and} \quad C := d \cdot \underline{C} \ \textit{for fixed}\ d \in \{1, 2, \ldots, \beta - 1\}. \tag{15}$$

Then $\mathrm{realmin}_{k,E} \le |x| < \beta^{E+1}(1 - \frac{1}{2}\beta^{-k})$ *implies*

$$\mathbf{fl}_{k,E}(x) = \mathbf{fl}_{m,E'}(\mathbf{fl}_{m,E'}(C + x) - C). \tag{16}$$

PROOF. By the symmetry of $\mathbb{F}_{k,E}$ and $\mathbb{F}_{m,E'}$ we may assume without loss of generality that $x \ge 0$. Note that by (10) the assumption $|x| < \beta^{E+1}(1 - \frac{1}{2}\beta^{-k})$ is equivalent to $|\mathbf{fl}_{k,E}(x)| < \infty$. Then $\mathrm{ufp}(x) \le \beta^E$ gives

$$\mathrm{realmin}_{m,E'} \le \mathrm{realmin}_{k,E} \le \beta^{m-k} \cdot \mathrm{ufp}(x) = \underline{C} \le \beta^{m-k+E} \le \beta^{E'}$$

and $\mathrm{realmin}_{m,E'} \le \underline{C} \le C \le (\beta - 1)\beta^{E'}$. Since all quantities in the computation of C and $\underline{C}$ except the factor d are powers of β, it follows $\underline{C}, C \in \mathfrak{F}_{m,E'}$. Then

$$C + x < \beta^{m-k+E}\big(\beta - 1 + \beta^{1+k-m}(1 - \frac{1}{2}\beta^{-k})\big) \le \beta^{E'}(\beta - 1 + 1 - \frac{1}{2}\beta^{1-m}) = \beta^{E'+1}(1 - \frac{1}{2}\beta^{-m})$$

using $E' \ge E + m - k$. Hence (10) implies that $\mathbf{fl}_{m,E'}(C + x)$ does not cause overflow in $\mathbb{F}_{m,E'}$, nor does $F := \mathbf{fl}_{m,E'}(\mathbf{fl}_{m,E'}(C + x) - C)$. Moreover,

$$C + x \ge \underline{C} + \mathrm{ufp}(x) = (\beta^{m-k} + 1)\mathrm{ufp}(x) \in \mathbb{F}_{m,E'} \quad \Rightarrow \quad \mathbf{fl}_{m,E'}(C + x) \ge C + \mathrm{ufp}(x)$$

by the monotonicity of the rounding, so that $F \ge \mathrm{ufp}(x) \ge \mathrm{realmin}_{k,E}$. Thus $E' \ge E$ and $x \ge \mathrm{realmin}_{k,E}$ imply that neither the operations in (15) nor the rounding $\mathbf{fl}_{k,E}(x)$ can cause underflow or overflow. Thus the roundings $\mathbf{fl}_{k,E}$ and $\mathbf{fl}_{m,E'}$ in (16) can safely be replaced by $\mathbf{fl}_k$ and $\mathbf{fl}_m$, respectively, corresponding to an infinite exponent range.

Denote by $x = \underline{x} + e$ the unique splitting of x into $\underline{x} \in 2\mathbf{u}_m\underline{C}\mathbb{Z}$ and $0 \le e < 2\mathbf{u}_m\underline{C}$, and set $\overline{x} := \mathrm{succ}_k(\underline{x})$. We first show

$$\underline{x}, \overline{x} \in \mathbb{F}_k \subseteq \mathbb{F}_m \quad \text{and} \quad C + \underline{x}, C + \overline{x} \in \mathbb{F}_m. \tag{17}$$

We have $\mathrm{ufp}(\underline{x}) = \mathrm{ufp}(x)$, $\overline{x} = \underline{x} + 2\mathbf{u}_k\mathrm{ufp}(x) = \underline{x} + 2\mathbf{u}_m\underline{C}$, and also $\underline{x}, \overline{x} \in 2\mathbf{u}_k\mathrm{ufp}(x)\mathbb{Z}$ and $0 \le e < 2\mathbf{u}_k\mathrm{ufp}(x)$. Thus $\underline{x} \le \overline{x} \le \beta\mathrm{ufp}(x) \in \mathbb{F}_{m,E'}$ and (8) imply the first statement in (17). Furthermore, $C + \underline{x}, C + \overline{x} \in 2\mathbf{u}_m\underline{C}\mathbb{Z}$ and

$$C + \underline{x} \le C + \overline{x} \le (\beta - 1)\underline{C} + \beta\mathrm{ufp}(x) = (\beta - 1 + \beta^{k-m+1})\underline{C} \le \beta\underline{C},$$

and using again (8) implies the second statement in (17).

We distinguish three cases. First, assume $0 \le e < \mathbf{u}_m\underline{C}$. For $g := C + \underline{x} \in \mathbb{F}_m$ and using $\underline{C} = \mathrm{ufp}(C)$ it follows

$$C + x - g = x - \underline{x} = e < \mathbf{u}_m\underline{C} = \mathbf{u}_k\mathrm{ufp}(x) \le \mathbf{u}_m\mathrm{ufp}(g),$$

and (14) gives $\mathrm{fl}_m(C + x) = g = C + \underline{x}$ as well as $\mathrm{fl}_k(x) = \underline{x} = \mathrm{fl}_m(g - C) = F$.

Second, assume $\mathbf{u}_m\underline{C} < e < 2\mathbf{u}_m\underline{C}$. Then for $g := C + \overline{x} \in \mathbb{F}_m$ we have

$$|C + x - g| = \overline{x} - x = 2\mathbf{u}_m\underline{C} - e < \mathbf{u}_m\underline{C} = \mathbf{u}_k\mathrm{ufp}(x) \le \mathbf{u}_m\mathrm{ufp}(g),$$

and again (14) gives $\mathrm{fl}_m(C + x) = g = C + \overline{x}$ and $\mathrm{fl}_k(x) = \overline{x} = \mathrm{fl}_m(g - C) = F$.

Third and finally, assume $e = \mathbf{u}_m\underline{C}$. The definition of $\underline{x}$ implies that the m-th digit of $C+\underline{x} \in \mathbb{F}_m$ and the k-th digit of $\underline{x} \in \mathbb{F}_k$ are the same, so that in particular they are either both odd or both even. The same arguments as before show that $\mathrm{fl}_m(C+x) = C+\widetilde{x}$ and $\mathrm{fl}_k(x) = \widetilde{x}$ for either $\widetilde{x} = \underline{x}$ or $\widetilde{x} = \overline{x}$. In any case $F = \widetilde{x} = \mathrm{fl}_k(x)$, finishing the proof. □

THEOREM 3.2. *For base $\beta \ge 2$, let $k, m \in \mathbb{N}$ and $E, E' \in \mathbb{N}^*$ be given such that $1 \le k < m$, $1 \le E$ and $E' \ge E + m - 1$. For $x \in \mathbb{F}_{m,E'}$ define*

$$\underline{C} := \mathrm{float}\big[\beta^{m-k}\mathrm{sign}(x) \cdot \max\big(\mathrm{ufp}(x), \mathrm{realmin}_{k,E}\big)\big] \tag{18}$$

and

$$C := \mathrm{float}\big[d \cdot \underline{C}\big] \quad \textit{for fixed } \; d \in \{1, 2, \ldots, \beta - 1\}, \tag{19}$$

where $\mathrm{float}[\cdot]$ *is computed with all operations being floating-point operations in* $\mathbb{F}^*_{m,E'}$.

