BIT manuscript No.
(will be inserted by the editor)

On the Definition of Unit Roundoff

Siegfried M. Rump · **Marko Lange**

Received: date / Accepted: date

Abstract The result of a floating-point operation is usually defined to be the floating-point number nearest to the exact real result together with a tie breaking rule. The analysis of numerical algorithms is often solely based on the first and/or the second standard model specifying the maximum relative error with respect to the true and/or to the computed result, respectively. In this note we take a more general perspective. For an arbitrary finite set of real numbers we identify the rounding to minimize the relative error in the first or the second standard model. The optimal "switching points" are the arithmetic or the harmonic means of adjacent floating-point numbers. Moreover, the maximum relative error of both models is minimized by taking the geometric mean. If the maximum relative error in one model is α, then $\alpha/(1-\alpha)$ is the maximum relative error in the other model. Those maximal errors, the unit roundoff, are characteristic constants of a given finite set of reals: The floating-point model to be optimized identifies the rounding and the unit roundoff.

Keywords floating-point number · IEEE 754 · rounding · tie

CR Subject Classification 65G50 · 65F05

S.M. Rump
Institute for Reliable Computing, Hamburg University of Technology, Schwarzenbergstraße 95, 21071 Hamburg, Germany,
and Faculty of Science and Engineering, Waseda University,
3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan
E-mail: rump@tuhh.de

M. Lange
Institute for Reliable Computing, Hamburg University of Technology, Schwarzenbergstraße 95, Hamburg 21071, Germany
E-mail: marko.lange@tuhh.de

1 Introduction

Usually floating-point arithmetic over a finite set $\mathbb{F} \subseteq \mathbb{R}$ is introduced by defining the result of a floating-point operation $\mathrm{op} \in \{+,-,*,/\}$ to be $\mathrm{fl}(a \,\mathrm{op}\, b)$ for $a,b \in \mathbb{F}$ and a rounding to nearest $\mathrm{fl}\colon \mathbb{R} \to \mathbb{F}$. The rounding is unique up to a tie breaking rule.

More generally, rounding error analysis is often based on the first or the second standard model [2, Section 2]. The *first standard model* of arithmetic declares

$$a,b \in \mathbb{F}: \qquad \mathrm{fl}(a \,\mathrm{op}\, b) = (a \,\mathrm{op}\, b)(1+\delta), \quad |\delta| \leqslant u, \quad \mathrm{op} = +,-,*,/, \tag{1.1}$$

while the *second standard model* assumes

$$\mathrm{fl}(a \,\mathrm{op}\, b) = \frac{a \,\mathrm{op}\, b}{1+\delta}, \qquad |\delta| \leqslant u. \tag{1.2}$$

This is true for all $a,b \in \mathbb{F}$ as long as no overflow or underflow occurs, our general assumption for this note. For $\mathbb{F}$ denoting the set of IEEE 754 binary64 floating-point numbers [3] the relative rounding error unit $u := 2^{-53}$ is used for both models.

In the following we assume an arbitrary finite set $\mathfrak{F} \subseteq \mathbb{R}$ to be given. Without further assumptions we analyze best possible approximation properties of real numbers by elements in $\mathfrak{F}$. For minimizing the relative error with respect to the true result, i.e. the first standard model, this is clearly rounding to nearest. However, we may alternatively and also simultaneously minimize the relative error with respect to the second standard model.

The finiteness of $\mathfrak{F}$ implies that a small relative error cannot be achieved for real numbers which are very large or very small in absolute value. To state that more precisely we define the *normal range*[1] of $\mathfrak{F}$ by

$$\varrho(\mathfrak{F}) := [\min \mathfrak{F}_{<0}, \max \mathfrak{F}_{<0}] \cup [\min \mathfrak{F}_{>0}, \max \mathfrak{F}_{>0}] \subset \mathbb{R} \tag{1.3}$$

and assume $\varrho(\mathfrak{F}) \neq \emptyset$. Zero may be in $\mathfrak{F}$ or not, but $0 \notin \varrho(\mathfrak{F})$ by (1.3). Here $x > \max \mathfrak{F}_{>0}$ or $x < \min \mathfrak{F}_{<0}$ corresponds to overflow, and $\max \mathfrak{F}_{<0} < x < \min \mathfrak{F}_{>0}$ to underflow. Note that $\varrho(\mathfrak{F})$ consists of all real numbers in one of the two intervals in (1.3).

