Floating point numbers are real numbers

Walter F. Mascarenhas [*1]

[1] IME, Universidade de São Paulo

October 5, 2017

Abstract

Floating point arithmetic allows us to use a finite machine, the digital computer, to reach conclusions about models based on continuous mathematics. In this article we work in the other direction, that is, we present examples in which continuous mathematics leads to sharp, simple and new results about the evaluation of sums, square roots and dot products in floating point arithmetic.

1 Introduction

According to Knuth [14], floating point arithmetic has been used since Babylonia (1800 B.C.). It played an important role in the beginning of modern computing, as in the work of Zuse in the late 1930s. Today we have several models for floating point arithmetic. Some of these models are based on algebraic structures, like Kulisch's Ringoids [13]. Others models validate numerical software and lead to automated proofs of results about floating point arithmetic [1].

There are also models based on continuous mathematics, which are used intuitively. For example, when analysing algorithms based on the floating point operations op $\in \{+, -, *, /\}$, executed with machine precision ε, one usually argues that

$$\mathrm{fl}(x \operatorname{op} y) = (x \operatorname{op} y)(1 + \delta) \qquad \text{with} \qquad |\delta| \leq \varepsilon, \tag{1}$$

where $\mathrm{fl}(z)$ is the rounded value of z. Equation (1) is called "the $(1 + \varepsilon)$ argument." It may not apply in the presence of underflow, but lead to many results in the hands of Wilkinson [23, 22]. The effectiveness of the $(1 + \varepsilon)$ argument is illustrated by Equation 3.4 in Higham [10], which expresses the dot product $\hat{d}_n$ of the vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ as

$$\hat{d}_n = x_1 y_1 (1 + \theta_n) + x_2 y_2 (1 + \theta_n') + x_3 y_3 (1 + \theta_{n-1}) + \cdots + x_n y_n (1 + \theta_2). \tag{2}$$

The θ_k above are bounded in terms of the unit roundoff u as

$$|\theta_k| \leq \frac{ku}{1 - ku} =: \gamma_k, \tag{3}$$

and Equations (2) and (3) are a good example of the use of continuous mathematics to analyze floating point operations. They express well the effects of rounding errors on dot products, and will suffice for most people interested in their numerical evaluation.

---

[*] walter.mascarenhas@gmail.com

The purpose of this article is to simplify and extend the results about floating point arithmetic obtained using the $(1 + \varepsilon)$ argument. We argue that by thinking of the set of floating point numbers as a subset of the real numbers we can use techniques from continuous mathematics to derive and prove non trivial results about floating point arithmetic. For instance, we show that in many circumstances we can replace Higham's γ_k by its linearized counterpart ku and still obtain rigorous bounds. For us, the replacement of γ_k by ku is interesting because it leads to simpler versions of our articles [15, 16, 17], and arguments by other people could be simplified as well. We could, for example, replace some of Wilkinson's 1.06 factors by 1. In fact, we can even replace γ_k by $ku/(1 + ku)$ when estimating the effects of rounding errors in the evaluation of the sum $\mathrm{fl}(\sum_{k=0}^{n} x_k)$ of $n + 1$ numbers. Instead of

$$\left| \mathrm{fl}\left(\sum_{k=0}^{n} x_k \right) - \sum_{k=0}^{n} x_k \right| \leq \frac{nu}{1 - nu} \sum_{k=0}^{n} |x_k|, \tag{4}$$

we prove the sharper bound

$$\left| \mathrm{fl}\left(\sum_{k=0}^{n} x_k \right) - \sum_{k=0}^{n} x_k \right| \leq \frac{nu}{1 + nu} \sum_{k=0}^{n} |x_k|, \tag{5}$$

for arithmetics with subnormal numbers, when we round to nearest with unit roundoff u and $20nu \leq 1$. This bound grows slightly less than linearly with nu, that is, the right hand side is a strictly concave function of nu. Due to this concavity, we can rigorously conclude from Equation (5) that

$$\left| \mathrm{fl}\left(\sum_{k=0}^{n} x_k \right) - \sum_{k=0}^{n} x_k \right| \leq nu \sum_{k=0}^{n} |x_k|, \tag{6}$$

and Equation (6) is simpler than Equation (4). When $x_k \geq 0$, Equation (6) can be improved to

$$x_k \geq 0 \Rightarrow \left| \mathrm{fl}\left(\sum_{k=0}^{n} x_k \right) - \sum_{i=0}^{n} x_k \right| \leq u \sum_{k=1}^{n} \sum_{i=0}^{k} x_i, \tag{7}$$

which is also simple and does not have higher order terms in u. We also analyze dot products, and derive simple and rigorous bounds like

$$\left| \mathrm{fl}\left(\sum_{k=0}^{n} x_k y_k \right) - \sum_{k=0}^{n} x_k y_k \right| \leq (n+1)u \sum_{k=0}^{n} |x_k y_k| \leq (n+1)u \sqrt{\sum_{k=0}^{n} x_k^2} \sqrt{\sum_{k=0}^{n} y_k^2}, \tag{8}$$

provided that $\sum_{k=0}^{n} |x_k y_k|$ is not too small, for the formal concepts of small presented in the statement of our results.

The bound in Equation (5) is new, and was derived with the theory introduced in the present article, but Equations (6) and (8) are not new. Actually, stronger versions of them were already proved by C.-P. Jeannerod and S. Rump in [18] and [19], and we present them here only as an improvement of the older results in [10]. When comparing the content of these references with the present work, please note that there is a slight difference in our bound (8) and similar bounds in them: our dot products involve $n + 1$ pairs of numbers, whereas the sums and dot products in [10, 18, 20] are defined for n numbers, or pairs of numbers. Therefore, to leading order in u, Equation (8) states exactly the same as the analogous equations (3.7) in [10]:

$$\left| \mathbf{x}^{\mathsf{T}} \mathbf{y} - \mathrm{fl}\left(\mathbf{x}^{\mathsf{T}} \mathbf{y} \right) \right| \leq nu \, |\mathbf{x}|^{\mathsf{T}} |\mathbf{y}| + O\left(u^2 \right), \tag{9}$$

2

because n in [10] is the same as $n+1$ for us. Our contribution is to show that there is no need for the $O(u^2)$ term in Equation (9) or the 1.06 factor in Wilkinson's Equation 6.11 [22], implicit in his exponent t_2. We also extend it to situations in which we may have underflow because, as one may expect after reading [6, 7, 9, 10, 21], the bounds above must be corrected in order to handle underflow or arithmetics without subnormal numbers. In the rest of the article we describe such corrections.

Equations (6) and (7) are simpler than Equation (4), but the proofs we present for them are definitely not. However, we hope that after our bounds are validated via the usual peer review process or by automated tools, people will be able to use them without reading their complicated proofs. For this reason we divided the article in three parts (besides this introduction.) In Section 2 we define the terms which allow us to treat floating point numbers as particular cases of real numbers. In Section 3 we illustrate the use of the definitions in Section 2 to derive sharper and simpler bounds for the effects on rounding errors in fundamental operations in floating point arithmetic, like sums, products, square roots and dot products.

Readers should focus on Section 3. It would be nice if they could find better proofs for the results stated in that section. In fact, we are glad that after we posted the first version of this manuscript M. Lange and S. Rump [20] derived stronger versions of some results presented here, using more direct arguments. This does not contradict the effectiveness of the use of continuous mathematics to analyze floating point arithmetic. Our point is that we can deduce the results thinking in continuous terms, and their formal proofs is just the last step in the discovery process.

In the last part of the article we prove our results. We try to handle all details in our proofs, and this makes them long and tedious. For this reason, we wrote two versions of the article. We plan to publish the long one, and the very long one will be available at arxiv.org. While reading any of these versions, we ask the reader not to underestimate how easily "short and intuitive" arguments about floating point arithmetic can be wrong. For example, in the appendix of our article [5] we argue that we can gain intuition about what would happen if we were to round upward instead of to nearest by replacing u by $2u$. This argument is correct in that context, because we verified each and every floating point operation in our computations. However, this intuitive argument is not rigorous in general. In fact, by replacing u by $2u$ in the bounds for rounding to nearest in the present article one will not obtain rigorous bounds for arithmetics which round upward or downward.

Finally, this extended version of the article is meant to be read using a software like the Adobe Acrobat Reader, so that you can click on the hyperlinks (anything in blue) and follow them. For example, the statement of our lemmas end with a blue triangle. By clicking on this triangle you will access the proof of the corresponding result, and by clicking on the "back button" you will return to the statement. Please, do use this feature of your reader in order to select which arguments to follow in more detail. Otherwise, you will find this article to be unbearably long.

2   Definitions

This section presents models of floating point arithmetic which extend the floating point operations to all real numbers. In the same way that one can use complex analysis to study integer arithmetic, and Sobolev spaces and distributions to learn about regular solutions of differential equations, by thinking of the set of floating point numbers as a subset of the set of real numbers we can use abstract arguments from optimization

theory, point set topology and convex analysis to reason about the floating point arithmetics implemented in real computers.

Most of our floating point numbers have the form $x = \pm\beta^e(\beta^\mu + r)$ where $\beta \in \{2, 3, 4, \dots\}$ is the base, e is an integer exponent, the exponent μ is a positive integer, and the remainder r is an integer in $[0, (\beta - 1)\beta^\mu)$. We also define zero as a floating point number and, finally, our models account for subnormal numbers $s = \pm\beta^e r$, for an integer $r \in [1, \beta^\mu)$. We now define floating point numbers more formally.

Definition 1 (Base) *A base is an integer greater than one.* ∎

Definition 2 (Unit roundoff) *The unit roundoff associated to the base β and the positive integer exponent μ is $u := u_{\beta,\mu} := 1/(2\beta^\mu)$ (We omit the subscript from $u_{\beta,\mu}$ when β and μ are evident given the context.)* ∎

The unit roundoff is our measure of rounding errors. It is equal to half the distance from 1 to the next floating point number. Some authors express their results in terms of ulps (units in the last place) or the machine precision, and our u correspond to half of the ulp or the machine epsilon used by them. However, the reader must be aware that there are conflicting definitions of these terms in the literature, and there is no universally accepted convention. A choice must be made, and we prefer to follow Higham [10] and use the unit roundoff u in Definition 2.

Our models are based on *floating point systems*, which are subsets of $\mathbb{R}$ to which we round real numbers. The simplest floating point systems are the perfect ones, which are defined below.

Definition 3 (Minus set) *For $\mathcal{A} \subset \mathbb{R}$, we define $-\mathcal{A} := \{-x, \text{ for } x \in \mathcal{A}\}$.* ∎

Definition 4 (Sign function) *The function $\text{sign} : \mathbb{R} \to \mathbb{R}$ is given by $\text{sign}(0) := 1$ and $\text{sign}(x) = |x|/x$ for $x \neq 0$, that is, we define the sign of 0 as one.* ∎

Definition 5 (Equally spaced range (E)) *The equally spaced range associated to the integer exponent e, the base β and the positive exponent μ is*

$$\mathcal{E}_{e,\beta,\mu} := \{\beta^e(\beta^\mu + r) \ \text{ for } \ r = 0, 1, 2, 3, \dots, (\beta - 1)\beta^\mu - 1\}$$

(We write simply $\mathcal{E}_e$ when β and μ are evident given the context.) ∎

Definition 6 (Perfect system) *The perfect floating point system associated to the base β and the positive exponent μ is*

$$\mathcal{P} := \mathcal{P}_{\beta,\mu} := \{0\} \bigcup \left(\bigcup_{e=-\infty}^{\infty} \mathcal{E}_{e,\beta,\mu} \right) \bigcup \left(\bigcup_{e=-\infty}^{\infty} -\mathcal{E}_{e,\beta,\mu} \right).$$
∎

Perfect floating point systems are convenient for proofs, but ignore underflow and overflow and are not practical. It is our opinion that the best compromise to handle overflow is to assume that it does not happen, that is, to formulate models which do not take overflow into account and shift the burden to handle overflow to the users of the model. This opinion is not due to laziness, but to the fact that verifying the absence of overflow in particular cases is simpler than dealing with floating point systems in which there is a maximum element.

Underflow is more subtle than overflow, and it may be difficult to avoid it even in simple cases. Therefore, handling underflow in each particular case would be too complicated, and it is a better compromise to have models that take underflow into account. Such models are formulated by limiting the range of the exponents e in Definition 6.