Then $|x| < \beta^{E+1}(1 - \frac{1}{2}\beta^{-k})$ implies

$$\mathrm{fl}_{k,E}(x) = \mathrm{fl}_{m,E'}(\mathrm{fl}_{m,E'}(C + x) - C). \tag{20}$$

REMARK. Note that $1 - \frac{1}{2}\beta^{-k} \notin \mathbb{F}_m$ for odd β and any $m \in \mathbb{N}$; using $x \in \mathbb{F}_{m,E'}$ the assumption $|x| < \beta^{E+1}(1 - \frac{1}{2}\beta^{-k})$ is equivalent to $|x| \le \beta^{E+1}(1 - \frac{\beta-1}{2}\beta^{-1-k}) \in \mathbb{F}_{m,E'}$.

PROOF. In Lemma 3.1 the result is proved if $\mathrm{realmin}_{k,E} \le |x|$, so we may assume without loss of generality that $0 \le x < \mathrm{realmin}_{k,E}$.

We distinguish three cases. First, suppose $\beta^{-E+2-k} \le x < \beta^{-E+1}$ such that x has a base-β representation

$$x = 0.0\ldots0m_1m_2\ldots m_{k-j}\ldots \cdot \beta^{-E+1} = m_1.m_2\ldots m_{k-j}\ldots \cdot \beta^{-E+1-j}$$

with $0 < j < k$. Then $\mathrm{fl}_{k,E}(x) = \mathrm{fl}_{k-j,E+j}(x)$. Setting $\tilde{k} := k - j$ and $\tilde{E} := E + j$ yields $1 \le \tilde{k} < m$, $\tilde{E} + m - k \le E + m - 1 \le E'$ and $x \ge \beta^{-E+1-j} = \mathrm{realmin}_{\tilde{k},\tilde{E}}$. Hence Lemma 3.1 is applicable and

$$f = \mathrm{fl}_{k,E}(x) = \mathrm{fl}_{\tilde{k},\tilde{E}}(x) = \mathrm{fl}_{m,E'}(\mathrm{fl}_{m,E'}(\tilde{C} + x) - \tilde{C}),$$

where

$$\tilde{C} = d \cdot \beta^{m-\tilde{k}}\mathrm{ufp}(x) = d \cdot \beta^{m-k+j} \cdot \beta^{-E+1-j} = d \cdot \beta^{m-k}\mathrm{realmin}_{k,E} = C.$$

Second, suppose $\frac{1}{2}\beta^{-E+2-k} < x < \beta^{-E+2-k}$. In that case $\mathrm{fl}_{k,E}(x) = \mathrm{subrealmin}_{k,E} = \beta^{-E+2-k} = 2\mathbf{u}_m\underline{C}$ for the corresponding $\underline{C}$ in (18). Defining $g := \mathrm{succ}_{m,E'}(C)$ and using $m \geq 2$ yields $\mathrm{ufp}(C) = \mathrm{ufp}(\underline{C}) = \mathrm{ufp}(g)$, and therefore $g = C + 2\mathbf{u}_m\underline{C}$ by (9). Then $|g - (C+x)| = 2\mathbf{u}_m\underline{C} - x < \mathbf{u}_m\underline{C} = \mathbf{u}_m\mathrm{ufp}(g)$, such that (14) implies $\mathrm{fl}_{m,E'}(C+x) = g$ and $F = \mathrm{fl}_{m,E'}(g - C) = 2\mathbf{u}_m\underline{C} = \mathrm{fl}_{k,E}(x)$.

In the remaining third case $0 \leq x \leq \frac{1}{2}\beta^{-E+2-k} = \mathbf{u}_m\underline{C}$ for $\underline{C}$ as in (18). Then $|C - (C+x)| = x \leq \mathbf{u}_m\underline{C} = \mathbf{u}_m\mathrm{ufp}(C)$, such that $m \geq 2$ and rounding tie to even implies $\mathrm{fl}_{m,E'}(C+x) = C$ and $F = \mathrm{fl}_{m,E'}(C - C) = 0 = \mathrm{fl}_{k,E}(x)$. This finishes the proof. □

To apply Theorem 3.2 we calculate a valid offset-constant C according to (18) and (19) in base-β precision-m arithmetic. For even base β this is done by[3] Algorithm 1, and for odd base Algorithm 2. Based on that, Algorithm 3 computes the rounding into precision-k.

ALGORITHM 1: Constant C for precision `m` and even base `beta`

```
function C = constC(p)
   f = 1 - 0.5*beta^(1-m)
   C = beta^(m-1)*(p/f - p)
```

LEMMA 3.3. *For even base $\beta \geq 2$, let $p \in \mathfrak{F}_{\beta,m,E'}$ with $1 \leq E \leq E' - m + 1$ and $m \geq 2$ be given, and suppose* $\mathrm{realmin}_{m,E} \leq |p| < \beta^{E+1}$.

Then Algorithm 1, executed in $\mathbb{F}^*_{\beta,m,E'}$ produces $C \in \mathbb{F}_{m,E}$ with $C = d \cdot \mathrm{sign}(p) \cdot \mathrm{ufp}(p)$ and $d \in \{1, 2, \ldots, \beta/2\}$.

Moreover, $C \leq \mathrm{realmin}_{m,E}$ when $|p| < \mathrm{realmin}_{m,E}$.

REMARK. Note that $C = \mathrm{sign}(p) \cdot \mathrm{ufp}(p)$ for $\beta = 2$.

PROOF. To prove the first assertion, note that $E' \geq 2$ and even β yield $0.5 \cdot \beta^{1-m} = \frac{\beta}{2}\beta^{-m} \geq \frac{\beta}{2} \cdot \mathrm{subrealmin}_{m,E'}$, and thus $f = 1 - \mathbf{u}_m = 1 - \frac{\beta}{2}\beta^{-m} \in \mathbb{F}_{m,E'}$. We define

$$q := p/f, \;\; \tilde{q} := \mathrm{fl}(q), \;\; s := \tilde{q} - p \quad \text{and} \quad \tilde{s} := \mathrm{fl}(s).$$

Because p is in the normalized range, (9) implies

$$\mathrm{succ}(p) = p + 2\mathbf{u}_m\mathrm{ufp}(p) \quad \text{and} \quad \mathrm{ufp}(p) \leq p \leq \beta(1 - \beta^{-m})\mathrm{ufp}(p).$$

Then

$$(p + \mathbf{u}_m\mathrm{ufp}(p))f = p + \mathbf{u}_m\mathrm{ufp}(p) - \mathbf{u}_m(p + \mathbf{u}_m\mathrm{ufp}(p)) < p$$

gives $p + \mathbf{u}_m\mathrm{ufp}(p) < p/f = q$ and therefore $\mathrm{succ}(p) \leq \tilde{q}$. Moreover, $m \geq 2$ shows

$$\beta(1 + \mathbf{u}_m)f = \beta(1 - \frac{1}{4}\beta^{2-2m}) > \beta(1 - \beta^{-m}) \geq \frac{p}{\mathrm{ufp}(p)},$$

so that $q = p/f < \beta(1 + \mathbf{u}_m)\mathrm{ufp}(p)$. Since $\beta\mathrm{ufp}(p) \leq \beta^{E+1} \in \mathfrak{F}_{m,E'}$ by $p < \beta^{E+1}$ we obtain $\mathrm{succ}(p) \leq \tilde{q} \leq \beta\mathrm{ufp}(p)$, and (4) yields $\tilde{q} = p + \alpha\beta^{1-m}\mathrm{ufp}(p)$ for some $1 \leq \alpha \in \mathbb{N}$. Now (6) implies

$$\begin{aligned} p \;&\leq\; p(1 + \tfrac{1}{2}\beta^{-m} - \tfrac{1}{2}\beta^{2-2m}) < p(1 - \tfrac{1}{4}\beta^{2-2m} + \tfrac{1}{2}\beta^{-m} - \tfrac{1}{4}\beta^{1-2m}) \\ &=\; p(1 - \tfrac{1}{2}\beta^{1-m})(1 + \tfrac{1}{2}\beta^{1-m} + \tfrac{1}{2}\beta^{-m}) < (1 - \mathbf{u}_m)p\left(1 + \tfrac{\beta+1}{2}\tfrac{\beta^{1-m}}{\beta(1-\beta^{-m})}\right) \\ &\leq\; f\left(p + \tfrac{\beta+1}{2}\beta^{1-m}\mathrm{ufp}(p)\right), \end{aligned}$$