2 Main results

The smallest possible constant u for the first standard model (1.1) is characterized by

$$\boldsymbol{\alpha} := \boldsymbol{\alpha}(\mathfrak{F}) = \sup_{x \in \varrho(\mathfrak{F})} \min_{f \in \mathfrak{F}} \left| \frac{f-x}{x} \right|. \tag{2.1}$$

Note that the minimum is taken over all elements $f \in \mathfrak{F}$, whereas x is restricted to the range of $\mathfrak{F}$. The constant $\boldsymbol{\alpha}$ minimizes the relative error with respect to x, i.e. the maximum error of the first standard model (1.1). It is a characteristic constant of a given set $\mathfrak{F}$.

[1] For the standard formats $\mathbb{F}$ in IEEE 754 the range could be slightly wider: For f denoting the rounded-to-nearest result in $\mathbb{F}$ with infinite exponent range, return this f if it belongs to $\mathbb{F}$ with the bounded exponent range. Since we are aiming on general sets $\mathfrak{F}$, there is no notion of "exponent range".

For a given $x \in \varrho(\mathfrak{F})$, denote a minimizing $f \in \mathbb{F}$ in (2.1) by $\mathrm{fl}(x)$. Definition (2.1), the continuity of $|(f-x)/x|$ and the finiteness of $\mathfrak{F}$ imply that $\mathrm{fl}(x)$ is unique up to finitely many real numbers in $\varrho(\mathfrak{F})$. The latter are called the *switching points* of the otherwise uniquely defined rounding $\mathrm{fl}\colon \varrho(\mathfrak{F}) \to \mathbb{F}$.

The concept can be extended to $\mathrm{fl}\colon \mathbb{R} \to \mathbb{F}$, so that any rounding to nearest $\mathrm{fl}\colon \mathbb{R} \to \mathfrak{F}$ with $|\mathrm{fl}(x)-x| = \min\{|f'-x| : f' \in \mathfrak{F}\}$ for all $x \in \mathbb{R}$ satisfies

$$\boldsymbol{\alpha} = \sup_{x \in \varrho(\mathfrak{F})} \left| \frac{\mathrm{fl}(x)-x}{x} \right|, \tag{2.2}$$

no matter how ties are rounded. The switching points are the arithmetic means of adjacent elements in $\mathfrak{F}$.

The rounding defines the floating-point operations. For rounding to nearest $\mathrm{fl}\colon \mathbb{R} \to \mathfrak{F}$, the maximum error for the second standard model (1.2) is given by

$$\boldsymbol{\beta} := \boldsymbol{\beta}(\mathfrak{F}) = \sup_{x \in \varrho(\mathfrak{F})} \left| \frac{\mathrm{fl}(x)-x}{\mathrm{fl}(x)} \right|. \tag{2.3}$$

Since the supremum is taken, the definition of $\boldsymbol{\beta}$ is independent of a tie breaking rule. Thus, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are well defined and are characteristic invariants of the set $\mathfrak{F}$, and they are related as follows.

Lemma 2.1 *Let finite $\mathfrak{F} \subseteq \mathbb{R}$ with $\varrho(\mathfrak{F}) \neq \emptyset$ be given. If the positive and the negative parts of $\varrho(\mathfrak{F})$ consist of at most one element, then $\boldsymbol{\alpha} = \boldsymbol{\beta} = 0$. Otherwise,*

$$\boldsymbol{\alpha} = \frac{\boldsymbol{\beta}}{1+\boldsymbol{\beta}} = \max\{\left|\frac{g-f}{g+f}\right| : f, g \text{ adjacent in } \varrho(\mathfrak{F}) \text{ with } fg > 0\} < 1 \tag{2.4}$$

for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ defined by (2.1) and (2.3), respectively.