Definition 7 (MPFR system) *The MPFR system associated to the base β, the positive integer μ and the integer exponent $e_\alpha < -\mu$ is*

$$\mathcal{M} := \mathcal{M}_{e_\alpha,\beta,\mu} := \{0\} \bigcup \left(\bigcup_{e=e_\alpha}^{\infty} \mathcal{E}_{\beta,\mu} \right) \bigcup \left(\bigcup_{e=e_\alpha}^{\infty} -\mathcal{E}_{\beta,\mu} \right). \quad\blacksquare$$

The name MPFR is a tribute to the MPFR library [8], which has been very helpful in our studies of floating point arithmetic. This library does not use subnormal numbers, but allows for very wide exponent ranges (the minimal exponent is $1 - 2^{30}$ in the default configuration.) As a result, underflow is very unlikely and when it does happen its consequences are minimal.

Definition 8 (Subnormal numbers) *The set of positive subnormal numbers associated to the base β, the positive integer exponent μ and the integer exponent e_α is* $\mathcal{S}_{e_\alpha} := \mathcal{S}_{e_\alpha,\beta,\mu} := \{\beta^{e_\alpha} r, \text{ with } r = 1, 2, \ldots, \beta^\mu - 1\}.$ $\quad\blacksquare$

Definition 9 (IEEE system) *The IEEE system associated to the base β, the positive exponent μ and the integer exponent e_α, with $e_\alpha < -\mu$, is*

$$\mathcal{I} := \mathcal{I}_{e_\alpha,\beta,\mu} := \{0\} \bigcup \mathcal{S}_{e_\alpha,\beta,\mu} \bigcup -\mathcal{S}_{e_\alpha,\beta,\mu} \bigcup \left(\bigcup_{e=e_\alpha}^{\infty} \mathcal{E}_{e,\beta,\mu} \right) \bigcup \left(\bigcup_{e=e_\alpha}^{\infty} -\mathcal{E}_{e,\beta,\mu} \right).$$

The elements of $\mathcal{S}_{e_\alpha,\beta,\mu} \cup -\mathcal{S}_{e_\alpha,\beta,\mu}$ are the subnormal numbers for $\mathcal{I}$. $\quad\blacksquare$

The name IEEE is due to the IEEE 754 Standard for floating point arithmetic [11], which contemplates subnormal numbers.

Definition 10 (Floating point system) *There are three kinds of floating point systems:*

- *The perfect ones in Definition 6, which do not contain subnormal numbers.*

- *The unperfect ones, which can be either*

  - *The IEEE systems in Definition 9, which have subnormal numbers, or*
  - *The MPFR systems in Definition 7, which do not have subnormal numbers.*

For brevity, we refer to "the floating point system $\mathcal{F}$" as "the system $\mathcal{F}$," and throughout the article the letter $\mathcal{F}$ will always refer to a floating point system. $\quad\blacksquare$

Please pay attention to the technical detail that, in order to avoid pathological cases, our definitions require that $\beta \geq 2$ and $\mu > 0$, so that the mantissas of our floating point numbers have at least two bits and $u \leq 1/4$. Additionally, the minimum exponent e_α for unperfect systems is smaller than $-\mu$, so that 1 and $1/\beta$ are floating point numbers. By limiting the exponent range, we also limit the size of the smallest positive floating point numbers, which are quantified by the numbers α and ν below.

Definition 11 (Alpha) *For a perfect system we define $\alpha := 0$; the IEEE system $\mathcal{I}_{e_\alpha,\beta,\mu}$ has $\alpha := \beta^{e_\alpha}$, and $\alpha := \beta^{e_\alpha+\mu}$ for the MFPR system $\mathcal{M}_{e_\alpha,\beta,\mu}$ (Informally, the set of non negative elements of a system begins at α.)* $\quad\blacksquare$

Definition 12 (Nu) *For a perfect system we define $\nu := 0$ and the unperfect system $\mathcal{F}_{e_\alpha,\beta,\mu}$ has $\nu := \beta^{e_\alpha+\mu}$. (Informally, the **N**ormalized range for a system is formed by the numbers z with $|z| \geq \nu$, and ν is the Greek **N**.)* $\quad\blacksquare$

Definition 13 (Exponent for F) *Any integer e is an exponent for a perfect system, and $e \in \mathbb{Z}$ is an exponent for the unperfect system $\mathcal{F}_{e_\alpha}$ if $e \geq e_\alpha$.* ∎

This article is about rounding to nearest, as we now formalize.

Definition 14 (Rounding to nearest) *A function $\mathrm{fl} : \mathbb{R} \to \mathbb{R}$ rounds to nearest in the floating point system $\mathcal{F}$ if $\mathrm{fl}(z) \in \mathcal{F}$ and $|\mathrm{fl}(z) - z| \leq |x - z|$ for $x \in \mathcal{F}$ and $z \in \mathbb{R}$.* ∎

Definition 15 (Breaking ties) *When fl rounds to nearest in $\mathcal{F}$, we say that fl breaks ties downward if, for $x \in \mathcal{F}$ and $z \in \mathbb{R}$, $|x - z| = |\mathrm{fl}(z) - z| \Rightarrow x \geq \mathrm{fl}(z)$. Similarly, fl breaks ties upward if $|x - z| = |\mathrm{fl}(z) - z| \Rightarrow x \leq \mathrm{fl}(z)$.* ∎

We now model the numerical sum $\mathrm{fl}(\sum_{k=0}^{n} y_k)$ of $n + 1$ real numbers. For technical reasons, it is important to allow for the use of different rounding functions in the evaluation of the partial sums $s_k = \left(\sum_{i=0}^{k-1} y_i\right) + y_k$. With this motivation, we state the last definitions in this section.

Definition 16 (Rounding tuples) *A tuple of functions $\mathrm{Fl} = \{\mathrm{fl}_1, \ldots, \mathrm{fl}_n\}$ rounds to nearest in $\mathcal{F}$ if its elements round to nearest in $\mathcal{F}$. In this case we say that Fl is a rounding n-tuple, n is Fl's dimension and $\mathcal{F}$ is Fl's range.* ∎

Definition 17 (Projection) *Let $\mathcal{A}$ be a set and $\mathcal{A}^n$ the Cartesian product $\mathcal{A} \times \cdots \times \mathcal{A}$ with n factors. For $k = 1, \ldots, n$, we define $\mathrm{P}_k : \mathcal{A}^n \to \mathcal{A}^k$ as the projection on the first k coordinates, that is $\mathrm{P}_k(x_1, \ldots, x_n) := (x_1, \ldots, x_k)$. When $\mathcal{A}$ is a vector space with zero element $\mathbf{0}$, we define $P_0 : \mathcal{A}^n \to \{\mathbf{0}\}$ as $P_0(x_1, \ldots, x_n) := \mathbf{0}$.* ∎

Definition 18 (Floating point sum) *Let $\mathcal{R}$ be the set of all functions from $\mathbb{R}$ to $\mathbb{R}$, and f_0 its zero element. We define $S_0 : \{0\} \times \{f_0\} \to \mathbb{R}$ as $S_0(0, f_0) := 0$. For $n > 0$ we define $S_n : \mathbb{R}^n \times \mathcal{R}^n \to \mathbb{R}$ recursively as $S_n(\mathbf{z}, \mathrm{Fl}) := \mathrm{fl}_n(S_{n-1}(\mathrm{P}_{n-1}\mathbf{z}, \mathrm{P}_{n-1}\mathrm{Fl}) + z_n)$.* ∎

As a convenient notation, given a rounding n-tuple Fl we write

$$\mathrm{Fl}\left(\sum_{k=0}^{n} x_k\right) := S_n((x_0 + x_1, x_2, x_3, x_4, \ldots, x_n), \mathrm{Fl}),$$

and when $\mathrm{Fl} = \{\mathrm{fl}, \mathrm{fl}, \ldots, \mathrm{fl}\}$ has all its elements equal to fl we write

$$\mathrm{fl}\left(\sum_{k=0}^{n} x_k\right) := \mathrm{Fl}\left(\sum_{k=0}^{n} x_k\right).$$

We ask the reader to forgive us for the inconsistency in these expressions: neither $\mathrm{Fl}(\sum_{k=0}^{n} x_k)$ nor $\mathrm{fl}(\sum_{k=0}^{n} x_k)$ is the value of a function $\mathrm{Fl}(s)$ at $s = \sum_{k=0}^{n} x_k$, but rather the value obtained by rounding the partial sums using the elements of Fl. Note also that $\mathrm{Fl}(\sum_{k=0}^{n} x_k)$ is defined in terms of $x_0 + x_1$, that is, the first term in the sum is treated differently from the others. The same detail is present in Equation (2), in which $x_1 y_1$ and $x_2 y_2$ are treat differently from the other terms.

We emphasize that we define "the floating point sum of $n + 1$ real numbers", and not the "the sum of $n + 1$ floating point numbers." As a result, our rounded sums apply to all real numbers, not only to the ones in the system $\mathcal{F}$, in the spirit of the first paragraph of this section. Dot products are similar to sums:

6

Definition 19 (Dot product) *The dot product of the vectors* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{n+1}$ *evaluated with the rounding tuples* $\mathrm{Fl} = \{\mathrm{fl}_1, \ldots, \mathrm{fl}_n\}$ *and* $\mathrm{R} = \{\mathrm{r}_0, \ldots, \mathrm{r}_n\}$ *is*

$$\mathrm{dot}_{\mathrm{Fl},\mathrm{R}} \left(\sum_{k=0}^{n} x_k y_k \right) := \mathrm{Fl} \left(\sum_{k=0}^{n} \mathrm{r}_k (x_k y_k) \right) \qquad\blacksquare$$

We also analyze dot products evaluated with the fused multiply add operations available in modern hardware and programming languages:

Definition 20 (Fma dot product) *The fma dot product of the vectors* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{n+1}$ *evaluated with the rounding tuple* $\mathrm{Fl} = \{\mathrm{fl}_0, \ldots, \mathrm{fl}_n\}$ *is*

$$\mathrm{fma}_{\mathrm{Fl}} \left(\sum_{k=0}^{n} x_k y_k \right) := S_{n+1}((x_0 y_0, x_1 y_1, x_2 y_2, \ldots, x_n y_n), \mathrm{Fl}). \qquad\blacksquare$$

3 Sharp error bounds

This section presents sharper versions of the $(1 + \varepsilon)$ argument. In summary, we argue that when rounding to nearest with unit roundoff u, in many situations we can use

$$\varepsilon = \frac{u}{1+u} \tag{10}$$

in the $(1 + \varepsilon)$ argument, and this value is better than u or $u/(1-u)$. The section has four parts. The first part describes the advantages of the ε in Equation (10) when dealing with a few floating point operations. The next one generalizes our results to sums of many numbers, by proving the bound (5). Section 3.3 presents bounds on the errors in sums which are expressed in terms of $\sum_{k=1}^{n} \left| \sum_{i=0}^{k} x_i \right|$. Section 3.4 is about dot products. It shows that by working with real numbers from the start it is easy to adapt results derived for sums in order to obtain bounds for the errors in dot products.

3.1 Basics

This section is about the $(1 + \varepsilon)$ argument for a few floating point operations. When rounding a floating point number, our first lemma states that the ε in Equation (10) can be used when the real number z is in the normal range, ie., the absolute value of z is greater than the number v in Definition 12.