[3] Originally I used `q=p/f` and `C=beta^(m-1)*(q-f*q)`; this simpler version was suggested by Marko Lange.

so that β even yields

$$q = p/f < p + \frac{\beta+1}{2}\beta^{1-m}\mathbf{ufp}(p) \qquad \text{and} \qquad \tilde{q} = \mathbf{fl}(p/f) \le p + \frac{\beta}{2}\beta^{1-m}\mathbf{ufp}(p).$$

Hence $\tilde{q} = p + d\beta^{1-m}\mathbf{ufp}(p)$ for an integer d with $1 \le d \le \beta/2$, and $s = d\beta^{1-m}\mathbf{ufp}(p) \in \mathbb{F}_{m,E}$ implies $\tilde{s} = s$. Finally $E' \ge m-1$ yields $\beta^{m-1} \in \mathbb{F}_{m,E'}$, and the result follows.

To prove the final assertion, we may assume without loss of generality that $0 < p < \text{realmin}_{m,E}$. Then (3) and (9) give

$$p \le q = p/f < \frac{1-\beta^{-m}}{1-\frac{1}{2}\beta^{1-m}}\,\beta^{-E+1} \le \beta^{-E+1} = \text{realmin}_{m,E},$$

so that $p \le \tilde{q} \le \text{realmin}_{m,E}$ and also $\tilde{q} - p \le \text{realmin}_{m,E}$. This finishes the proof. □

ALGORITHM 2: Constant C for precision `m` and odd base `beta`

```
function C = constC(p)
   f = (beta+1)/2*beta^(-m)
   C = beta^(m-1) * ( (p + f*p) - p )
```

LEMMA 3.4. *For odd base $\beta \ge 3$, let $p \in \mathfrak{F}_{\beta,m,E'}$ with $1 \le E \le E'-m+1$ and $m \ge 2$ be given, and suppose* $\text{realmin}_{m,E} \le |p| < \beta^{E+1}$.

Then Algorithm 2, executed in $\mathbb{F}^*_{\beta,m,E'}$ produces $C \in \mathbb{F}_{m,E}$ with $C = d \cdot \text{sign}(p) \cdot \mathbf{ufp}(p)$ and $d \in \{1, 2, \ldots, (\beta+1)/2\}$.

Moreover, $C \le \text{realmin}_{m,E}$ when $|p| < \text{realmin}_{m,E}$.

PROOF. To prove the first assertion we have $f = \frac{\beta+1}{2}\beta^{-m} \in \mathbb{F}_{m,E'}$ by $E' \ge 2$. Note that $1/2 \notin \mathbb{F}_m$ for any $m \ge 1$. We define

$$q := f \cdot p,\ \tilde{q} := \mathbf{fl}(q),\ r := p + \tilde{q},\ \tilde{r} := \mathbf{fl}(r),\ s := \tilde{r} - p \quad \text{and} \quad \tilde{s} := \mathbf{fl}(s).$$

Because p is in the normalized range, (9) implies

$$\text{succ}(p) = p + 2\mathbf{u}_m\mathbf{ufp}(p) \quad \text{and} \quad \mathbf{ufp}(p) \le p \le \beta(1-\beta^{-m})\mathbf{ufp}(p).$$

Then

$$\mathbf{u}_m\mathbf{ufp}(p) < B_1 := \frac{\beta+1}{2}\beta^{-m}\mathbf{ufp}(p) \le f \cdot p = q < \frac{\beta+1}{2}\beta^{1-m}\mathbf{ufp}(p) =: B_2$$

and $B_1, B_2 \in \mathbb{F}_m$ imply $B_1 \le \tilde{q} \le B_2$, so that

$$\mathbf{u}_m\mathbf{ufp}(p) < \tilde{q} \le \frac{\beta+1}{2}\beta^{1-m}\mathbf{ufp}(p). \tag{21}$$

Furthermore,

$$\begin{aligned}
\tfrac{p+\mathbf{succ}_m(p)}{2} \;&=\; p + \mathbf{u}_m\mathbf{ufp}(p) < \\
p + \tilde{q} \;&\le\; \big(\beta(1-\beta^{-m}) + \tfrac{\beta+1}{2}\beta^{1-m}\big)\mathbf{ufp}(p) = \big(\beta + \tfrac{\beta-1}{2}\beta^{1-m}\big)\mathbf{ufp}(p) \\
&<\; \beta(1+\mathbf{u}_m)\mathbf{ufp}(p) = \tfrac{\beta+\mathbf{succ}(\beta)}{2}\mathbf{ufp}(p)
\end{aligned}$$

and therefore

$$\text{succ}_m(p) \le \mathbf{fl}_m(p+\tilde{q}) = \tilde{r} \le \beta\mathbf{ufp}(p) \le \beta^{E+1} \in \mathbb{F}_{m,E'}.$$

Similar to the proof of Lemma 3.3 we use (4) to conclude $\tilde{r} = p + d\beta^{1-m}\mathrm{ufp}(p)$ for some $1 \leq d \in \mathbb{N}$, and (21) yields $1 \leq d \leq (\beta+1)/2$. Therefore $s = d\beta^{1-m}\mathrm{ufp}(p) = \tilde{s} \in \mathbb{F}_{m,E'}$, and using $\beta^{m-1} \in \mathbb{F}_{m,E'}$ by $E' \geq m-1$ finishes this part of the proof.

To prove the final assertion, we may assume without loss of generality that $0 < p < \mathrm{realmin}_{m,E}$. Then $\mathrm{ufp}(p) \leq \beta^{-E}$ and

$$q = f \cdot p < \frac{\beta+1}{2}\beta^{1-m}\mathrm{ufp}(p) < \beta^{2-m-E} = \mathrm{subrealmin}_{m,E} =: \sigma \in \mathfrak{F}_{m,E'},$$

so that $\tilde{q} \leq \sigma$. It follows $\tilde{r} \leq \mathrm{succ}_{m,E'}(p) \leq p + \sigma$ and $\tilde{s} \leq \sigma$, so that $\beta^{m-1}\tilde{s} \leq \beta^{1-E} = \mathrm{realmin}_{m,E}$ finishes the proof. □

ALGORITHM 3: Rounding of $d \in \mathbb{F}^*_{\beta,m,E'}$ into $\mathbb{F}^*_{\beta,k,E}$ for $1 \leq k < m$, $1 \leq E \leq E' - m + 1$ and $\beta \in \mathbb{N}$; all operations in $\mathbb{F}_{\beta,m,E'}$. Here `Sovfl` $= \beta^{E+1}(1 - \alpha\beta^{-1-k}) \in \mathfrak{F}_{m,E'}$ with $\alpha := \frac{\beta}{2}$ if β is even, and $\alpha := \frac{\beta+1}{2}$ if β is odd.

```
function f = flround(d,k,E)
   if abs(d) >= Sovfl
      d = sign(d)*inf;
   else
      C = sign(d) * max(constC(d),realmin(k,E)) * beta^(m-k);
      f = ( C + d ) - C;
   end
   if f == 0
      f = 0 * d;                        % preserve signed zero
   end
```

THEOREM 3.5. *For even base $\beta \geq 2$, let $k, m \in \mathbb{N}$ and $E, E' \in \mathbb{N}$ be given with $1 \leq k < m$, $1 \leq E \leq E' - m + 1$. Then for $d \in \mathbb{F}^*_{\beta,m,E'}$ the result f of Algorithm 3 satisfies $f = \mathrm{fl}_{\beta,k,E}(d)$. This includes overflow and underflow, NaN, infinity and signed zero.*

REMARK 1. Note that `constC(d)` refers to Algorithm 1 for even, and to Algorithm 2 for odd base β.