Proof If the positive and/or negative part of $\varrho(\mathfrak{F})$ is empty or consists of only one element, there is nothing to prove for that part. Let fixed but arbitrary positive adjacent elements $f, g \in \mathfrak{F} \cap \varrho(\mathfrak{F})$ be given with $0 < f < g$. For $x \in [f, g]$ we have $\mathrm{fl}(x) \in \{f, g\}$, and clearly $|\mathrm{fl}(x)-x|/|x|$ is maximized for $x = m := (f+g)/2$. This maximum is

$$\alpha := \frac{m-f}{m} = \frac{g-m}{m} = \frac{(g-f)/2}{(g+f)/2} = \frac{g-f}{g+f}.$$

This implies in particular $\alpha < 1$. For increasing $x \in [f, m)$ the function $|f-x|/|f|$ increases, and for increasing $x \in (m, g]$ the function $|g-x|/|g|$ decreases. For $x := m - \varepsilon$ and sufficiently small $\varepsilon > 0$ we have $\mathrm{fl}(x) = f$ and

$$\frac{|f-x|}{|f|} = \frac{(g+f)/2 - \varepsilon - f}{f} = \frac{g-f}{2f} - O(\varepsilon).$$

For $x := m + \varepsilon$ and sufficiently small $\varepsilon > 0$ we have $\mathrm{fl}(x) = g$ and

$$\frac{|g-x|}{|g|} = \frac{g - (g+f)/2 - \varepsilon}{g} < \frac{g-f}{2g} < \frac{g-f}{2f}.$$

Thus

$$\beta := \sup\{\left|\frac{\mathrm{fl}(x)-x}{\mathrm{fl}(x)}\right| : f \leqslant x \leqslant g\} = \frac{g-f}{2f}.$$

Moreover,

$$\frac{\alpha}{1-\alpha} = \frac{(g-f)/(g+f)}{(g+f-g+f)/(g+f)} = \frac{g-f}{2f} = \beta \quad \text{and therefore} \quad \alpha = \frac{\beta}{1+\beta}.$$

A similar argument applies to negative adjacent elements $f, g \in \varrho(\mathfrak{F})$, so that on every such interval $[f,g]$ the maximal α and β are related by $\beta = \alpha/(1-\alpha)$. Since f, g are chosen arbitrarily and since the mapping $\alpha \to \alpha/(1-\alpha)$ is increasing for $0 \leqslant \alpha < 1$, the proof is finished. □

For $\mathfrak{F}$ denoting the set $\mathbb{F}$ of IEEE 754 binary64 floating-point numbers barring underflow we may restrict x in Definition (2.1) to the interval $[1,2)$. Abbreviating $u := 2^{-53}$, the maximum is achieved for $x = 1+u$, the midpoint of 1 and its successor in $\mathbb{F}$. It follows $\boldsymbol{\alpha} = u/(1+u)$ for the first standard model, and $\boldsymbol{\beta} = u$ for the second one, the characteristic constants of $\mathbb{F}$. In the literature (e.g. [2]) the same constant u is used for both standard models. This improvement for the first model was also noted in [4].

Next we ask for a rounding realizing the smallest possible constant u for the second standard model (1.2). The corresponding characteristic constant is defined by

$$\mathbf{w} := \mathbf{w}(\mathfrak{F}) = \sup_{x \in \varrho(\mathfrak{F})} \min_{f \in \mathfrak{F} \setminus \{0\}} \left|\frac{f-x}{f}\right|. \tag{2.5}$$