Lemma 1 (A better epsilon) *If* fl *rounds to nearest in* $\mathcal{F}$ *and* $|z| \geq v_{\mathcal{F}}$ *then*

$$|\mathrm{fl}(z) - z| \leq \frac{|z| u}{1 + u}. \tag{11}$$

In particular, if $\mathcal{F}$ *is perfect then Equation (11) holds for all* $z \in \mathbb{R}$. $\blacktriangle$

Lemma 1 is sharp in the sense that for any ε smaller than $u/(1+u)$ there exists a real number z near $1+u$ for which Equation (11) does not hold. It has been known for a long time [14], but it leads to bounds slightly stronger than the ones in [10] for instance, because

$$\frac{u}{1+u} < u < \frac{u}{1-u}$$

and when the result of the operation $x\,\mathrm{op}\,y \neq 0$ is in the normal range we have the bound

$$\frac{1}{1+u} \leq \frac{\mathrm{fl}(x\,\mathrm{op}\,y)}{x\,\mathrm{op}\,y} \leq \frac{1+2u}{1+u}, \tag{12}$$

instead of the usual bound

$$1 - u \leq \frac{\mathrm{fl}(x\,\mathrm{op}\,y)}{x\,\mathrm{op}\,y} \leq \frac{1}{1-u}. \tag{13}$$

As a result, we could use the same argument as Higham to conclude that in a perfect floating point system

$$\mathrm{Fl}\left(\sum_{k=0}^{n} x_k\right) = (x_0 + x_1)\,\xi_0^n + \sum_{i=2}^{n} x_k \xi_k^{n-i+1} \quad \text{with} \quad \frac{1}{1+u} \leq \xi_k \leq \frac{1+2u}{1+u} \tag{14}$$

and

$$\mathrm{Fl}\left(\sum_{k=0}^{n} x_k y_k\right) = x_0 y_0 \xi_0^{n+1} + \sum_{i=1}^{n} x_k y_k \xi_k^{n-i+2} \quad \text{with} \quad \frac{1}{1+u} \leq \xi_k \leq \frac{1+2u}{1+u}.$$

The underlying reason as to why

$$f(u) := \frac{1+2u}{1+u} \quad \text{is a better upper bound than} \quad h(u) := \frac{1}{1-u} \tag{15}$$

is the difference between concavity and convexity. The function $f_\tau(x) := f(u)^\tau$ has second derivative

$$f_\tau''(u) = \frac{\tau f_{\tau-2}(u)}{(1+u)^4}(\tau - 3 - 4u)$$

and is concave for $\tau \leq 3 + 4u$. On the other hand, $h_\tau(x) := h(u)^\tau$ has second derivative

$$h_\tau''(u) = \tau(\tau+1)h_{\tau-2}(u)$$

and is convex for all $\tau > 0$ and $0 < u < 1$. As a result, we can linearize rigorously an upper bound based on f_τ, with $0 \leq \tau \leq 3 + 4u$, whereas linearizing an upper bound based on h_τ is correct only to leading order. For instance, using the bound (12) we can prove the next corollary, and similar results combining multiplications and divisions, but we could not prove such results based only on the usual bound (13).

Corollary 1 (Three products) *Let x, y, z and w be real numbers. If $\hat{p}_1 := \mathrm{fl}(x * y)$, $\hat{p}_2 := \mathrm{fl}(\hat{p}_1 * z)$ and $\hat{p}_3 := \mathrm{fl}(\hat{p}_2 * w)$, $p_i \neq 0$, and $|p_i|$ satisfy Equation (11) for $k = 1$, 2 and 3 then*

$$1 - ku \leq \frac{\hat{p}_k}{p_k} \leq 1 + ku,$$

for $p_1 := x * y$, $p_2 := x * y * z$ and $p_3 := x * y * z * w$. ▲

Lemma 1 also yields a simple proof of a well known result about square roots when $\beta = 2$ [4, 12], and solves an open problem for arbitrary bases β [2]:

Corollary 2 (Square roots) *For the base $\beta = 2$, if $x \in \mathcal{F}$ is such that $x^2 \geq v$ and $\mathrm{fl}(x^2)$ and $\mathrm{fl}\left(\sqrt{\mathrm{fl}(x^2)}\right)$ are evaluated rounding to nearest then $\mathrm{fl}\left(\sqrt{\mathrm{fl}(x^2)}\right) = |x|$. Moreover,*

$$\mathrm{fl}\left(\frac{|x|}{\mathrm{fl}\left(\sqrt{\mathrm{fl}(x^2)}\right)}\right) \leq 1 \tag{16}$$

for a general base β, under the same hypothesis on fl and x. ▲

The next two lemmas show that there are other conditions besides $|z| \geq v$ in which we can use the bound in Equation (11):

Lemma 2 (Exact sums) *If $x, y \in \mathcal{F}$ are such that $\alpha \leq |x+y| \leq \beta v$ then $z := x+y \in \mathcal{F}$, that is, the sum $x+y$ is exact. In particular, z satisfies Equation (11).* ▲

Lemma 3 (IEEE sums) *Let $\mathcal{I}$ be an IEEE system and $x, y \in \mathcal{I}$. If $0 < |x+y| \leq \beta v$ then $|x+y| \geq \alpha$, and $z := x+y \in \mathcal{I}$ and satisfies Equation (11).* ▲

The last two lemmas combined with Lemma 1 imply that we can use the bound (11) for every real number z which is the sum of two floating point numbers in an IEEE system. This is yet another instance in which subnormal numbers lead to simpler results, and corroborates Demmel's arguments [6, 7] and the soundness of the decision to include subnormal numbers in the IEEE standard for floating point arithmetic [11]. Another instance is the fundamental Sterbenz's Lemma, which must be modified by the inclusion of the term α in its hypothesis in order to hold for MPFR systems:

Lemma 4 (Sterbenz's Lemma) *If $a, b \in \mathcal{F}$ and $\alpha \leq b-a \leq a$ then $b-a \in \mathcal{F}$.* ▲

3.2   Norm one bounds

This subsection extends Lemma 1 to sums with many parcels. Our results are described by the next lemma and its corollaries. In particular, we show that underflow does not affect sums of positive numbers. Therefore, there is no need for terms involving the smallest positive floating point number when bounding the errors in such sums.

Lemma 5 (Norm one bound) *If $\mathrm{Fl} = \{\mathrm{fl}_1, \ldots, \mathrm{fl}_n\}$ rounds to nearest in a perfect system, $20nu \leq 1$ and $y_0, \ldots y_n \in \mathbb{R}$ then*

$$\left| \mathrm{Fl}\left(\sum_{k=0}^{n} y_k \right) - \sum_{k=0}^{n} y_k \right| \leq \frac{nu}{1+nu} \sum_{k=0}^{n} |y_k|. \tag{17}$$

▲

Corollary 3 (IEEE norm one bound) *If $\mathrm{Fl} = \{\mathrm{fl}_1, \ldots, \mathrm{fl}_n\}$ rounds to nearest in an IEEE system $\mathcal{I}$, $20nu \leq 1$ and $y_0, \ldots, y_n \in \mathcal{I}$ then Equation (17) is satisfied.* ▲

Corollary 4 (MPFR norm one bound) *If Fl rounds to nearest in a MPFR system $\mathcal{M}$, $20nu \leq 1$, $\mathbf{y} \in \mathcal{M}^{n+1}$ and $y_k \geq 0$ for all k then Equation (17) holds.* ▲

Corollary 5 (Unperfect norm one bound) *If $\mathrm{Fl} = \{\mathrm{fl}_1, \ldots, \mathrm{fl}_n\}$ rounds to nearest in an unperfect system, $y_0, \ldots, y_n \in \mathbb{R}$ and $20nu \leq 1$ then*

$$\left| \mathrm{Fl}\left(\sum_{k=0}^{n} y_k \right) - \sum_{k=0}^{n} y_k \right| \leq \frac{n\alpha}{2} + \frac{nu}{1+nu} \left(\frac{n\alpha}{2} + \sum_{k=0}^{n} |y_k| \right). \tag{18}$$

If, additionally, $u \sum_{k=0}^{n} |y_k| \geq \alpha$ then Equation (17) is satisfied. ▲

Note that Corollaries 3 and 4 have different hypothesis regarding the floating point numbers $y_0, \ldots, y_n$: in the IEEE case, in which we have subnormal numbers, Equation (17) holds for all such y_k. In the MPFR case, due to the absence of subnormal numbers, we must assume that $y_k \geq 0$, for Equation (17) does not hold for instance when $\beta = 2$, $x_0 = 3\alpha/2$, $x_1 = -\alpha$, $n = 1$ and we break ties upward. Note also that the number α in Equation (18) for an IEEE system is much smaller than the α for the corresponding MPFR system. The next example shows that the bound (17) is sharp:

Example 1 (The norm one bound is sharp) *If* fl *rounds to nearest in the perfect system* $\mathcal{P}_{2,\mu}$, *breaking ties downward, and* $x_0 := 1$ *and* $x_k := u$ *for* $k = 1, \ldots, n$ *then*

$$\text{fl}\left(\sum_{k=0}^{n} x_k\right) = 1 = \sum_{k=0}^{n} x_k - nu = \sum_{k=0}^{n} x_k - \frac{nu}{1+nu} \sum_{k=0}^{n} x_k.$$

If fl *breaks ties upward for the same* x_k *and* $2nu < 1$ *then*

$$\text{fl}\left(\sum_{k=0}^{n} x_k\right) = 1 + 2nu = \sum_{k=0}^{n} x_k + nu = \sum_{k=0}^{n} x_k + \frac{nu}{1+nu} \sum_{k=0}^{n} x_k. \qquad \blacksquare$$

As in Lemma 1, the bound (17) has concavity properties which allow us to linearize rigorously bounds resulting from a couple of its applications:

Lemma 6 (Convexity) *For* $k \in \mathbb{N}$ *and* $i = 1, \ldots, k$, *let* n_i *be a positive number and define functions* $f_k, g_k : (0, \infty) \to \mathbb{R}$ *by*

$$f_k(u) = \prod_{i=1}^{k} \frac{1+2n_i u}{1+n_i u} \qquad \text{and} \qquad g_k(u) = \prod_{i=1}^{k} \frac{1}{1+n_i u}.$$

The functions f_k *are strictly concave for* $k = 1$, 2 *and* 3 *and the functions* g_k *are convex for all k. In particular, for* $k \le 3$,

$$1 - \left(\sum_{i=1}^{k} n_i\right) u \le g_k(u) \le f_k(u) \le 1 + \left(\sum_{i=1}^{k} n_i\right) u. \qquad \blacktriangle$$

As a final point for this section, we note that Lemma 5 implies that

$$\left|\text{Fl}\left(\sum_{k=0}^{n} y_k\right) - \sum_{k=0}^{n} y_k\right| \le \frac{n(n+1)u}{1+nu} \max_{k=0,\ldots,n} |y_k|, \qquad (19)$$

and it is natural to ask whether the quadratic term in n in the right hand side of Equation (19) is necessary. The next example shows that bounds in terms of $\max |y_k|$ do need a quadratic term in n (or large constant factors):

Example 2 (Quadratic growth) *If* fl *rounds to nearest in the perfect system* $\mathcal{P}_{2,\mu}$, *breaking ties downward,* $y_0 := 1 + u$ *and* $y_k := 1 + 2^{\lfloor \log_2(k+1) \rfloor} u$ *for* $k = 1, \ldots, n :=$ $2^m - 1$, *where* $m \in \mathbb{N}$ *is such that* $2^m u < 1$, *then*

$$\text{fl}\left(\sum_{k=0}^{n} y_k\right) = \sum_{k=0}^{n} y_k - \frac{n^2+2n+3}{3} u \le \sum_{k=0}^{n} y_k - \frac{n^2+2n+3}{6} u \max_{k=0,\ldots,n} |y_k|. \qquad \blacktriangle$$

3.3 Cumulative bounds

Although Lemma 5 leads to the simple bound (6) on the error in the evaluation of sums, it is not as good from the qualitative view as the result one would obtain from the version of Higham's Equation 3.4 for sums, or from our Equation (2). We believe that Higham and Wilkinson would write this equation as

$$\text{fl}\left(\sum_{k=1}^{n} x_k\right) = (x_1 + x_2)(1+\theta_{n-1}) + x_3(1+\theta_{n-2}) + \cdots + x_n(1+\theta_1). \qquad (20)$$

10

for θ_n in Equation (3). In fact, Wilkinson presents an expression similar to Equation (20) for sums using a double precision accumulator in page 117 of [22]. Equation (20) gives a better intuition regarding the effects of rounding errors in the corresponding sum than the bound in Lemma 5. Therefore, it is natural to look for bounds that take into account the stronger relative influence of the first parcels in Equation (20). The next examples are relevant in this context.