REMARK 2. With a little effort, the restriction $E' \geq E + m - 1$ can be relaxed into $E' \geq E + m - k$. For that in particular the multiplication by β^{m-1} in Algorithms 1 and 2 is replaced by a division by β^{1-m} to ensure the constant is in $\mathbb{F}_{m,E'}$. In any case even the same exponent range $E' = E$ can be covered, at some cost, by a suitable scaling.

PROOF. The assertion is true for the special cases $d \in \{\pm\infty, \mathrm{NaN}\}$, and by (10) also for $|d| \geq \beta^{E+1}(1 - \alpha\beta^{-k}) =$ `Sovfl` $\in \mathbb{F}_{m,E'}$ and for $d = 0$. The latter includes a signed zero.

As before, we may assume without loss of generality that $d \geq 0$. Denote by $\tilde{C} \in \mathbb{F}_{m,E'}$ the computed value of `constC(d)`. If $0 < d < \mathrm{realmin}_{k,E}$, then Lemma 3.3 and Lemma 3.4 imply that $\tilde{C} \leq \mathrm{realmin}_{k,E}$. Hence for $0 < d < \mathrm{realmin}_{k,E}$ and for $\mathrm{realmin}_{k,E} \leq d < \mathrm{realmax}_{k,E}$ the computed `C` satisfies the assumptions of Theorem 3.2. This implies (20) and proves the theorem. □

4. OPERATIONS IN K-BIT FORMAT

For a given base β, we will show in the following that $\mathrm{fl}_k(a \text{ op } b) = \mathrm{fl}_k(\mathrm{fl}_m(a \text{ op } b))$ for all $a, b \in \mathbb{F}_k$ and large enough m. The only exception to that statement is division in case of odd base β. In particular, we will characterize the minimal values of m. It will turn out that $m = 2k + 1$ is sufficient for the four basic arithmetic operations and, with one tiny exception if $\beta = 2$, also for the square root.

For the IEEE 754 binary64 format it follows that for k up to 26 and suitable restriction of the exponent range floating-point operations can be executed in double precision and then rounded to k bits by Algorithm 3 to obtain the correctly rounded IEEE754 k-bit result. Most of those results can be found for binary in [Figueroa 1995; Figueroa Del Cid 2000] barring overflow, underflow and exceptional values. Here, however, we treat a general base $\beta \geq 2$.

We formulate the results first for infinite exponent range. Later we identify the exponent range to ensure

$$\mathrm{fl}_k(\mathrm{fl}_m(a \text{ op } b)) = \mathrm{fl}_k(a \text{ op } b). \tag{22}$$

For better readability we write $\mathbb{F}_k, \mathrm{fl}_k$ rather than $\mathbb{F}_{\beta,k}, \mathrm{fl}_{\beta,k}$ etc. omitting the index β.

LEMMA 4.1. *Let fixed but arbitrary $k, m \in \mathbb{N}$ with $1 \leq k < m$ and $x \in \mathbb{R}$ be given. If $\mathrm{fl}_m(x) \notin \mathbb{S}_k$ for even base β, then $\mathrm{fl}_k(\mathrm{fl}_m(x)) = \mathrm{fl}_k(x)$.*

If $x \notin \mathbb{S}_k$ for odd base β, then $\mathrm{fl}_k(\mathrm{fl}_m(x)) = \mathrm{fl}_k(x)$. If $x := \frac{1}{2}(t+\mathrm{succ}_k(t))$ when $\beta = 4p+1$ and $x := \frac{1}{2}(1+t)$ when $\beta = 4p+3$ for $t := \mathrm{succ}_k(1)$ and $p \in \mathbb{N}$, then $\mathrm{fl}_k(\mathrm{fl}_m(x)) \neq \mathrm{fl}_k(x)$.

PROOF. By the symmetry of the rounding we may assume without loss of generality that $x > 0$. The set of switching points $\mathbb{S}_k$ is characterized by (11).

For even base β define $g := \mathrm{fl}_m(x)$. If $g = \mathrm{fl}_k(g)$, then $g \in \mathbb{F}_k \subset \mathbb{F}_m$ and rounding to nearest imply $g = \mathrm{fl}_k(x)$. Suppose $\mathrm{fl}_k(g) < g$. Then there is a unique switching point $s \in \mathbb{S}_k$ with $\mathrm{fl}_k(g) < g \leq s$ minimizing $|\mathrm{fl}_k(g) - s|$. Since $g = \mathrm{fl}_m(x) \notin \mathbb{S}_k$ it follows $g \neq s$, so that $\mathrm{fl}_k(g) < g < s \in \mathbb{S}_k \subset \mathbb{F}_{k+1} \subseteq \mathbb{F}_m$ and $\mathrm{fl}_k(g), g, s \in \mathbb{F}_m$. Hence $\mathrm{fl}_k(g) = \mathrm{fl}_k(x)$; the case $\mathrm{fl}_k(g) > g$ is handled similarly.

For odd base β if follows $\mathbb{S}_k \subset \mathbb{S}_m$ and $\mathbb{F}_p \cap \mathbb{S}_q = \emptyset$ for any $p, q \in \mathbb{N}$. If $x \notin \mathbb{S}_k$, then there is a unique switching point $s \in \mathbb{S}_k$ minimizing $|\mathrm{fl}_k(x) - s|$. If $\mathrm{fl}_k(x) < s$, then $\mathrm{fl}_k(x) \leq x < s$. Thus $s \in \mathbb{S}_m$ and $\mathbb{F}_k \subset \mathbb{F}_m$ imply $\mathrm{fl}_k(x) \leq \mathrm{fl}_m(x) < s$, and therefore $\mathrm{fl}_k(x) = \mathrm{fl}_k(\mathrm{fl}_m(x))$. The case $\mathrm{fl}_k(x) > s$ is treated similarly.

Finally, for $\beta := 2b+1$ and odd b, we have $t = 1 + \beta^{1-k}$ and $\frac{1}{2} = (0.bb\bar{b})_\beta$, so that

$$x = \frac{1}{2}(1+t) = 1 + b\sum_{i=k}^{\infty} \beta^{-i} \quad \text{and} \quad y := 1 + b\sum_{i=k}^{m-1} \beta^{-i} = 1 + \frac{1}{2}\left(\beta^{1-k} - \beta^{1-m}\right) \in \mathbb{F}_m$$

yield $x = \frac{1}{2}(y + \mathrm{succ}_m(y))$. Since b is odd, rounding tie-to-even implies $\mathrm{fl}_m(x) = \mathrm{succ}_m(y) > x$, so that $\mathrm{fl}_k(\mathrm{fl}_m(x)) = \mathrm{succ}_k(1) \neq 1 = \mathrm{fl}_k(x)$. The case $\beta = 2b+1$ for even b follows similarly, and this completes the proof. □

This suffices to completely characterize the situation concerning (22) for odd base β; for even base it will be more involved.