For given $x \in \varrho(\mathfrak{F})$ we first identify $\varphi \in \mathfrak{F}$ minimizing $|\varphi - x|/|\varphi|$. If $x \in \mathfrak{F}$, then clearly it suffices to take $\varphi := x$. Otherwise the definition (1.3) of the range implies the existence of adjacent $f, g \in \mathfrak{F} \cap \varrho(\mathfrak{F})$ with $f < x < g$. The function $|\varphi - x|/|\varphi|$ decreases on $\varphi \in (0, x]$ and increases on $\varphi \in [x, \infty)$, so that $\varphi \in \{f, g\}$. For increasing $y \in [f,g]$ the ratio $|f-y|/|f|$ increases and $|g-y|/|g|$ decreases. Equality is achieved for $y := h(f,g) := 2(\frac{1}{f} + \frac{1}{g})^{-1}$, the harmonic mean of f and g, because

$$\left|\frac{f-y}{f}\right| = \frac{2fg/(f+g)-f}{f} = \frac{g-f}{g+f} = \frac{g-2fg/(f+g)}{g} = \left|\frac{g-y}{g}\right|. \tag{2.6}$$

Thus φ minimizing $|\varphi - x|/|\varphi|$ is f for $x \in [f, h(f,g))$, it is g for $x \in (h(f,g), g]$, and is any of f or g if $x = h(f,g)$. Hence a rounding $\mathrm{gl}\colon \varrho(\mathfrak{F}) \to \mathfrak{F}$ with the property

$$\left|\frac{\mathrm{gl}(x)-x}{\mathrm{gl}(x)}\right| = \min_{f \in \mathfrak{F}} \left|\frac{f-x}{f}\right|$$

is characterized by being a projection and satisfying

$$\mathrm{gl}(x) \begin{cases} = f & \text{for } x \in [f,h) \\ \in \{f,g\} & \text{for } x = h \\ = g & \text{for } x \in (h,g] \end{cases} \tag{2.7}$$

for $x \in [f,g]$ with adjacent $f, g \in \varrho(\mathfrak{F})$ and h denoting the harmonic mean of f and g.

For rounding to nearest the switching point is the arithmetic mean of adjacent elements $f, g \in \mathfrak{F}$, now it is the harmonic mean. In both cases there is freedom about the tie without jeopardizing the minimization properties.

For a rounding $\mathrm{gl}\colon \varrho(\mathfrak{F}) \to \mathfrak{F}$ with (2.7) we showed

$$\mathbf{w} = \mathbf{w}(\mathfrak{F}) = \sup_{x \in \varrho(\mathfrak{F})} \left| \frac{\mathrm{gl}(x) - x}{\mathrm{gl}(x)} \right|, \tag{2.8}$$

minimizing the error of the second standard model (1.2). Note that $\mathbf{w}$, just like $\boldsymbol{\alpha}$, solely depends on $\mathfrak{F}$. Using the rounding $\mathrm{gl}\colon \varrho(\mathfrak{F}) \to \mathfrak{F}$ with (2.7) to minimize the second standard model (1.2), the maximum error for the first standard model (1.1) is given by

$$\mathbf{v} := \mathbf{v}(\mathfrak{F}) = \sup_{x \in \varrho(\mathfrak{F})} \left| \frac{\mathrm{gl}(x) - x}{x} \right|. \tag{2.9}$$

As before we will show that $\mathbf{v}$ is invariant for any rounding $\mathrm{gl}\colon \mathbb{R} \to \mathfrak{F}$ with (2.7). Thus $\mathbf{v}$ and $\mathbf{w}$ are well defined and are again characteristic invariants of the set $\mathfrak{F}$. For any $\mathfrak{F}$, the invariants $\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{v}, \mathbf{w}$ are related as follows.

Lemma 2.2 *Let finite $\mathfrak{F} \subseteq \mathbb{R}$ with $\varrho(\mathfrak{F}) \neq \emptyset$ be given. Let $\mathbf{v}$, $\mathbf{w}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ as defined in (2.9), (2.5), (2.1) and (2.3), respectively. If the positive and the negative parts of $\varrho(\mathfrak{F})$ consist of at most one element, then $\mathbf{v} = \mathbf{w} = 0$. Otherwise,*

$$\mathbf{v} = \frac{\mathbf{w}}{1 - \mathbf{w}} = \max\left\{ \left| \frac{g - f}{2\min\{|f|, |g|\}} \right| : f, g \text{ adjacent in } \varrho(\mathfrak{F}) \text{ with } fg > 0 \right\} \tag{2.10}$$

and therefore

$$\mathbf{v} = \boldsymbol{\beta} \qquad \text{and} \qquad \mathbf{w} = \boldsymbol{\alpha} < 1. \tag{2.11}$$