Example 3 (Minimum cumulative bound) *If* fl *rounds to nearest in the perfect system* $\mathcal{P}_{2,\mu}$, *breaking ties downward, and* $x_k := u^{-k}$ *for* $k = 1,\dots,n$ *then*

$$\mathrm{fl}\left(\sum_{k=0}^{n} x_k\right) = u^{-n} = \sum_{k=0}^{n} x_k - \kappa_n u \sum_{k=1}^{n} \sum_{i=0}^{k} x_i$$

for

$$1 - u < \kappa_n := \frac{(1-u)(1-u^n)}{1 - u^n - nu^{n+1}(1-u)} < (1-u)(1+u^n) < 1. \qquad \blacktriangle$$

Example 4 (Maximum cumulative bound) *If* fl *rounds to nearest in the perfect system* $\mathcal{P}_{\beta,\mu}$, *breaking ties upward,* $1 = e_1 < e_2 \cdots < e_n$ *are integer exponents,* $x_0 := u$, $x_1 := 1$, *and* $x_k := \beta^{e_k}(1+u) - \beta^{e_{k-1}}(1+2u)$ *for* $k = 2,\dots,n$ *then*

$$\mathrm{fl}\left(\sum_{k=0}^{n} x_k\right) \le \sum_{i=0}^{n} x_i + \tau_n u \sum_{k=1}^{n} \sum_{i=0}^{k} x_k, \qquad (21)$$

for

$$\tau_n := \frac{1}{1 + u\left(\frac{\beta-2}{\beta-1} + \frac{n}{\beta^n-1}\right)}.$$

Additionally, if $e_k = k - 1$ *for* $k \ge 1$ *then we have equality in Equation (21).* $\blacktriangle$

These examples indicate that there is an asymmetry between the upper and lower bounds on the errors $\delta := \mathrm{fl}(\sum_{k=0}^{n} x_k) - \sum_{i=0}^{n} x_i$ in terms of $\sum_{k=1}^{n} \sum_{i=0}^{k} x_k$: The constants κ_n in Example 3 and τ_n in Example 4 are equal to $1/(1+u)$ for $n = 1$ but as n increases, κ_n decreases toward $1 - u$ whereas τ_n increases toward 1. Therefore, the worst lower and upper values for δ are reached in different situations, and are due to distinct causes. In fact, the lower bound for δ in the next Lemma is a straightforward consequence of the convexity of the functions $(1+u)^{-k}$ and Equation (14), whereas the upper bound is a non trivial consequence of the concavity of $(1+2u)/(1+u)$.

Lemma 7 (Positive cumulative bound) *If* $\mathrm{Fl} = \{\mathrm{fl}_1,\dots,\mathrm{fl}_n\}$ *rounds to nearest in a perfect system,* $y_0, y_1, \dots y_n \in \mathbb{R}$, *with* $y_k \ge 0$ *for* $k = 0,\dots,n$, *and* $20nu \le 1$ *then*

$$-\frac{u}{1+u} \sum_{k=1}^{n} \sum_{i=0}^{k} y_k \le \mathrm{Fl}\left(\sum_{k=0}^{n} y_k\right) - \sum_{k=0}^{n} y_k \le \tau_n u \sum_{k=1}^{n} \sum_{i=0}^{k} y_k, \qquad (22)$$

for τ_n *in Example 4.* $\blacktriangle$

Corollary 6 (Unperfect cumulative bound) *If* Fl *rounds to nearest in an unperfect system* $\mathcal{F}$, $20nu \le 1$, $\mathbf{y} \in \mathcal{F}^{n+1}$ *and* $y_k \ge 0$, *then Equation (22) holds.* $\blacktriangle$

The next example shows that Lemma 7 does not apply to sums of numbers with mixed signs, and Lemma 8 and its corollary show that the example is nearly worst possible.

11

Example 5 (Mixed signs) *If* fl *rounds to nearest in a perfect system* $\mathcal{P}_{2,\mu}$, *breaking ties upward,* $x_0 := u$, $x_1 := 1$, $x_k := -2^{1-k}(1+3u)$ *for* $k > 1$ *and* $2^n u \leq 1$ *then*

$$\mathrm{fl}\left(\sum_{k=0}^{n} x_k\right) - \sum_{k=0}^{n} x_k = 2\left(1 - 2^{-n}\right)u = \frac{\kappa_n u}{1 - (n-2)u} \sum_{k=1}^{n}\left|\sum_{i=0}^{k} x_i\right|, \tag{23}$$

for

$$1 - u \leq \kappa_n := \frac{\left(1 - 2^{-n}\right)\left(1 - (n-2)u\right)}{\left(1 - 2^{-n}\right)\left(1 + 3u\right) - nu} \leq 1. \qquad \blacktriangle$$

Lemma 8 (Signed cumulative bound) *If* $\mathrm{Fl} = \{\mathrm{fl}_1, \ldots, \mathrm{fl}_n\}$ *rounds to nearest in a perfect system,* $y_0, y_1, \ldots y_n \in \mathbb{R}$ *and* $20nu < 1$ *then*

$$\left|\mathrm{Fl}\left(\sum_{k=0}^{n} y_k\right) - \sum_{k=0}^{n} y_k\right| \leq \frac{u}{1 - (n-2)u} \sum_{k=1}^{n}\left|\sum_{i=0}^{k} y_i\right|. \tag{24}$$

$\blacktriangle$

Corollary 7 (Unperfect signed cumulative bound) *If* $\mathrm{Fl} = \{\mathrm{fl}_1, \ldots, \mathrm{fl}_n\}$ *rounds to nearest in an unperfect system,* $y_0, y_1, \ldots y_n \in \mathbb{R}$ *and* $20nu \leq 1$ *then*

$$\left|\mathrm{Fl}\left(\sum_{k=0}^{n} y_k\right) - \sum_{k=0}^{n} y_k\right| \leq (1 + 2nu)\frac{n\alpha}{2} + \frac{u}{1 - (n-2)u} \sum_{k=1}^{n}\left|\sum_{i=0}^{k} y_i\right|, \tag{25}$$

If, additionally, $u\sum_{k=1}^{n}\left|\sum_{i=0}^{k} y_i\right| \geq n\alpha$ *then*

$$\left|\mathrm{Fl}\left(\sum_{k=0}^{n} y_k\right) - \sum_{k=0}^{n} y_k\right| \leq \frac{3}{2}\left(1 + \frac{nu}{2}\right)u\sum_{k=1}^{n}\left|\sum_{i=0}^{k} y_i\right|. \tag{26}$$

$\blacktriangle$

3.4 Dot products

This section presents bounds on the errors in the numerical evaluation of dot products. These bounds are derived from the ones for sums presented in Section 3.2. This derivation is possible because some of our previous bounds apply to general real numbers, and a numerical dot product is simply a numerical sum of real numbers, which may or may not be floating point numbers. If our analysis of sums were restricted to floating point numbers then the extensions presented here would be harder to derive. For example, the next corollaries follow directly from Lemma 5 and the Definition 20 of numerical dot products using fused multiply adds (these corollaries are proved in the extended version of the article):

Corollary 8 (Dot prod. with fma) *If* $\mathrm{Fl} = \{\mathrm{fl}_0, \ldots, \mathrm{fl}_n\}$ *rounds to nearest in a perfect system,* $20nu \leq 1$ *and* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{n+1}$ *then*

$$\left|\mathrm{fma}_{\mathrm{Fl}}\left(\sum_{k=0}^{n} x_k y_k\right) - \sum_{k=0}^{n} x_k y_k\right| \leq \frac{(n+1)u}{1 + (n+1)u} \sum_{k=0}^{n}|x_k y_k|. \tag{27}$$

$\blacktriangle$

Corollary 9 (Unperfect Dot prod. with fma) *If* $Fl = \{fl_0, \ldots, fl_n\}$ *rounds to nearest in an unperfect system,* $20nu \leq 1$ *and* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{n+1}$ *then*

$$\left| \text{fma}_{Fl} \left(\sum_{k=0}^{n} x_k y_k \right) - \sum_{k=0}^{n} x_k y_k \right| \leq (n+1) \frac{\alpha}{2}$$

$$+ \frac{(n+1)u}{1+(n+1)u} \left(\frac{(n+1)\alpha}{2} + \sum_{k=0}^{n} |x_k y_k| \right). \tag{28}$$

If, additionally, $u \sum_{k=0}^{n} |x_k y_k| \geq \alpha$ *then Equation (27) holds.* ▲

When we evaluate dot products rounding each product $x_k y_k$, the bounds are slightly worse, but can still be obtained with the theory in Section 3.2:

Corollary 10 (Dot prod.) *If* $Fl = \{fl_1, \ldots, fl_n\}$ *and* $R = \{r_0, \ldots, r_n\}$ *round to nearest in a perfect system,* $20nu \leq 1$ *and* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{n+1}$ *then*

$$\left| \text{dot}_{Fl,R} \left(\sum_{k=0}^{n} x_k y_k \right) - \sum_{k=0}^{n} x_k y_k \right| \leq \beta_n u \sum_{k=0}^{n} |x_k y_k| \leq (n+1)u \sum_{k=0}^{n} |x_k y_k|, \tag{29}$$

where

$$\beta_n := \frac{n+1+3nu}{1+(n+1)u+nu^2} \leq \frac{n+1}{1+nu/2} \qquad \text{and} \qquad \beta_n \leq \frac{n+1}{1+(n-3)u}. \quad ▲$$

Corollary 11 (IEEE dot prod.) *If* $Fl = \{fl_1, \ldots, fl_n\}$ *and* $R = \{r_0, \ldots, r_n\}$ *round to nearest in an IEEE system,* $20nu \leq 1$ *and* $\mathbf{x}, \mathbf{z} \in \mathbb{R}^{n+1}$ *then*

$$\left| \text{dot}_{Fl,R} \left(\sum_{k=0}^{n} x_k y_k \right) - \sum_{k=0}^{n} x_k y_k \right| \leq 1.05 (n+1) \frac{\alpha}{2} + \beta_n u \sum_{k=0}^{n} |x_k y_k|, \tag{30}$$

for β_n *in Corollary 10. If, additionally,* $u \sum_{k=0}^{n} |x_k y_k| \geq \alpha$ *then*

$$\left| \text{dot}_{Fl,R} \left(\sum_{k=0}^{n} x_k y_k \right) - \sum_{k=0}^{n} x_k y_k \right| \leq \frac{3}{2} (n+1)u \sum_{k=0}^{n} |x_k y_k|. \quad ▲$$

Corollary 12 (MPFR dot prod.) *If* $Fl = \{fl_1, \ldots, fl_n\}$ *and* $R = \{r_0, \ldots, r_n\}$ *round to nearest in a MPFR system,* $20nu \leq 1$ *and* $\mathbf{x}, \mathbf{z} \in \mathbb{R}^{n+1}$ *then*

$$\left| \text{dot}_{Fl,R} \left(\sum_{k=0}^{n} x_k y_k \right) - \sum_{k=0}^{n} x_k y_k \right| \leq \frac{(2.05n+1.05)\alpha}{2} + \beta_n \sum_{k=0}^{n} |x_k y_k|, \tag{31}$$

for β_n *in Corollary 10. If, additionally,* $u \sum_{k=0}^{n} |x_k y_k| \geq \alpha$ *then the last equation in Corollary 11 is satisfied.* ▲

Finally, in all bounds above we can use the Cauchy-Schwarz inequality and replace the terms $\sum_{k=0}^{n} |x_k y_k|$ by $\|\mathbf{x}\|_2 \|\mathbf{y}\|_2$. With this replacement, we can compare our bounds to the ones in [6] and [21].

4 Proofs

This section we prove our main results. Section 4.1 contains more definitions and Section 4.2 presents more lemmas. In Section 4.3 we state basic results about floating point systems and rounding to nearest. We call such results by "Propositions," because they are obvious and readers should be able to deduce them with little effort. Section 4.4 begins with the proofs of the main lemmas, and after that we prove some of the corollaries. The extended version of the article contains the proofs of the remaining lemmas and corollaries and the propositions, and the verification of the examples.