LEMMA 4.2. *For odd base $\beta \geq 3$ let fixed but arbitrary $k, m \in \mathbb{N}$ with $1 \leq k < m$ and $a, b \in \mathbb{F}_k$ be given. Then*

$$\mathrm{fl}_k(\mathrm{fl}_m(x)) = \mathrm{fl}_k(x) \qquad \textit{for} \quad x \in \{a+b, a-b, a\cdot b, \sqrt{|a|}\}, \tag{23}$$

and, abbreviating $t := \mathrm{succ}_k(1)$,

$$\mathrm{fl}_k(\mathrm{fl}_m(a/b)) \neq \mathrm{fl}_k(a/b) \qquad \textit{for} \quad \begin{cases} \beta = 4p+1, a := t + \mathrm{succ}_k(t),\ b := 2 \\ \beta = 4p+3, a := 1+t,\ b := 2. \end{cases} \tag{24}$$

PROOF. Concerning the first statement we have to show $x \notin \mathbb{S}_k$. Without loss of generality assume that $a, b > 0$. Then $a, b \in \mathbb{F}_k$ implies $x \in \mathbb{F}_n$ for $x \in \{a+b, a-b, a\cdot b\}$

and suitably large $n \in \mathbb{N}$. But $S_p \cap \mathbb{F}_q = \emptyset$ for all $p, q \in \mathbb{N}$ in case of odd β, so that $x \notin \mathbb{S}_k$ and Lemma 4.1 imply $\mathrm{fl}_k(\mathrm{fl}_m(x)) = \mathrm{fl}_k(x)$.

For the square root we note

$$0 \neq a \in \mathbb{F}_k \qquad \Leftrightarrow \qquad a = A\beta^e \quad \text{with} \quad A, e \in \mathbb{Z},\ \beta^{k-1} \leq |A| \leq \beta^k - 1.$$

Hence (11) implies that $s \in \mathbb{S}_k$ is equivalent to $s = (S + \frac{1}{2})\beta^e$ with $S \in \mathbb{N}$ and $\beta^{k-1} \leq S \leq \beta^k - 1$. Thus $s^2 = (S^2 + S + \frac{1}{4})\beta^{2e}$ cannot be in any $F_n, n \in \mathbb{N}$ because 4 does not divide any power of β for odd β. This concludes the proof of (23).

For division $a, b \in \mathbb{F}_k$ as defined in (24) and Lemma 4.1 finishes the proof. □

For odd base β we proved that, on the one hand, for addition, subtraction, multiplication and square root double rounding never occurs for any $m \geq k \geq 1$. On the other hand, cases were identified where for division double rounding is unavoidable for any $m > k$.

The problem with division a/b can be resolved by computing, in precision-m, another quantity q such that $\mathrm{fl}_k(q) = \mathrm{fl}_k(a/b)$. This method was suggested by Marko Lange.

LEMMA 4.3. *For even or odd base* $\beta \geq 2$ *let* $1 \leq k, m \in \mathbb{N}$ *and* $a, b \in \mathbb{F}_k$ *be given,* $a, b \neq 0$. *Denote* $a = s_a A \cdot \mathrm{ufp}(a)$ *and* $b = s_b B \cdot \mathrm{ufp}(b)$ *with* $s_a, s_b \in \{-1, 1\}$ *and* $A, B \in \mathbb{F}_k$, $1 \leq A, B < \beta$. *Define*

$$C := \beta^e B + A \qquad \textit{with} \qquad e := \begin{cases} m - k & \textit{if } A \geq B \\ m - k - 1 & \textit{otherwise} . \end{cases}$$

If $m \geq 2k$, *then*

$$\big(\mathrm{fl}_m(C/B) - \beta^e\big)\frac{s_a \mathrm{ufp}(a)}{s_b \mathrm{ufp}(b)} = fl_k(a/b).$$

REMARK. It will be clear from the proof that with some case distinctions the weaker assumption $m \geq k + 1$ suffices to derive a similar method using a different exponent e. Moreover, the cases $a = 0$ and/or $b = 0$ are trivial.

PROOF. The assumptions imply $1 \leq A, B \leq \beta(1 - \beta^{-k})$ and $1 \leq \beta^{m-2k}$. Furthermore, $C = \beta^e B + A \in \beta^{1-k}\mathbb{Z}$ because $0 \leq m - k - 1 \leq e$, so that

$$C \leq (\beta^e + 1)\beta(1 - \beta^{-k}) \leq \beta^{m-k}(1 + \beta^{-k})\beta(1 - \beta^{-k}) < \beta^{m+1-k}$$

together with (8) yields $C \in \mathbb{F}_m$ and $\mathrm{ufp}(C) \leq \beta^{m-k}$.

First, assume $A \geq B$, so that $1 \leq A/B < \beta$ and $\mathrm{ufp}(A/B) = 1$. Then $C/B = \beta^{m-k} + A/B$ and $\mathrm{ufp}(A/B) = 1$ imply that the last k digits of $Q := \mathrm{fl}_m(C/B)$ are equal to the mantissa digits of $\mathrm{fl}_k(A/B)$. Hence $fl_m(Q - \beta^{m-k})$ causes no rounding error and the result follows.

Second, assume $A < B$ so that $\beta^{-1} \leq A/B < 1$. Then $C/B = \beta^{m-k-1} + A/B$ and $\mathrm{ufp}(A/B) = \beta^{-1}$ imply that again the last k digits of $Q := \mathrm{fl}_m(C/B)$ are equal to the mantissa digits of $\mathrm{fl}_k(A/B)$. This finishes the proof. □

To treat an even base β we will, besides (5). . .(8), frequently use

$$p, q \in \mathbb{Z} : \qquad p \geq q \quad \Rightarrow \quad \beta^p\mathbb{Z} \subseteq \beta^q\mathbb{Z}. \tag{25}$$

LEMMA 4.4. *For even base* $\beta \geq 2$ *let* $1 \leq k, m \in \mathbb{N}$ *and* $a, b \in \mathbb{F}_k$ *be given. If* $m \geq 2k+1$ *for* $\beta = 2$ *and* $m \geq 2k$ *for* $\beta \geq 4$, *then*

$$\mathrm{fl}_k(\mathrm{fl}_m(a + b)) = \mathrm{fl}_k(a + b).$$

Both lower bounds on m *cannot be improved.*

PROOF. Without loss of generality assume that $a \ge |b|$. Suppose $|b| \ge \beta^e \mathbf{ufp}(a)$ for $e \in \mathbb{Z}$, then $\mathbf{ufp}(b) \ge \beta^e \mathbf{ufp}(a)$ and

$$a, b, a+b \in \beta^{1-k}\mathbf{ufp}(b)\mathbb{Z} \subseteq \beta^{1-k+e}\mathbf{ufp}(a)\mathbb{Z}$$

by (7) and (25), and therefore

$$|b| \ge \beta^e \mathbf{ufp}(a) \quad \Rightarrow \quad a, b, a+b \in \begin{cases} \beta^{2+k-m+e}\mathbf{ufp}(a)\mathbb{Z} & \text{if } m \ge 2k+1 \\ \beta^{1+k-m+e}\mathbf{ufp}(a)\mathbb{Z} & \text{if } m \ge 2k. \end{cases} \tag{26}$$

We distinguish several cases, where the first two cases are treated together for $\beta = 2$ and for $\beta \ge 4$ by assuming only $m \ge 2k$ and $\beta \ge 2$.

First, suppose $|b| \ge \beta^{-k+1}\mathbf{ufp}(a)$, then $a, b, a+b \in \beta^{2-m}\mathbf{ufp}(a)\mathbb{Z}$ by (25). Hence $|a+b| \le 2|a| < 2\beta\mathbf{ufp}(a) \le \beta^2\mathbf{ufp}(a)$ and (8) prove $a+b \in \mathbb{F}_m$ and therefore $\mathbf{fl}_m(a+b) = a+b$.