Proof If the positive and/or negative part of $\varrho(\mathfrak{F})$ is empty or consists of only one element, there is nothing to prove for that part. Let fixed but arbitrary positive adjacent elements $f, g \in \mathfrak{F} \cap \varrho(\mathfrak{F})$ be given with $0 < f < g$. By (2.6) the ratio $|\mathrm{gl}(x) - x|/|\mathrm{gl}(x)|$ is maximal for x being the harmonic mean of f and g. That maximum value is $(g - f)/(g + f)$, so that applying this argument to the negative part of $\varrho(\mathfrak{F})$ and using (2.4) shows $\mathbf{w} = \boldsymbol{\alpha} < 1$.

The supremum of the ratio $|\mathrm{gl}(x) - x|/|x|$ is also achieved for the harmonic mean $x := 2fg/(f + g)$, and using $x - f < g - x$ it is equal to

$$\frac{g - x}{x} = \frac{g}{2fg/(f + g)} - 1 = \frac{g - f}{2f} = \frac{\alpha}{1 - \alpha}$$

for $\alpha := (g - f)/(g + f)$. Applying the same argument to the negative part of $\varrho(\mathfrak{F})$ and using (2.9) and (2.4) implies

$$\mathbf{v} = \sup\left\{ \left| \frac{g - f}{2\min\{|f|, |g|\}} \right| : f, g \text{ adjacent in } \varrho(\mathfrak{F}) \text{ with } fg > 0 \right\} = \frac{\boldsymbol{\alpha}}{1 - \boldsymbol{\alpha}}.$$

A computation using (2.4) yields $\boldsymbol{\beta} = \boldsymbol{\alpha}/(1 - \boldsymbol{\alpha})$, and the results follow. □

For a given set $\mathfrak{F} \subseteq \mathbb{R}$, knowing one of the invariants $\mathbf{v}$, $\mathbf{w}$, $\boldsymbol{\alpha}$ or $\boldsymbol{\beta}$ means knowing the others. Choosing the appropriate rounding fl or gl, i.e. the switching points being the arithmetic or the harmonic mean, the maximum error for both models is the same.

For any finite set $\mathfrak{F}$ one invariant can be computed using (2.4) or (2.10), then the other invariants follow. As an example consider a logarithmic number system (LNS)

$$\mathfrak{F} := \{c^k : k \in \mathbb{Z}, k_1 \leqslant k \leqslant k_2\} \qquad \text{for } 1 < c \in \mathbb{R}. \tag{2.12}$$

Earliest references of this well studied concept include [5,7,6]; see [1] and the literature cited over there for recent publications. Assume the integers k_1, k_2 satisfy $k_1 < k_2$ to ensure that $\varrho(\mathfrak{F})$ has at least two elements. Then $\varrho(\mathfrak{F}) = [c^{k_1}, c^{k_2}]$, and (2.10) yields

$$\mathbf{v} = \frac{c^{k+1} - c^k}{2c^k} = \frac{c-1}{2} \qquad \text{and} \qquad \mathbf{w} = \frac{\mathbf{v}}{1+\mathbf{v}} = \frac{c-1}{c+1}.$$

These characteristic constants are the same for $\mathfrak{F} \cup (-\mathfrak{F})$ or for $\mathfrak{F} \cup \{0\} \cup (-\mathfrak{F})$. The elements could be stored by the integer exponent. If the product or quotient of elements in $\mathfrak{F}$ is in $\varrho(\mathfrak{F})$, then they are in $\mathfrak{F}$. In terms of floating-point operations this means that multiplication and division are error-free if no over- or underflow occurs.

Moreover, those sets bear the advantage that the maximum relative rounding error for the first and for the second standard model is attained on each interval $[f, g]$ of adjacent elements in $\varrho(\mathfrak{F})$. In contrast, there is a factor of up to β for floating-point systems to base β, depending on whether the interval $[f, g]$ is slightly larger or smaller than a power of β.