4.1 More definitions

The proofs of our bounds on the errors in sums use the following definitions:

Definition 21 (Tight function) *Let $\mathcal{A}$ and $\mathcal{B}$ be topological spaces and $\mathcal{R}$ a set. A function $f : \mathcal{A} \times \mathcal{R} \to \mathcal{B}$ is tight if for every sequence $\{(a_k, r_k), k \in \mathbb{N}\} \subset \mathcal{A} \times \mathcal{R}$ such that $\lim_{k \to \infty} a_k$ there exists $r \in \mathcal{R}$ and a subsequence $\left\{\left(a_{n_k}, r_{n_k}\right), k \in \mathbb{N}\right\}$ with $\lim_{k \to \infty} f\left(a_{n_k}, r_{n_k}\right) = f(a, r)$.* ∎

Definition 22 (Tight set of functions) *Let $\mathcal{A}$ and $\mathcal{B}$ be topological spaces and let $\mathcal{R}$ be a set of functions from $\mathcal{A}$ to $\mathcal{B}$. We say that $\mathcal{R}$ is tight if the function $f : \mathcal{A} \times \mathcal{R} \to \mathcal{B}$ given by $f(a, r) := r(a)$ is tight.* ∎

4.2 More lemmas

Lemma 9 (Sharp epsilons) *Suppose* fl *rounds to nearest in $\mathcal{F}$ and e is an exponent for $\mathcal{F}$. If $|z| = \beta^e (\beta^\mu + w)$ with $w \in [0, (\beta - 1)\beta^\mu]$ then*

$$\mathrm{fl}(z) = \mathrm{sign}(z)\beta^e(\beta^\mu + r) \quad \text{for} \quad r \in [0, (\beta - 1)\beta^\mu] \cap \mathbb{Z} \quad \text{and} \quad |r - w| \le 1/2. \quad (32)$$

Moreover,

$$\left| \frac{\mathrm{fl}(z) - z}{z} \right| \le \frac{u}{1 + \max\{1, 2w\}u} \le \frac{u}{1 + u} \quad (33)$$

and

$$\left| \frac{\mathrm{fl}(z) - z}{z} \right| \le \frac{u}{1 + (2r - 1)u} \quad \text{and} \quad \left| \frac{\mathrm{fl}(z) - z}{\mathrm{fl}(z)} \right| \le \frac{u}{1 + 2ru}. \quad \blacktriangle$$

Lemma 10 (Compactness) *Let $\mathcal{R}$ be a set, $\mathcal{L} \subset \mathbb{R} \setminus \{0\}$, $\mathcal{A}, \mathcal{B} \subset \mathbb{R}^n$, $\mathcal{X} \subset \mathbb{R}^m$ and $\mathcal{K} \subset \mathcal{A}$. Define $\mathcal{Z} := \mathcal{A} \cup \mathcal{B}$. If the functions $f : \mathcal{Z} \times \mathcal{X} \to \mathbb{R}$, $h : \mathcal{Z} \times \mathcal{R} \to \mathcal{X}$, and $g(\mathbf{z}, r) := f(\mathbf{z}, h(\mathbf{z}, r))$ and $\varphi \in \mathbb{R}$ are such that*

- *$\mathcal{K}$ is compact and for $\mathbf{z} \in \mathcal{A}$ there exist $\lambda \in \mathcal{L}$ such that $\lambda \mathbf{z} \in \mathcal{K}$.*

- *If $\lambda \in \mathcal{L}$, $\mathbf{z} \in \mathcal{A}$, $\lambda \mathbf{z} \in \mathcal{K}$ and $r \in \mathcal{R}$ then $h(\lambda \mathbf{z}, r') = \lambda h(\mathbf{z}, r)$ for some $r' \in \mathcal{R}$.*

- *f is upper semi-continuous and $f(\lambda \mathbf{z}, \lambda \mathbf{x}) \ge f(\mathbf{z}, \mathbf{x})$ for $\mathbf{z} \in \mathcal{A}$ and $\lambda \in \mathcal{L}$.*

- *h is tight, in the sense of Definition 21.*

- *$g(\mathbf{z}, r) \le \varphi$ for $(\mathbf{z}, r) \in \mathcal{B} \times \mathcal{R}$.*

then either $g(\mathbf{z}, r) \le \varphi$ for all $(\mathbf{z}, r) \in \mathcal{Z} \times \mathcal{R}$ or there exist $(\mathbf{z}^*, r^*) \in \mathcal{K} \times \mathcal{R}$ such that $g(\mathbf{z}^*, r^*) \ge g(\mathbf{z}, r)$ for all $(\mathbf{z}, r) \in \mathcal{Z} \times \mathcal{R}$. ▲

Lemma 10 is a compactness argument. Its purpose is to show that either there exists examples for which the relative effects of rounding errors are the worst possible or these errors are small. It is necessary because floating point systems are infinite and we cannot take this existence for granted. The intuition behind Lemma 10 is simple. The vector $\mathbf{z}$ represents the input to computation. The vector $h(\mathbf{z}, r)$ is obtained by rounding functions of $\mathbf{z}$ using the rounding functions $r \in \mathcal{R}$. The bad set $\mathcal{B}$ represents situations like underflow or very poor scaling, and its elements are handled separately. For $\mathbf{z}$ outside of the bad set, we can use scaling by powers of β (represented by $\lambda \in \mathcal{L}$) to reduce the analysis of $f(\mathbf{z}, \mathbf{x})$ to real numbers $\mathbf{z}$ in the compact set $\mathcal{K}$. We can then deal with the discontinuity in rounding by analyzing all functions which round to nearest (represented by $\mathcal{R}$) instead of a single function. In the end, as in the applications of the classic Banach-Alaoglu Theorem, by using compactness and continuity in their full generality, we can analyze the existence of maximizers for the relative effects of rounding errors. We can then exploit the implications of maximality in order to describe precisely such maximizers.

4.3 Propositions

In this section we present auxiliary results about floating point systems. We believe readers will find most of them to be trivial, and they are presented only to make our arguments more precise. In all propositions β is a base, μ is a positive integer, $\mathcal{F}$ is a floating point system associated to β and μ, $z \in \mathbb{R}$, $x \in \mathcal{F}$, and u, e_α, α and ν are the numbers related to this system in Definitions 2, 6, 7, 9, 11 and 12, Finally the function fl rounds to nearest in $\mathcal{F}$.

Proposition 1 (Order by the exponent) *Let d and e be integers and $v, w \in \mathbb{R}$, with $v < (\beta - 1)\beta^\mu$ and $w \geq 0$. If $d < e$ then $\beta^d (\beta^\mu + v) < \beta^e (\beta^\mu + w)$.* ▲

Proposition 2 (Normal form) *If $z \in \mathbb{R}$ is different from zero then there exist unique $e \in \mathbb{Z}$ and $w \in [0, (\beta - 1)\beta^\mu)$ such that $z = \text{sign}(z)\beta^e (\beta^\mu + w)$.* ▲

Proposition 3 (Integer form) *If $\mathcal{F}$ is unperfect and $x \in \mathcal{F}$ then there exists $e, r \in \mathbb{Z}$ with $e \geq e_\alpha$ such that $x = \beta^e r$ and*

- *$r = 0$ if and only if $x = 0$.*

- *$0 < |r| < \beta^\mu$ if and only if x is subnormal and $\mathcal{F}$ is an IEEE system.*

- *$\beta^\mu \leq |r| < \beta^{1+\mu}$ if and only if $|x| \in \mathcal{E}_e$.*

▲

Proposition 4 (Symmetry) *$\mathcal{F}$ is symmetric, that is, $x \in \mathcal{F} \Leftrightarrow -x \in \mathcal{F} \Leftrightarrow |x| \in \mathcal{F}$.* ▲

Proposition 5 (The minimality of nu) *Let x be a floating point number. If $|x| \geq \nu$ and $x \neq 0$ then x is normal, that is, there exists an exponent e for $\mathcal{F}$ such that $|x| \in \mathcal{E}_e$. If $0 < |x| < \nu$ then $\mathcal{F}$ is an IEEE system $\mathcal{I}_{e_\alpha}$ and x is subnormal, that is, $|x| \in \mathcal{S}_{e_\alpha}$. Conversely, if $e, r \in \mathbb{Z}$ and $z = \beta^e r$ with $|r| \leq \beta^{1+\mu}$ and $|z| \geq \nu$ then $z \in \mathcal{F}$.* ▲

Proposition 6 (Subnormal sum) *Let $\mathcal{I}$ be an IEEE system. If $x, y \in \mathcal{I}$ are subnormal then $x + y \in \mathcal{I}$.* ▲

Proposition 7 (Critical sum) *If e is an exponent for $\mathcal{F}$ and $x \in \mathcal{F}$ and $z \in \mathbb{R}$ are such that $|x + z| = \beta^e (\beta^\mu + r + 1/2)$ with $r \in [0, (\beta - 1)\beta^\mu) \cap \mathbb{Z}$ then $|z| \geq \beta^e / 2$.* ▲

Proposition 8 (Identity) *If $x \in \mathcal{F}$ then* $\mathrm{fl}(x) = x$. ▲

Proposition 9 (Monotonicity) *If $z \leq w$ then $\mathrm{fl}(z) \leq \mathrm{fl}(w)$, and if $x \in \mathcal{F}$ then*

- $x > \mathrm{fl}(z) \Rightarrow x > z$,

- $x < \mathrm{fl}(z) \Rightarrow x < z$,

- $|x| > |\mathrm{fl}(z)| \Rightarrow |x| > |z|$,

- $|x| < |\mathrm{fl}(z)| \Rightarrow |x| < |z|$.

▲

Proposition 10 (Symmetric rounding) $\mathrm{m}(z) := -\mathrm{fl}(-z)$ *rounds to nearest in $\mathcal{F}$.* ▲

Proposition 11 (Normal rounding) *Let e be an exponent for $\mathcal{F}$. If $|z| = \beta^e (\beta^\mu + w)$, with $w \in [0, (\beta - 1)\beta^\mu)$, then $\mathrm{fl}(z) \in \{a, b\}$ for*

$$a := \mathrm{sign}(z)\beta^e (\beta^\mu + \lfloor w \rfloor) \in \mathcal{F} \qquad \text{and} \qquad b := \mathrm{sign}(z)\beta^e (\beta^\mu + \lceil w \rceil) \in \mathcal{F},$$

and

$$|\mathrm{fl}(z) - z| = \min\{|z - a|, |z - b|\} \leq \frac{|b - a|}{2} \leq \beta^e/2. \tag{34}$$

If $z < m := (a + b)/2$ then $\mathrm{fl}(z) = \min\{a, b\}$, and if $z > m$ then $\mathrm{fl}(z) = \max\{a, b\}$.
    In particular, if $r \in \mathbb{Z}$ and $|r - w| < 1/2$ then $\mathrm{fl}(z) = \mathrm{sign}(z)\beta^e (\beta^\mu + r)$. ▲

Proposition 12 (Subnormal rounding) *Let $\mathcal{I} = \mathcal{I}_{e_\alpha}$ be an IEEE system. If $|z| \leq \nu$ then $\mathrm{fl}(z) \in \{a, b\}$ for*

$$a := \beta^{e_\alpha} \lfloor \beta^{-e_\alpha} z \rfloor \in \mathcal{I} \qquad \text{and} \qquad b := \beta^{e_\alpha} \lceil \beta^{-e_\alpha} z \rceil \in \mathcal{I}$$

and

$$|\mathrm{fl}(z) - z| = \min\{z - a, b - z\} \leq \frac{b - a}{2} \leq \alpha/2.$$

If $z < m := (a + b)/2$ then $\mathrm{fl}(z) = a$, and if $z > m$ then $\mathrm{fl}(z) = b$.
    If $r \in [-\beta^\mu, \beta^\mu) \cap \mathbb{Z}$ then $\mathrm{fl}(z) = \beta^{e_\alpha} r$ for $\beta^{e_\alpha} (r - 1/2) < z < \beta^{e_\alpha} (r + 1/2)$ and $\mathrm{fl}(\beta^{e_\alpha} (r + 1/2)) \in \{\beta^{e_\alpha} r, \beta^{e_\alpha} (r + 1)\}$. ▲

Proposition 13 (Rounding below alpha) *If $|z| < \alpha/2$ then $\mathrm{fl}(z) = 0$. If $|z| = \alpha/2$ then $\mathrm{fl}(z) \in \{0, \mathrm{sign}(z)\alpha\}$ and if $\alpha/2 < |z| \leq \alpha$ then $\mathrm{fl}(z) = \mathrm{sign}(z)\alpha$. In particular, if $|z| \leq \alpha$ then $|\mathrm{fl}(z) - z| \leq \alpha/2$.* ▲

Proposition 14 (Perfect adapter) *Let $\mathcal{P}_{\beta,\mu}$ be a perfect system and $\mathcal{F}_{e_\alpha,\beta,\mu}$ an unperfect one. If fl rounds to nearest in $\mathcal{F}$ then there exists $\tilde{\mathrm{fl}}$ which rounds to nearest in $\mathcal{P}$ and is such that $\tilde{\mathrm{fl}}(z) = \mathrm{fl}(z)$ for z with $|z| \geq \nu_{\mathcal{F}}$.* ▲

Proposition 15 (IEEE adapter) *If $\mathrm{Fl} = \{\mathrm{fl}_1, \ldots, \mathrm{fl}_n\}$ rounds to nearest in the IEEE system $\mathcal{I}_{e_\alpha,\beta,\mu}$ and $\mathcal{P}_{\beta,\mu}$ is a perfect system then there exists $\tilde{\mathrm{Fl}} = \{\tilde{\mathrm{fl}}_1, \ldots, \tilde{\mathrm{fl}}_n\}$ which rounds to nearest in $\mathcal{P}$ and is such that $\mathrm{Fl}(\sum_{k=0}^n x_k) = \tilde{\mathrm{Fl}}(\sum_{k=0}^n x_k)$ for all $\mathbf{x} \in \mathcal{I}^{n+1}$. In particular, $\tilde{\mathrm{Fl}}(\sum_{k=0}^n x_k) \in \mathcal{I}$.* ▲