Second, suppose $\beta^{-k}\mathbf{ufp}(a) \le |b| < \beta^{-k+1}\mathbf{ufp}(a)$ so that $a, b, a+b \in \beta^{1-m}\mathbf{ufp}(a)\mathbb{Z}$ and

$$|b| < \beta^{-k+1}\mathbf{ufp}(a) \quad \Rightarrow \quad |a+b| < \left(\beta(1-\beta^{-k}) + \beta^{-k+1}\right)\mathbf{ufp}(a) = \beta\mathbf{ufp}(a) \tag{27}$$

by (6), so that again (8) proves $\mathbf{fl}_m(a+b) = a+b$.

Now we need a case distinction on β. First, assume $\beta \ge 4$ and the remaining case $|b| < \beta^{-k}\mathbf{ufp}(a)$. Then $|b| < \mathbf{u}_k\mathbf{ufp}(a)$, and (14) yields $\mathbf{fl}_k(a+b) = a$. By (27) we get $\mathbf{ufp}(a+b) \le \mathbf{ufp}(a)$, and (13), $\beta \ge 4$, $m \ge 2k$ and $k \ge 1$ give

$$\begin{aligned} |\mathbf{fl}_m(a+b) - a| &< \mathbf{u}_m\mathbf{ufp}(a+b) + \beta^{-k}\mathbf{ufp}(a) < (\beta^{1-m} + \beta^{-k})\mathbf{ufp}(a) \\ &\le (\beta^{-k} + \beta^{-1})\beta^{1-k}\mathbf{ufp}(a) \le \mathbf{u}_k\mathbf{ufp}(a). \end{aligned}$$

Hence (14) proves $\mathbf{fl}_m(a+b) = a = fl_k(a+b)$.

It remains the case $\beta = 2$ and $|b| < \beta^{-k}\mathbf{ufp}(a)$. For that we need a final case distinction. First, assume $\beta^{-k-1}\mathbf{ufp}(a) \le |b| < \beta^{-k}\mathbf{ufp}(a)$, then $a, b, a+b \in \beta^{1-m}\mathbf{ufp}(a)\mathbb{Z}$ by (26). Then (6) gives

$$|a+b| < [\beta(1-\beta^{-k}) + \beta^{-k}]\mathbf{ufp}(a) < \beta\mathbf{ufp}(a),$$

and again $\mathbf{fl}_m(a+b) = a+b$ by (8). It remains the final case $\beta = 2$ and $|b| < \beta^{-k-1}\mathbf{ufp}(a)$. In that case $|b| < \mathbf{u}_k\mathbf{ufp}(a)$ and (14) give $\mathbf{fl}_k(a+b) = a$. Now (13), (9), $\mathbf{ufp}(a) \le \beta\mathbf{ufp}(a+b)$ and (12) yield

$$\begin{aligned} |\mathbf{fl}_m(a+b) - a| &\le |\mathbf{fl}_m(a+b) - (a+b)| + |b| \le \mathbf{u}_m\mathbf{ufp}(a+b) + \mathbf{pred}_k(\beta^{-k-1}\mathbf{ufp}(a)) \\ &\le \tfrac{1}{2}\beta^{-2k}\mathbf{ufp}(a+b) + (1-\beta^{-k})\beta^{-k-1}\mathbf{ufp}(a) \\ &\le \left(\tfrac{1}{2}\beta^{-2k} + (1-\beta^{-k})\beta^{-k}\right)\mathbf{ufp}(a+b) \\ &< \beta^{-k}\mathbf{ufp}(a+b) \le \mathbf{u}_k\mathbf{ufp}(\mathbf{fl}_m(a+b)). \end{aligned}$$

Thus $a \in \mathbb{F}_k$ and (14) imply $\mathbf{fl}_k(\mathbf{fl}_m(a+b)) = a = \mathbf{fl}_k(a+b)$.

For $\beta = 2$, $k = 2$, $m = 2k$, $a = 24 = (11000)_2$ and $b = 3 = (11)_2$ it is $\mathbf{fl}_k(a+b) = a$, but $\mathbf{fl}_k(\mathbf{fl}_m(a+b)) = \mathbf{fl}_2((11100)_2) = (100000)_2 = 32$.

For $\beta = 2\xi \ge 4$, $k = 2$, $m = 2k-1$, $a = \beta^3 = (1000)_\beta$ and $b = \xi\beta + 1 = (\xi 1)_\beta$ it is $\mathbf{fl}_k(a+b) = (1100)_\beta$, but $\mathbf{fl}_k(\mathbf{fl}_m(a+b)) = \mathbf{fl}_k((10\xi 0)_\beta) = (1000)_\beta$. Examples for other values of β and k are easily constructed. □

LEMMA 4.5. *For even base $\beta \ge 2$ let $1 \le k \in \mathbb{N}$ and $a, b \in \mathbb{F}_k$ be given. If $2k \le m \in \mathbb{N}$, then*

$$\mathbf{fl}_k(\mathbf{fl}_m(a \cdot b)) = \mathbf{fl}_k(a \cdot b).$$

The statement is not true for $m = 2k-1$.

PROOF. Obviously, $\mathrm{fl}_m(ab) = ab$ for $m \geq 2k$, so there is nothing to prove. For $k = 4$, $m = 2k-1$ and $a = b = 13 = (1101)_2$ it is $ab = 169 = (10101001)_2$, $fl_4(ab) = 176$, but $\mathrm{fl}_4(\mathrm{fl}_7(ab)) = \mathrm{fl}_4(168) = 160$. For base $\beta = 2$ this is, up to scaling, the only counterexample for $k = 4$; there is no counterexample for $k \leq 3$. For base $\beta = 10$ and $k = 2$ define $a = 14$ and $b = 82$. Then $ab = 1148$ and $\mathrm{fl}_2(ab) = 1100$, whereas $\mathrm{fl}_2(\mathrm{fl}_3(ab)) = \mathrm{fl}_2(1150) = 1200$. Examples for other values of β and k are easily constructed. □

LEMMA 4.6. *For even base $\beta \geq 2$ let $1 \leq k \in \mathbb{N}$ and $a, b \in \mathbb{F}_k$ be given. If $2k \leq m \in \mathbb{N}$, then*

$$\mathrm{fl}_k(\mathrm{fl}_m(a/b)) = \mathrm{fl}_k(a/b).$$

This statement is not true for $m = 2k-1$.

PROOF. After suitable scaling with a power of β we may assume without loss of generality that $1 \leq a, b < \beta$. Abbreviating $g := \mathrm{fl}_m(a/b)$ assume $\mathrm{fl}_k(g) \neq \mathrm{fl}_k(a/b)$, so that Lemma 4.1 and (11) imply $g = f + \mathbf{u}_k\mathrm{ufp}(f)$ for some $f \in \mathbb{F}_k$. In particular (7) implies $f \in \beta^{1-k}\mathrm{ufp}(f)\mathbb{Z}$, and therefore $g \in \mathbf{u}_k\mathrm{ufp}(f)\mathbb{Z}$.

We distinguish two cases. First, assume $1 \leq a < b < \beta$. Then

$$\frac{a}{b} \leq \frac{b-\beta^{1-k}}{b} < 1 - \beta^{1-k} \in \mathbb{F}_k \subset \mathbb{F}_m$$

by (9). Hence $g \leq 1-\beta^{1-k}$ and $\beta^{-1} \leq f < 1$, so that $\mathrm{ufp}(f) = \mathrm{ufp}(a/b) = \beta^{-1}$. Moreover, $g = a/b + \varepsilon$ with $|\varepsilon| \leq \mathbf{u}_m\mathrm{ufp}(a/b) = \beta^{-1}\mathbf{u}_m$. Furthermore, $g \in \frac{1}{2}\beta^{-k}\mathbb{Z}$ and $a, b \in \beta^{1-k}\mathbb{Z}$ yield $\varepsilon b = gb - a \in \frac{1}{2}\beta^{-k} \cdot \beta^{1-k}\mathbb{Z} \subseteq \mathbf{u}_m\mathbb{Z}$, so that $|\varepsilon b| < \mathbf{u}_m$ implies $\varepsilon = 0$ and $g = a/b$.