A possible choice is $c := 2^{\varepsilon}$ for $\varepsilon := 2^{-52}$. Spending one bit for the sign, integers up to about $\pm 2^{62}$ can be stored in 64-bit, so that the positive range is about $[2^{-1024}, 2^{1024}]$, similar to IEEE 754 double precision. The relative rounding error unit is then about $0.77 \cdot 10^{-16}$ compared to about $1.11 \cdot 10^{-16}$ in double precision. With this choice of c powers of 2 are in $\mathfrak{F}$, however, as a disadvantage, $\mathfrak{F}$ contains no other integers.

3 Conclusion

From the perspective of a rounding function the situation for binary64 can be summarized as follows. If the switching point for the rounding is set to the *arithmetic mean* of adjacent elements, that is the usual rounding to nearest, then we obtain the bound $\boldsymbol{\beta} = 2^{-53}$ for the second standard model, and $\boldsymbol{\alpha} = 2^{-53}/(1 + 2^{-53})$ is the optimal error bound for the first standard model.

Defining the switching points to be the *harmonic mean* of adjacent elements, the invariants change places: The maximum relative error with respect to x (the first standard model) is $\mathbf{v} = 2^{-53}$, whereas the maximum error relative to its rounded value f (the second standard model) becomes $\mathbf{w} = 2^{-53}/(1 + 2^{-53})$. Now $\mathbf{w} = \boldsymbol{\alpha}$ becomes the optimal error bound for the second standard model.

By continuity there must be a switching point between the harmonic and the arithmetic mean for which both the maximum relative error of all x with respect to both standard models is bounded by the same constant. Using the arithmetic-geometric mean inequality one verifies that for adjacent floating-point numbers $0 < f < g$ the

maximum relative error for both models is equal to $\sqrt{g/f} - 1$ for x equal to the *geometric mean* $\sqrt{fg}$, and that this relative error is maximized for $f = 2^e$ and its successor $g = (1 + 2\mathbf{v})2^e$.

It follows that for the switching points $x := \mathrm{sign}(f)\sqrt{|fg|}$ the relative errors of both the first and the second standard models are bounded by the same characteristic constant $\sqrt{1 + 2\mathbf{v}} - 1$, slightly less than $\mathbf{v}$ and slightly larger than $\mathbf{v}/(1 + \mathbf{v})$. The relation harmonic $\leqslant$ geometric $\leqslant$ arithmetic mean finds its counterpart in the corresponding maximal relative errors $\mathbf{v}/(1 + \mathbf{v}) \leqslant \sqrt{1 + 2\mathbf{v}} - 1 \leqslant \mathbf{v}$.

4 Acknowledgement

Our dearest thanks go to Claude-Pierre Jeannerod from Lyon for his many detailed comments and for very helpful discussions and suggestions.

References

1. M.G. Arnold and S. Collange. The denormal logarithmic number system. In *24th IEEE International Conference on Application-specific Systems, Architectures and Processors (ASAP)*, pages 117–124, 2013.
2. N. J. Higham. *Accuracy and stability of numerical algorithms*. SIAM Publications, Philadelphia, 2nd edition, 2002.
3. IEEE, New York. *ANSI/IEEE 754-2008: IEEE Standard for Floating-Point Arithmetic*, 2008.
4. C.-P. Jeannerod and S.M. Rump. On relative errors of floating-point operations: optimal bounds and applications. Preprint, 2014.
5. N.G. Kingsburg and P.J.W. Rayner. Digital Filtering using logarithmic arithmetic. *Electron. Lett.*, 7:56–58, 1971.
6. S.C. Lee and A.D. Edgar. The focus number system. *IEEE Trans. Comput.*, C-26:1167–1170, 1977.
7. E.E. Swartzlander Jr. and A.G. Alexopoulos. The sign/logarithm number system. *IEEE Trans. Comput.*, C-24:1238–1243, 1975.