Proposition 16 (MPFR adapter) *If* $\mathrm{Fl} = \{\mathrm{fl}_1,\dots,\mathrm{fl}_n\}$ *rounds to nearest in the MPFR system* $\mathcal{M}_{e_\alpha,\beta,\mu}$ *and* $\mathcal{P}_{\beta,\mu}$ *is a perfect system then there exists* $\tilde{\mathrm{Fl}} = \{\tilde{\mathrm{fl}}_1,\dots,\tilde{\mathrm{fl}}_n\}$ *which rounds to nearest in* $\mathcal{P}$ *and is such that* $\mathrm{Fl}(\sum_{k=0}^n x_k) = \tilde{\mathrm{Fl}}(\sum_{k=0}^n x_k)$ *for* $\mathbf{x} \in \mathcal{M}^{n+1}$ *with* $\mathrm{Fl}(\sum_{i=0}^{k-1} x_i) + x_k \geq 0$ *for* $k = 0,\dots,n$. ▲

Proposition 17 (Flatness) *Let* e *be an exponent for* $\mathcal{F}$ *and* z *with* $|z| = \beta^e(\beta^\mu + w)$ *for* $w \in [0,(\beta-1)\beta^\mu)$. *On the one hand, if* $w = \lfloor w \rfloor + 1/2$ *then*

$$|w - y| < \beta^e/2 \;\Rightarrow\; \mathrm{fl}(y) \in \{\mathrm{sign}(z)\beta^e(\beta^\mu + \lfloor w \rfloor), \mathrm{sign}(z)\beta^e(\beta^\mu + \lceil w \rceil)\}.$$

On the other hand, if $w - \lfloor w \rfloor \neq 1/2$ *then there exists* $\delta > 0$ *such that if* fl_1 *and* fl_2 *round to nearest in* $\mathcal{F}$ *and* $|y - z| < \delta$ *then* $\mathrm{fl}_1(y) = \mathrm{fl}_2(z)$. ▲

Proposition 18 (Scaled sums) *Suppose* $\mathrm{Fl} = \{\mathrm{fl}_1,\dots,\mathrm{fl}_n\}$ *rounds to nearest in a perfect system* $\mathcal{P}$ *and* S_k *is the sum in Definition* 18. *If* $\mathbf{z} \in \mathbb{R}^n$, $\sigma \in \{-1,1\}$ *and* $m \in \mathbb{Z}$ *then there exist* $\tilde{\mathrm{Fl}} = \{\tilde{\mathrm{fl}}_1,\dots,\tilde{\mathrm{fl}}_n\}$ *which round to nearest in* $\mathcal{P}$ *such that* $S_k(\sigma\beta^m\mathbf{z}, \mathrm{Fl}) = \sigma\beta^m S_k(\mathbf{z}, \tilde{\mathrm{Fl}})$ *for* $k = 1,\dots,n$. ▲

Proposition 19 (Whole is tight) *The set of all functions which round to nearest in* $\mathcal{F}$ *is tight.* ▲

Proposition 20 (Sums are tight) *Let* $\mathcal{R}$ *be a tight set of functions which round to nearest in* $\mathcal{F}$ *and* S_k *the sum in Definition* 18. *The function* $T_n : \mathbb{R}^n \times \mathcal{R}^n \to \mathbb{R}^{n+1}$ *given by*

$$T_n(\mathbf{z}, \mathrm{Fl}) := (S_0(\mathbf{z}, \mathrm{Fl}), S_1(\mathbf{z}, \mathrm{Fl}), S_2(\mathbf{z}, \mathrm{Fl}), \dots, S_n(\mathbf{z}, \mathrm{Fl}))$$

is tight. ▲

4.4 Lemmas

This section presents the proofs of the Lemmas other than 4 and 6, which are proved in the extended version of the article.

Proof of Lemma 1 If $z = 0$ then $\mathrm{fl}(z) = z = 0$ by Prop. 8 and Equation (11) holds. If $z \neq 0$ then, by Prop. 2, $z = \mathrm{sign}(z)\beta^e(\beta^\mu + w)$, with $e \in \mathbb{Z}$ and $r \in \mathbb{R}$ with $w \in [0,(\beta-1)\beta^\mu)$. When $\mathcal{F}$ is unperfect $\nu = \beta^{e_\alpha+\mu}$ and, by Prop. 1, $e \geq e_\alpha$ because $|z| \geq \nu$. Therefore, e is an exponent for $\mathcal{F}$ and Lemma 1 follows from Lemma 9. Finally, Lemma 1 applies to all z when $\mathcal{F}$ is perfect because $\nu = 0$ in this case. □

Proof of Lemma 2 If $\mathcal{F}$ is perfect then $\alpha = \nu = 0$ and Lemma 2 holds because $0 \in \mathcal{F}$. It is clear that the Lemma also holds when $x = 0$ or $y = 0$, and from now on we suppose that $x,y \neq 0$ and $\mathcal{F}$ is unperfect. In this case $\nu = \beta^{e_\alpha+\mu}$, and we can assume that $|y| \geq |x|$ because $x + y = y + x$. Moreover, $x + y \in \mathcal{F} \Leftrightarrow -(x+y) \in \mathcal{F}$ by Prop. 4 and we can also assume that

$$\alpha \leq x + y \leq \beta\nu = \beta^{1+e_\alpha+\mu} \qquad \text{and} \qquad y > 0. \tag{35}$$

If y is subnormal then $0 < |x| \leq y < \nu$, x is also subnormal by Prop. 5 and Lemma 2 follows from Prop. 6. Therefore, we can assume that y is normal, that is,

$$y = \beta^{e_\alpha+e}(\beta^\mu + r_y) \qquad \text{with} \qquad e \geq 0 \qquad \text{and} \qquad r_y \in [0,(\beta-1)\beta^\mu) \cap \mathbb{Z}. \tag{36}$$

On the one hand, if $x > 0$ then Equation (35) leads to $y < \beta^{1+e\alpha+\mu}$ and Equation (36) yields $e = 0$. Prop. 3 and the assumption $0 < |x| \leq y$ lead to

$$x = \beta^{e\alpha} r_x \qquad \text{with} \qquad r_x \in \mathbb{Z} \quad \text{and} \quad 1 \leq r_x \leq \beta^{\mu} + r_y,$$

and Equations (35) and (36) imply that

$$x + y = \beta^{e\alpha} (\beta^{\mu} + r_x + r_y) \leq \beta^{1+e\alpha+\mu} \; \Rightarrow x + y \geq \nu \; \text{ and } \; r_x + r_y \leq (\beta - 1)\beta^{\mu},$$

and Prop. 5 with $r = \beta^{\mu} + r_x + r_y$ shows that $x + y \in \mathcal{F}$.

On the other hand, if $x < 0$ then Prop. 3 and the assumption $0 < |x| \leq y$ lead to

$$x = -\beta^{e\alpha+d} r_x \quad \text{with} \quad 0 \leq d \leq e, \quad r_x \in \mathbb{Z} \quad \text{and} \quad 1 \leq r_x < \beta^{1+\mu}.$$

It follows that $x + y = \beta^{e\alpha+d} \left(\beta^{e-d} (\beta^{\mu} + r_y) - r_x \right) = \beta^{e\alpha} r$ for

$$r := \beta^{d} \left(\beta^{e-d} (\beta^{\mu} + r_y) - r_x \right) \in \mathbb{Z}.$$

Since $x < 0$, using (35) and the identity $\nu = \beta^{e\alpha+\mu}$ we deduce that

$$0 < r = \beta^{-e\alpha} (x + y) < \beta^{-e\alpha} \beta \nu = \beta^{1+\mu}.$$

When $\mathcal{F}$ is a MPFR system we have that $\alpha = \nu$, Equation (35) implies that $x + y \geq \nu$ and the equation above and Prop. 5 show that $x + y \in \mathcal{F}$. Finally, when $\mathcal{F}$ is an IEEE system we either have (i) $r \geq \beta^{\mu}$, in which case $x + y \geq \beta^{e\alpha+\mu} = \nu$ and $x + y \in \mathcal{F}$ by Prop. 5, or (ii) $r < \beta^{\mu}$, and $x + y \in \mathcal{S}_{e\alpha}$ is a subnormal number, which belongs to $\mathcal{F}$. Therefore, $x + y \in \mathcal{F}$ in all cases and we are done. $\qquad\square$

Proof of Lemma 3 By Prop. 3, $x = \beta^{d} r$ and $y = \beta^{e} s$ for $d, e, r, s \in \mathbb{Z}$ such that $d, e \geq e_\alpha$. It follows that $z = \beta^{e\alpha} t$ for $t := \beta^{d-e\alpha} r + \beta^{e-e\alpha} s \in \mathbb{Z}$. We have that $|t| \geq 1$ because $t \in \mathbb{Z} \setminus \{0\}$ and $|z| = \beta^{e\alpha} |t| \geq \beta^{e\alpha} = \alpha$, and $z \in \mathcal{F}$ by Lemma 2. $\qquad\square$

Proof of Lemma 5 This proof illustrates the use of optimization to bound rounding errors. We define $z_1 := y_0 + y_1$, $z_k := y_k$ for $k > 1$ and use the sums S_k in Definition 18, the set $\mathcal{R}$ of all $n-$tuples which round to nearest and the function

$$\eta(\mathbf{z}, \text{Fl}) := \sum_{k=1}^{n} |S_k(\mathbf{z}, \text{Fl}) - (S_{k-1}(\mathbf{z}, \text{Fl}) + z_k)|, \tag{37}$$

from $\mathbb{R}^n \times \mathcal{R}$ to $\mathbb{R}$. We show that Example 1 is the worst case for the ratio

$$q_n(\mathbf{z}, \text{Fl}) := \frac{\eta(\mathbf{z}, \text{Fl})}{\sum_{k=1}^{n} |z_k|}. \tag{38}$$

This ratio is related to Equation (17) because

$$\left| \text{Fl}\left(\sum_{k=0}^{n} y_k \right) - \sum_{k=0}^{n} y_k \right| = \left| \sum_{k=1}^{n} (S_k(\mathbf{z}, \text{Fl}) - S_{k-1}(\mathbf{z}, \text{Fl}) - z_k) \right| \leq \eta(\mathbf{z}, \text{Fl})$$

and

$$\left| \text{Fl}\left(\sum_{k=0}^{n} y_k \right) - \sum_{k=0}^{n} y_k \right| \leq q_n(\mathbf{z}, \text{Fl}) \sum_{k=1}^{n} |z_k| \leq q_n(\mathbf{z}, \text{Fl}) \sum_{k=0}^{n} |y_k|.$$

18

Therefore, to prove Lemma 5 it suffices to show that

$$\sup_{(\mathbf{z},\mathrm{Fl})\in(\mathbb{R}^n\setminus\{0\})\times\mathcal{R}} q_n(\mathbf{z},\mathrm{Fl}) = \theta_n u \qquad \text{for} \qquad \theta_n := \frac{n}{1+nu}. \qquad (39)$$

The ratio q_n can be written as $q_n(\mathbf{z},\mathrm{Fl}) = f(\mathbf{z},h(\mathbf{z},\mathrm{Fl}))$ for

$$h(\mathbf{z},\mathrm{Fl}) := (S_0(\mathbf{z},\mathrm{Fl}),S_1(\mathbf{z},\mathrm{Fl}),\ldots,S_n(\mathbf{z},\mathrm{Fl})) \in \mathbb{R}^{n+1}$$

and

$$f(\mathbf{z},\mathbf{x}) := \frac{\sum_{k=1}^n |x_k - (x_{k-1}+z_k)|}{\sum_{k=1}^n |z_k|}.$$

The function f is continuous for $\mathbf{z} \neq 0$ and satisfies $f(\lambda\mathbf{z},\lambda\mathbf{x}) = f(\mathbf{z},\mathbf{x})$ for $\lambda \neq 0$, and Prop. 18, 19 and 20 show that h satisfies the requirements of Lemma 10. We can then apply this Lemma to prove that either (i) $\sup q_n \leq \theta_n u$ or (ii) q_n has a maximizer $(\mathbf{z}^*,\mathrm{Fl}^*)$. In case (ii) we use the properties of this maximizer to prove that it is no worse than what is described in Example 1. For instance, this example tells us that $q_n(\mathbf{z}^*,\mathrm{Fl}^*) \geq \theta_n u$ and if q has a partial derivative with respect to z_k at $\mathbf{z}^*$ then this derivative is zero.