Second, assume $1 \leq b < a < \beta$. Similarly $1 \leq f < \beta$ and $|\varepsilon| \leq \mathbf{u}_m$, but $gb - a \in \mathbf{u}_k \cdot \beta^{1-k}\mathbb{Z} \subseteq \beta\mathbf{u}_m\mathbb{Z}$. Again it follows $\varepsilon = 0$ and the assertion.

For base $\beta = 2$ and $k = 4$, $m = 2k-1$, $a = 16$ and $b = 15$ it follows $a/b = (1.\overline{0001})_2$, so that $\mathrm{fl}_4(a/b) = (1.001)_2$ but $\mathrm{fl}_4(\mathrm{fl}_7(a/b)) = \mathrm{fl}_4((1.0001)_2) = 1$. For base $\beta = 10$, $k = 2$ and $a = 10, b = 22$ it follows $a/b = 0.45\overline{45}$, so that $\mathrm{fl}_2(a/b) = 0.45$ but $\mathrm{fl}_2(\mathrm{fl}_3(a/b)) = \mathrm{fl}_2(0.455) = 0.46$. Examples for other values of β and k are easily constructed. □

LEMMA 4.7. *For even base $\beta \geq 2$ let $1 \leq k \in \mathbb{N}$ and $0 \leq a \in \mathbb{F}_k$ be given. If $m \geq 2k+2$ for $\beta = 2$ and $m \geq 2k+1$ for $\beta \geq 4$, then*

$$\mathrm{fl}_k(\mathrm{fl}_m(\sqrt{a})) = \mathrm{fl}_k(\sqrt{a}). \tag{28}$$

For $\beta = 2$, $k \geq 1$ and $m = 2k+1$, the unique exception, up to scaling by even powers of 2, that (28) is not true is $a = \mathrm{pred}_k(4) = 4(1-\mathbf{u}_k)$. Moreover, (28) is also not true for $\beta \geq 4$ and $m = 2k$.

REMARK. For $\beta \geq 4$ and $m = 2k$ the example needs not to be unique. For example, for $\beta = 10$ and $k = 2$ we have

$$\mathrm{fl}_2(\mathrm{fl}_4(\sqrt{99})) = \mathrm{fl}_2(\mathrm{fl}_4(9.9498\ldots)) = \mathrm{fl}_2(9.950) = 10 \neq 9.9 = \mathrm{fl}_2(\sqrt{99})$$

and

$$\mathrm{fl}_2(\mathrm{fl}_4(\sqrt{57})) = \mathrm{fl}_2(\mathrm{fl}_4(7.5498\ldots)) = \mathrm{fl}_2(7.550) = 7.6 \neq 7.5 = \mathrm{fl}_2(\sqrt{57}).$$

PROOF. Set $g := \mathrm{fl}_m(\sqrt{a})$ and assume without loss of generality that $1 \leq a < \beta^2$, so that $1 \leq a \leq \beta^2(1-\beta^{-k})$ by (6). If $g \notin \mathbb{S}_k$ the result follows by Lemma 4.1. Henceforth suppose $g \in \mathbb{S}_k$, so that (11) implies $g = f + \mathbf{u}_k$ for some $f \in \mathbb{F}_k$, $1 \leq f < 2$, and $f + \mathbf{u}_k = \sqrt{a} + \varepsilon$ with $|\varepsilon| \leq \mathbf{u}_m\mathrm{ufp}(\sqrt{a}) = \mathbf{u}_m$.

Then $f \in \beta^{1-k}\mathbb{Z}$ implies $2\varepsilon\sqrt{a} + \varepsilon^2 = (f+\mathbf{u}_k)^2 - a \in \mathbf{u}_k^2\mathbb{Z}$. Furthermore, $|2\varepsilon\sqrt{a} + \varepsilon^2| < 2\mathbf{u}_m \cdot \beta(1-\frac{1}{2}\beta^{-k}) + \mathbf{u}_m^2 < 2\beta\mathbf{u}_m$. A short computation verifies $2\beta\mathbf{u}_m \leq \mathbf{u}_k^2$ both for $\beta = 2$ with $m \geq 2k+2$ and for $\beta \geq 4$ with $m \geq 2k+1$. It follows that $\varepsilon(2\sqrt{a}+\varepsilon) = 0 = \varepsilon$. Hence $f + \mathbf{u}_k = g = \sqrt{a} \in \mathbb{F}_m$ proves the first assertion (28).

To prove the second assertion suppose $\mathrm{fl}_k(g) \neq \mathrm{fl}_k(\sqrt{a})$ for $\beta = 2$ and $m = 2k+1$. Then as before $\Delta := 2\varepsilon\sqrt{a} + \varepsilon^2 \in \mathbf{u}_k^2\mathbb{Z}$ and $|\Delta| < 4\mathbf{u}_m = 2\mathbf{u}_k^2$ implies $\Delta \in \{-1,0,1\} \cdot \mathbf{u}_k^2$. If $\Delta = 0$, then $\varepsilon = 0$ and $g = \sqrt{a}$, contradicting $\mathrm{fl}_k(g) \neq \mathrm{fl}_k(\sqrt{a})$. If $\Delta = -\mathbf{u}_k^2$, then $-2\mathbf{u}_k^2 = f^2 + 2\mathbf{u}_k f - a \in 4\mathbf{u}_k^2\mathbb{Z}$, a contradiction.

It remains the case $\Delta = \mathbf{u}_k^2$, so that $f^2 + 2\mathbf{u}_k f = a$. Then, $f \in 2\mathbf{u}_k\mathbb{Z}$ implies $2^k(f + \mathbf{u}_k) = 2^k\sqrt{a + \mathbf{u}_k^2} \in \mathbb{N}$, or

$$2^{2k} \cdot a + 1 = p^2 \quad \text{for} \quad p \in \mathbb{N}. \tag{29}$$

Furthermore, $a \in 2\mathbf{u}_k\mathbb{Z}$ so that $2^{2k}a = 2^{k+1}q$ for some $q \in \mathbb{N}$. Hence

$$2^{k+1}q + 1 = p^2 \quad \text{for} \quad p, q \in \mathbb{N}. \tag{30}$$

Moreover, (29) yields $2^{2k} + 1 \le p^2 \le 2^{2k} \cdot 4(1 - \mathbf{u}_k) + 1 = 2^{2k+2} - 2^{k+2} + 1$ and

$$2^k + 1 \le p \le 2^{k+1} - 1.$$

Now (30) gives

$$2^{k-1} + 1 = 2^{-k-1}((2^k + 1)^2 - 1) \le q = \frac{p^2 - 1}{2^{k+1}} \le 2^{k+1} - 2.$$

Set $e := p - 2^k$, so that $1 \le e \le 2^k - 1$. Then $p^2 = 2^{2k} + 2^{k+1}e + e^2 = 2^{k+1}q + 1$ implies

$$2^{k+1} | (e^2 - 1). \tag{31}$$

If $e = 1$, then $p = 2^k + 1$, $a = 2^{-2k}(2^{2k} + 2^{k+1}) = 1 + 2^{-k+1} = \mathrm{succ}_k(1)$, $(1 + 2^{-k})^2 > 1 + 2^{-k+1} = a$ and therefore $\mathrm{fl}_k(\sqrt{a}) = 1$. But $\mathrm{fl}_m(\sqrt{a}) = g = 1 + \mathbf{u}_k$ implies $\mathrm{fl}_k(g) = \mathrm{fl}_k(\sqrt{a})$, a contradiction.

Finally we show that $m = 2k+1$ for $\beta = 2$ and $m = 2k$ for $\beta \ge 4$ is not sufficient. First, suppose $2 \le e \le 2^k - 1$. Let j be the largest power of 2 dividing $e+1$ or $e-1$, then $2^{j+2} \nmid (e^2 - 1)$. Now (31) gives $j \ge k$, and combining $2^k | (e+1)$ or $2^k | (e-1)$ with $2 \le e \le 2^k - 1$ implies $k \ge 2$ and $e = 2^k - 1$. This corresponds to $p = 2^{k+1} - 1$ and

$$a = 2^{-2k}(2^{2k+2} - 2^{k+2}) = 4(1 - \mathbf{u}_k).$$

Now $[2(1 - \frac{1}{2}\mathbf{u}_k)]^2 > a$ gives $\mathrm{fl}_k(\sqrt{a}) = 2(1 - \mathbf{u}_k)$. Furthermore, $2(1 - \frac{1}{2}\mathbf{u}_k) \in \mathbb{F}_m$ and $\mathrm{pred}_m(2(1 - \frac{1}{2}\mathbf{u}_k)) = 2(1 - \frac{1}{2}\mathbf{u}_k - \mathbf{u}_m)$, so that $\mathbf{u}_m = \frac{1}{2}\mathbf{u}_k^2$ yields

$$[2(1 - \frac{1}{2}\mathbf{u}_k - \frac{\mathbf{u}_m}{2})]^2 < a = 4(1 - \mathbf{u}_k).$$

Therefore, $g = \mathrm{fl}_m(\sqrt{a}) = 2(1 - \frac{1}{2}\mathbf{u}_k)$, and this implies $\mathrm{fl}_k(g) = 2 \neq \mathrm{fl}_k(\sqrt{a})$, the predicted unique exception for $m = 2k+1$.

For $\beta \ge 4$ and $m = 2k$ define $a = 1 - \beta^{-k} = \mathrm{pred}_k(1)$. Then

$$a = \mathrm{pred}_k(1) < \sqrt{a} < 1 - \frac{1}{2}\beta^{-k} = \frac{1}{2}\big(\mathrm{pred}_k(1) + 1\big) =: s \quad \Rightarrow \quad \mathrm{fl}_k(\sqrt{a}) = a,$$

but

$$\frac{1}{2}\big(\mathrm{pred}_m(s) + s\big) = 1 - \frac{1}{2}\beta^{-k} - \frac{1}{2}\beta^{-2k} < \sqrt{1 - \beta^{-k}} = \sqrt{a} < s$$

and rounding tie to even yields $\mathrm{fl}_k(\mathrm{fl}_m(\sqrt{a})) = \mathrm{fl}_k(s) = 1$. This completes the proof. □

For a computer implementation, a simple way is to restrict the exponent range for the anticipated arithmetic in $\mathbb{F}_{\beta,k,E}$ so that $f := \mathrm{fl}_m(a \text{ op } b)$ is always a finite normalized double precision floating-point number in$\mathbb{F}_{\beta,m,E'}$ for all $a, b \in \mathbb{F}_{\beta,k,E}$.

5. DIRECTED ROUNDING

The difficulty in the proof of correctness of Algorithm 1 (`flround`) was the behavior near switching points. In case of directed rounding, the floating-point numbers themselves are the switching points. Hence for directed rounding downwards, upwards or towards zero there is no double rounding, and computing $\mathrm{fl}^{\nabla}_{k,E}, \mathrm{fl}^{\Delta}_{k,E}$ or $\mathrm{fl}^{><}_{k,E}$ reduces to compute the correct rounding from working into target precision by Algorithm 3.

Moreover, in that case there cannot be double rounding when first computing the floating-point result in double precision with directed rounding, and then use a second directed rounding into k-bit format. This is even true for just $m \geq k$ because the switching points are the floating-point numbers themselves and $\mathbb{F}_k \subseteq \mathbb{F}_m$. This allows a simple implementation of k-bit interval arithmetic, also for vectors and matrices.

6. SUMMARY

Suppose an m-digit working precision in base $\beta \geq 2$ following the IEEE 754 standard is given, and a k-digit target precision with respect to the same base β, and $1 \leq k < m$.

For given x in working precision we presented algorithms using solely operations in working precision to compute the correctly rounded image of x into the target precision including exceptional values and signed zero.

Moreover, sharp estimates on m relative to k were presented such that computing addition, subtraction, multiplication, division and/or square root in working precision and rounding the result into the target precision produces the same result as if computing the result directly in target precision. The only exception to that statement is division for odd base β, in which case for every $k \geq 1$ examples $a, b \in \mathbb{F}_k$ exist such that $\mathrm{fl}_k(\mathrm{fl}_m(a/b)) \neq \mathrm{fl}_k(a/b)$ for any $m > k$. We handle division for odd β by computing an auxiliary quantity q in working precision such that $\mathrm{fl}_k(q) = \mathrm{fl}_k(a/b)$.

A toolbox for k-bit binary arithmetic for $k \leq 26$ including directed rounding and intervals of scalars, vectors and matrices is part Version 8 of INTLAB [Rump 1999], the Matlab toolbox for reliable computing. The unique exception for the square root for $k = 26$ is explicitly checked for.

ACKNOWLEDGMENTS

I want to thank the anonymous referees for their thorough reading and most constructive and very valuable comments. The first version of the paper covered only binary arithmetic and one referee asked about decimal arithmetic; this lead to the treatment of general bases $\beta \geq 2$ in this note.

Moreover many thanks to Marko Lange for many fruitful discussions and in particular for pointing to the simpler version of Algorithm 1 and suggesting the method in Lemma 4.3.

REFERENCES

Samuel A. Figueroa. 1995. When is Double Rounding Innocuous? *SIGNUM Newsl.* 30, 3 (1995), 21–26.

Samuel Arturo Figueroa Del Cid. 2000. *A Rigorous Framework for Fully Supporting the Ieee Standard for Floating-point Arithmetic in High-level Programming Languages*. Ph.D. Dissertation. New York University, New York, NY, USA. Advisor(s) Dewar, Robert B.

N. J. Higham. 2002. *Accuracy and stability of numerical algorithms* (2nd ed.). SIAM Publications, Philadelphia.

IEEE 2008. *ANSI/IEEE 754-2008: IEEE Standard for Floating-Point Arithmetic*. IEEE, New York.

C.-P. Jeannerod and S.M. Rump. 2014. On relative errors of floating-point operations: optimal bounds and applications. Preprint. (2014).

D.E. Knuth. 1998. *The Art of Computer Programming: Seminumerical Algorithms* (third ed.). Vol. 2. Addison Wesley, Reading, Massachusetts.

Vincent Lefèvre. 2013. Sipe: a Mini-Library for Very Low Precision Computations with Correct Rounding. (2013). http://hal.inria.fr/hal-00864580 submitted.

J.-M. Muller, N. Brisebarre, F. de Dinechin, C.-P. Jeannerod, V. Lefèvre, G. Melquiond, R. Revol, D. Stehlé, and S. Torres. 2009. *Handbook of Floating-Point Arithmetic*. Birkhäuser, Boston.

S.M. Rump. 1999. INTLAB - INTerval LABoratory. In *Developments in Reliable Computing*, Tibor Csendes (Ed.). Kluwer Academic Publishers, Dordrecht, 77–104. http://www.ti3.tu-harburg.de/rump/intlab/index.html

S.M. Rump. 2009. Ultimately Fast Accurate Summation. *SIAM Journal on Scientific Computing (SISC)* 31, 5 (2009), 3466–3502.

G.W. Stewart. 2009. Flap: A Matlab Package for Adjustable Precision Floating-Point Arithmetic. `http://www.cs.umd.edu/ stewart/flap/flap.html`. (2009).

G.W. Stewart. 2014. private communication. (2014).

Received February 0000; revised March 0000; accepted June 0000