We prove Equation (39) by induction. For $n = 1$, this equation follows from Lemma 1. Let us then assume that $n > 1$ and Equation (39) is valid for $\mathbf{z} \in \mathbb{R}^m$ and rounding tuples $\mathrm{Fl} = \{\mathrm{fl}_1,\ldots,\mathrm{fl}_m\}$ when $m < n$ and show that it also holds for n. To apply Lemma 10, let us define the numbers

$$a := \frac{1+(1+2u)\theta_{n-1}-(1+u)\theta_n}{1+u} = \frac{(n-1)(3+2nu)u}{(1+nu)(1+(n-1)u)(1+u)} > 0 \qquad (40)$$

(recall that $n \geq 2$) and

$$b := \theta_n - \theta_{n-1} = \frac{1}{(1+nu)(1+(n-1)u)} > 0, \qquad (41)$$

and split $\mathbb{R}^n \setminus \{0\}$ as the union of the set

$$\mathcal{B} := \left\{ \mathbf{z} \in \mathbb{R}^n \text{ with } b\sum_{k=2}^n |z_n| > a|z_1| \right\} \qquad (42)$$

and the cone $\mathcal{A} := \{\lambda\mathbf{z}, \text{ with } \mathbf{z} \in \mathcal{K}, \lambda \in \mathbb{R}\setminus\{0\}\}$, for

$$\mathcal{K} := \left\{ \mathbf{z} \in \mathbb{R}^n \text{ with } 2/3 \leq z_1 \leq 2\beta/3 \text{ and } b\sum_{k=2}^n |z_k| \leq az_1 \right\}. \qquad (43)$$

We claim that $q_n(\mathbf{z},\mathrm{Fl}) \leq \theta_n u$ for $\mathbf{z} \in \mathcal{B}$ and $\mathrm{Fl} \in \mathcal{R}$. In fact, writing $\hat{s}_k := S_k(\mathbf{z},\mathrm{Fl})$ and $s_k := \sum_{i=1}^k z_i$ for $k = 0,\ldots,n$ and using Equation (39) with $\tilde{\mathrm{Fl}} := \{\mathrm{fl}_2,\ldots,\mathrm{fl}_n\}$ and $\tilde{\mathbf{z}} := (\hat{s}_1+z_2,z_3,\ldots,z_n)$ we obtain by induction that

$$\sum_{k=2}^n |\hat{s}_k - \hat{s}_{k-1} - z_k| \leq \theta_{n-1}u\left(|\hat{s}_1+z_2| + \sum_{k=3}^n |z_k|\right). \qquad (44)$$

Keeping in mind that $z_1 = s_1$, we have that

$$|\hat{s}_1 - s_1| + \sum_{k=2}^n |\hat{s}_k - \hat{s}_{k-1} - z_k| \leq (|\hat{s}_1 - s_1| + \theta_{n-1}u|\hat{s}_1|) + \theta_{n-1}u\sum_{k=2}^n |z_k|,$$

19

and Lemma 1, the definitions (37), (40) and (41) of η, a and b and $\hat{s}_0 = 0$ yield

$$\eta(\mathbf{z},\mathrm{Fl}) = \sum_{k=1}^{n} |\hat{s}_k - \hat{s}_{k-1} - z_k| \leq (1 + \theta_{n-1}(1 + 2u)) \frac{u\,|s_1|}{1+u} + \theta_{n-1}u \sum_{k=2}^{n} |z_k|$$

$$= u \left((1 + (1+2u)\theta_{n-1} - (1+u)\theta_n) \frac{|z_1|}{1+u} - (\theta_n - \theta_{n-1}) \sum_{z=2}^{n} |z_k| \right) + \theta_n u \sum_{k=1}^{n} |z_k|$$

$$= u \left(a\,|z_1| - b \sum_{k=2}^{n} |z_k| \right) + \theta_n u \sum_{k=1}^{n} |z_k|.$$

The definitions (38) and (42) of q and $\mathcal{B}$ and this equation imply that $q_n(\mathbf{z},\mathrm{Fl}) \leq \theta_n u$, and, indeed, $q_n(\mathbf{z},\mathrm{Fl}) \leq \theta_n u$ for $\mathbf{z} \in \mathcal{B}$ and $\mathrm{Fl} \in \mathcal{R}$. As a result, Lemma 10 shows that either (i) the supremum of q_n is at most $\theta_n u$ or (ii) there exists $\mathbf{z}^* \in \mathcal{K}$ and $\mathrm{Fl}^* \in \mathcal{R}$ with

$$q_n(\mathbf{z}^*,\mathrm{Fl}^*) = \sup_{(\mathbf{z},\mathrm{Fl}) \in (\mathbb{R}^n \setminus \{\mathbf{0}\}) \times \mathcal{R}} q_n(\mathbf{z},\mathrm{Fl}^*).$$

In case (i) we are done and we now analyze case (ii). Let us define $\hat{s}_k^* := S_k(\mathbf{z}^*,\mathrm{Fl}^*)$, and $s_k^* := \sum_{i=1}^{k} z_k^*$, for $k = 0, \dots, n$. Since $\mathbf{z}^* \in \mathcal{K}$, the definitions of a and b lead to

$$\sum_{k=2}^{n} |z_k^*| \leq \frac{(n-1)(3+2nu)}{1+u} u z_1^*.$$

Using Lemma 1, the hypothesis $20nu \leq 1$ and induction we deduce that

$$|\hat{s}_k^* - z_1^*| \leq \left| \hat{s}_k^* - \left((\hat{s}_1^* + z_2^*) + \sum_{i=3}^{n} z_i^* \right) \right| + |\hat{s}_1^* - s_1^*| + \sum_{i=2}^{n} |z_i^*|$$

$$\leq \frac{(n-1)u}{1+(n-1)u} \left(|\hat{s}_1^* + z_2^*| + \sum_{i=3}^{n} |z_i^*| \right) + \frac{u}{1+u} z_1^* + \frac{(n-1)(3+2nu)}{1+u} u z_1^*$$

$$\leq \left(\frac{n-1}{1+(n-1)u}(1 + 2u + (n-1)(3+2nu)u) + 1 + (n-1)(3+2nu) \right) \frac{u}{1+u} z_1^*$$

and, since $s_1^* = z_1^*$ and $20nu \leq 1$,

$$|\hat{s}_k^* - z_1^*| \leq \kappa n u z_1^* \leq \kappa z_1^*/20, \tag{45}$$

for

$$\kappa := \left(\frac{1}{1+nu}(1 + (3+2nu)nu) + 3 + 2nu \right) \frac{1}{1+u}$$

$$\leq \frac{1}{1+\frac{1}{20}} \left(1 + \frac{1}{20}\left(3 + \frac{1}{10}\right) \right) + 3 + \frac{1}{10} = \frac{21}{5}. \tag{46}$$

Since $2/3 \leq z_1^* \leq 2\beta/3$, Equations (45) and (46) lead to

$$\frac{1}{\beta} \leq \frac{1}{2} < \frac{158}{300} \leq \frac{79}{100} z_1^* \leq \hat{s}_k^* \leq \frac{121}{100} z_1^* \leq \frac{121}{150}\beta < \beta$$

for $1 < k \leq n$, and since $\hat{s}_1^* = \mathrm{fl}_1(z_1^*)$ and $2/3 \leq z_1^* \leq 2/3\beta$ this equation also holds for $k = 1$. Monotonicity (Prop. 9) and the fact that $\hat{s}_k^* = \mathrm{fl}_k(\hat{s}_{k-1} + z_k)$ lead to

$$1/\beta < \hat{s}_{k-1} + s_k^* < \beta \quad \text{for } 1 \leq k \leq n. \tag{47}$$

20

We now explore the implications of $(\mathbf{z}^*,\mathrm{Fl}^*)$ being a maximizer of q_n. Example 1 shows that $q_n(\mathbf{z}^*,\mathrm{Fl}^*) \geq \theta_n u$ and this implies that $z_k \neq 0$ for all k, because if $z_k = 0$ for some k then we would have $q_n(\mathbf{z}^*,\mathrm{Fl}^*) = q_{n-1}(\tilde{\mathbf{z}},\tilde{\mathrm{Fl}})$ for $\tilde{\mathbf{z}} \in \mathbb{R}^{n-1}$ and $\tilde{\mathrm{Fl}}$ obtained by removing the kth coordinate of $\mathbf{z}^*$ and fl_k from Fl^*, and $q_{n-1}(\tilde{\mathrm{Fl}},\tilde{\mathbf{z}}) \leq \theta_{n-1}u < \theta_n u$, contradicting the maximality of $(\mathbf{z}^*,\mathrm{Fl}^*)$. Therefore, $z_k^* \neq 0$ for $k = 1,\ldots,n$, and the denominator of q_n has non zero partial derivatives at $\mathbf{z}^*$. Equation (47) shows that $\hat{s}_{k-1}^* + z_k^* \neq 0$, and Prop. 17 implies that the numerator of q_n will have a zero partial derivative with respect to z_k if $\hat{s}_{k-1}^* + z_k^*$ is not of the form

$$\hat{s}_{k-1}^* + z_k^* = \beta^{e_k}\left(\beta^\mu + r_k + 1/2\right) \qquad \text{with} \qquad e_k \in \mathbb{Z} \text{ and } r_k \in [0,(\beta-1)\beta^\mu), \quad (48)$$

and this would imply that the derivative of q_n is well defined and different from zero. By the maximality of $(\mathbf{z}^*,\mathrm{Fl}^*)$, we conclude that Equation (48) is valid. Combining this equation with Equation (47) we conclude that we can write $\{1,2,\ldots,n\} = \mathcal{L} \cup \mathcal{H}$ (for low and high) so that the exponents in e_k Equation (48) are $e_k = -\mu - 1$ for $k \in \mathcal{L}$ and $e_k = -\mu$ for $k \in \mathcal{H}$. Since $\beta^{-\mu}/2 = u$, this leads to

$$
\begin{aligned}
k \in \mathcal{L} &\quad\Rightarrow\quad \frac{1+u}{\beta} \leq \hat{s}_{k-1}^* + z_k^* \leq \frac{\beta - u}{\beta}, \\
k \in \mathcal{U} &\quad\Rightarrow\quad 1 + u \leq \hat{s}_{k-1}^* + z_k^* \leq \beta - u,
\end{aligned}
$$

As a result, Prop. 7 implies that

$$k \in \mathcal{L} \Rightarrow |z_k^*| \geq u/\beta \qquad \text{and} \qquad k \in \mathcal{H} \Rightarrow |z_k^*| \geq u, \quad (49)$$

and Prop. 11 yields

$$k \in \mathcal{L} \Rightarrow \left|\hat{s}_k^* - \left(\hat{s}_{k-1}^* + z_k^*\right)\right| = u/\beta \quad \text{and} \quad k \in \mathcal{H} \Rightarrow \left|\hat{s}_k^* - \left(\hat{s}_{k-1}^* + z_k^*\right)\right| = u. \quad (50)$$

We now show that if $1 \in \mathcal{L}$ then we obtain a contradiction to the maximality of $(\mathbf{z}^*,\mathrm{Fl}^*)$. Indeed, let $m \in [1,n]$ be the last index such that $k \in \mathcal{L}$ for $1 \leq k \leq m$. If $m = n$ then $k \in \mathcal{L}$ for all $k \in [1,n]$ and the inequality $z_1^* \geq 2/3$ and Equations (49) and (50) and the fact that $2\beta/3 - u > 1$ imply that

$$q_n(\mathbf{z}^*,\mathrm{Fl}^*)/(\theta_n u) = \frac{\frac{nu/\beta}{2/3+(n-1)u/\beta}}{\frac{nu}{1+nu}} = \frac{1+nu}{\left(\frac{2\beta}{3} - u\right) + nu} < 1,$$

and this contradicts the maximality of $(\mathbf{z}^*,\mathrm{Fl}^*)$. For $m < n$ we have

$$\sum_{k=1}^m |z_k^*| \geq \sum_{k=1}^m z_k^* = \left(\hat{s}_m^* + z_{m+1}^*\right) - \left(\sum_{k=1}^m \left(\hat{s}_k^* - \left(\hat{s}_{k-1}^* + z_k^*\right)\right)\right) - z_{m+1}^*$$

$$\geq (1 + u) - (mu/\beta) - \left|z_{m+1}^*\right|.$$

Let ℓ be the size of $\mathcal{L}$ and h the size of $\mathcal{H}$. Equations (49) and (50), the identity $n = \ell + h$ and the hypothesis $20nu \leq 1$ lead to

$$q_n(\mathbf{z}^*,\mathrm{Fl}^*) - \theta_n u \leq \frac{\ell u/\beta + hu}{1 + u - mu/\beta - \left|z_{m+1}^*\right| + (\ell - m)u/\beta + \left|z_{m+1}^*\right| + (h-1)u} - \frac{nu}{1+nu}$$

$$= -u\frac{\xi}{(1+nu)(\beta - 2mu + \ell u + \beta hu)},$$

21

for

$$\xi := (\beta - 1)\ell - 2hmu - 2\ell mu = \ell\left((\beta - 1) - \left(\frac{m}{\ell}\right)(2hu) - 2mu\right) \geq 0.8\ell > 0,$$

and, again, $q_n(\mathbf{z}^*, \mathrm{Fl}^*) < \theta_n u$. Therefore, by the maximality of $(\mathrm{Fl}^*, \mathbf{z}^*)$ we must have $z_1^* \geq 1$, and Equation (48) shows that $z_1^* \geq 1 + u$ and Equations (49) and (50) lead to

$$q_n(\mathbf{z}^*, \mathrm{Fl}^*) \leq \frac{\ell u/\beta + hu}{1 + u + \ell u/\beta + (h-1)u} = \frac{\ell u/\beta + hu}{1 + \ell u/\beta + hu}. \tag{51}$$

Since $n = \ell + h$, $\theta_n = (\ell + h)/(1 + (\ell + h)u)$ and

$$\frac{\ell + h}{1 + (\ell + h)u} - \frac{\ell/\beta + h}{1 + \ell u/\beta + hu} = \frac{(\beta - 1)\ell}{(1 + (\ell + h)u)(\beta + \ell u + \beta hu)} \geq 0,$$

Equation (51) implies that $q_n(\mathbf{z}^*, \mathrm{Fl}^*) \leq \theta_n u$ and we are done. $\qquad\square$

Proof of Lemma 7 Let us define $z_1 := y_0 + y_1$ and $z_k := y_k$ for $k > 1$. Using Lemma 1 and induction in n we can show that

$$S_n(\mathbf{z}, \mathrm{Fl}) \geq \sum_{k=1}^{n}(1 + u)^{-(n-k+1)} z_k = \frac{1}{1 + u}\sum_{k=1}^{n}(1 + u)^{-(n-k)} z_k.$$

The convexity of the functions $(1 + u)^{-(n-k)}$, which have value 1 and derivative $-(n - k)$ at $u = 0$, lead to

$$S_n(\mathbf{z}, \mathrm{Fl}) \geq \frac{1}{1 + u}\left(\sum_{k=1}^{n} z_k - u\sum_{k=1}^{n}(n - k)z_k\right)$$

$$= \frac{1}{1 + u}\left((1 + u)\sum_{k=1}^{n} z_k - u\sum_{k=1}^{n}(n - k + 1)z_k\right) = \sum_{k=1}^{n} z_k - \frac{u}{1 + u}\sum_{k=1}^{n}(n - k + 1)z_k,$$

and the lower bound in Equation (22) follows from the identities

$$\sum_{i=0}^{k} y_i = \sum_{i=1}^{k} z_i, \qquad \mathrm{Fl}\left(\sum_{k=0}^{n} y_k\right) = S_n(\mathbf{z}, \mathrm{Fl}) \qquad \text{and} \qquad \sum_{k=1}^{n}\sum_{i=0}^{k} y_i = \sum_{k=1}^{n}(n - k + 1)z_k.$$

In order to prove the second inequality in Equation (22), we proceed as in the proof of Lemma 5 (We ask the reader to look at the first two paragraphs of that proof.) This time we consider only the rounding tuple $\mathrm{Fl} := \{\mathrm{fl}, \ldots, \mathrm{fl}\}$ where fl rounds to nearest and breaks all ties upward, because our function

$$q_n(\mathbf{z}) := \frac{\eta(\mathbf{z})}{\sum_{k=1}^{n}(n - k + 1)z_k} \tag{52}$$

for

$$\eta(\mathbf{z}) := S_n(\mathbf{z}, \mathrm{Fl}) - \sum_{k=1}^{n} z_k = \sum_{k=1}^{n}(S_k(\mathbf{z}, \mathrm{Fl}) - S_{k-1}(\mathbf{z}, \mathrm{Fl}) - z_k) \tag{53}$$

is clearly maximized by the rounding tuple Fl for which all ties are broken upward.

We prove by induction that

$$q_n(\mathbf{z}) \leq \tau_n u := \frac{u}{1 + u\left(\frac{\beta - 2}{\beta - 1} + \frac{n}{\beta^n - 1}\right)}. \tag{54}$$

For $n = 1$ Equation (54) follows from Lemma 1. Let us then assume it holds for $n - 1$ and prove it for n using Lemma 10 to show that either Equation (54) holds or there exists a maximizer for q_n, which we then analyze. With this purpose, define

$$a := \frac{1 + (n-1)(1+2u)\tau_{n-1} - n(1+u)\tau_n}{1+u} \quad \text{and} \quad b := \tau_n - \tau_{n-1}. \quad (55)$$

In order to prove that a and b are positive, note that

$$\tau_n = \frac{1}{1+u\phi_n} \quad \text{and} \quad \tau_{n-1} := \frac{1}{1+u(\phi_n+\delta_n)} \quad \text{for} \quad \phi_n := \frac{\beta-2}{\beta-1} + \frac{n}{\beta^n-1}$$

and

$$\delta_n := \frac{n-1}{\beta^{n-1}-1} - \frac{n}{\beta^n-1} = \frac{n(\beta-1)-(\beta-\beta^{1-n})}{\beta^n(1-\beta^{1-n})(1-\beta^{-n})} > 0.$$

For $\beta, n \geq 2$ we have that $\delta_n > 0$, and the positivity of δ_n implies that

$$b = \tau_n - \tau_{n-1} = u\delta_n \tau_{n-1} \tau_n > 0.$$

For $n = 2$ the software Mathematica shows that

$$a = u\frac{\beta - 1 + u(\beta - 2)}{(1+u)^2(\beta+1+\beta u)} > 0.$$

Mathematica also shows that when $\beta = 2$

$$a = u\frac{(2^n(n-2)+2)(2^n-n-1)}{(1+u)(2^n-1+nu)(2^n-2+2(n-1)u)},$$

which is positive for $n \geq 3$. For $\beta = 3$ we have

$$a = \frac{u(u+2)(3^n-2n-1)((2n-3)3^n+3)}{(1+u)(2(3^n-1)+u(3^n+2n-1))(2\times 3^n-6+u(3^n+6n-9))},$$

which is also positive for $n \geq 3$. Finally, for $\beta \geq 4$ and $n \geq 3$

$$n\delta_n \leq n\frac{(n-1)(\beta-1)}{(1-\beta^{1-n})(1-\beta^{-n})}\beta^{-n} \leq \frac{3\times 2\times 4^{-3}}{(1-4^{-2})(1-4^{-3})} = \frac{32}{315} < 0.2,$$

and the software Mathematica also shows that

$$a = \frac{(n-3)+(1-(n-1+nu)\delta_n)+\phi_n(1+(\delta_n+\phi_n+n-2)u)}{(1+u)(1+u\phi_n)(1+u(\phi_n+\delta_n))}$$

and this number is positive for $n \geq 3$ because $(n-1+nu)\delta_n \leq n\delta_n \leq 0.2$. Therefore, a and b are positive and the set

$$\mathcal{K} := \left\{ \mathbf{z} \in \mathbb{R}^n \setminus \{0\} \text{ with } 2/3 \leq z_1 \leq 2\beta/3, \ z_k \geq 0 \text{ and } b\sum_{k=2}^n (n-k+1)z_k \leq az_1 \right\}$$

is compact. We now split $\{\mathbf{x} \in \mathbb{R}^n \setminus \{0\} \text{ with } x_k \geq 0\}$ as the union of the set

$$\mathcal{B} := \left\{ \mathbf{z} \in \mathbb{R}^n \text{ with } z_k \geq 0 \text{ and } b\sum_{k=2}^n (n-k+1)z_k > az_1 \right\} \quad (56)$$

23

and the cone
$$\mathcal{A} := \{\lambda \mathbf{x} \text{ with } \mathbf{x} \in \mathcal{K} \text{ and } \lambda \in \mathbb{R}, \ \lambda > 0\}$$

and show that $q_n(\mathbf{z}) \leq \tau_n u$ for $\mathbf{z} \in \mathcal{B}$. In fact, for such $\mathbf{z}$, let us write $\hat{s}_k := S_k(\mathbf{z}, \mathrm{Fl})$ for $k = 0, \ldots, n$. Using induction, Lemma 1, and the definitions of a, b, and keeping in mind that $s_1 = z_1$, we deduce that

$$\sum_{k=1}^{n} \left(\hat{s}_k - (\hat{s}_{k-1} + z_k) \right) = (\hat{s}_1 - s_1) + \left((\hat{s}_2 - (\hat{s}_1 + z_2)) + \sum_{k=3}^{n} (\hat{s}_k - (s_{k-1} - z_k)) \right)$$

$$\leq \frac{u}{1+u} z_1 + \tau_{n-1} u \left((n-1)(\hat{s}_1 + z_2) + \sum_{k=3}^{n} (n-k+1) z_k \right)$$

$$= \frac{u}{1+u} z_1 + \tau_{n-1} u \left((n-1)\hat{s}_1 + \sum_{k=2}^{n} (n-k+1) z_k \right)$$

$$\leq (1 + (1+2u)(n-1)\tau_{n-1}) \frac{u z_1}{1+u} + \tau_{n-1} u \sum_{k=2}^{n} (n-k+1) z_k$$

$$= (1 + (1+2u)(n-1)\tau_{n-1} - n(1+u)\tau_n) \frac{u z_1}{1+u}$$

$$- (\tau_n - \tau_{n-1}) u \sum_{k=2}^{n} (n-k+1) z_k + \tau_n u \left(n z_1 + \sum_{k=2}^{n} (n-k+1) z_k \right),$$

and it follows that

$$\eta(\mathbf{z}) \leq \left(a z_1 - b \sum_{k=2}^{n} (n-k+1) z_k \right) u + \tau_n u \sum_{k=1}^{n} (n-k+1) z_k.$$

By the definition of $\mathcal{B}$ the term in parenthesis above is negative and this equation shows that $q_n(\mathbf{z}) \leq \tau_n u$ for $\mathbf{z} \in \mathcal{B}$. According to Lemma 10 we have that either (i) Equation (54) holds or (ii) q_n has a maximizer $\mathbf{z}^* \in \mathcal{K}$. In case (i) we are done and we now suppose that there exists such $\mathbf{z}^*$. Define $\hat{s}_k^* := S_k(\mathbf{z}^*, \mathrm{Fl})$ for $k = 0, \ldots, n$. The same argument used in the proof of Lemma 5 to deduce that $z_k^* \neq 0$ and Equation (48) shows that $z_k^* \neq 0$ for $k = 1, \ldots, n$, and

$$\hat{s}_{k-1}^* + z_k^* = \beta^{d_k}(\beta^\mu + r_k + 1/2) \quad \text{with} \quad d_k \in \mathbb{Z} \quad \text{and} \quad r_k \in [0, (\beta-1)\beta^\mu) \cap \mathbb{Z}. \quad (57)$$

Since fl break ties upward, we have that

$$\hat{s}_k^* = \beta^{d_k}(\beta^\mu + r_k + 1), \quad (58)$$

If the r_k in Equation (57) were all zero then, since $\hat{s}_0^* = 0$ and

$$\frac{1}{\beta} < 2/3 \leq z_1^* \leq \frac{2\beta}{3} < \beta,$$

Equation (57) would yield $z_1^* = \beta^{-\mu}(\beta^\mu + 1/2) = 1 + u$ and, for $k > 1$, Equations (57) and (58) would lead to $\hat{z}_k^* = \beta^{d_k+\mu}(1+u) - \beta^{d_{k-1}+\mu}(1+2u)$ and the z_k^* would correspond to the x_k in Example 4 with $e_k = d_k + \mu$ (take $x_0 = 0$ and $x_1 = z_1^*$). Therefore, by the last line in the statement of Example 4, in order to complete this proof it suffices to show that $r_k = 0$ for all k, and this is what we do next.

24

We start with $k < n$ and after that we handle the case $k = n$. Let us define $r_0 := 0$, assume that $r_i = 0$ for $i < k < n$ and show that $r_k = 0$. Take $\delta_k := \min\{1, r_k\}$ and $\mathbf{z}' \in \mathbb{R}^n$ given by $z_i' := z_i^*$ for $i < k$ or $i > k+1$ and

$$z_k' := z_k^* - \beta^{d_k}\delta_k \qquad \text{and} \qquad z_{k+1}' := z_{k+1}^* + \beta^{d_k}\delta_k.$$

We now prove that $\delta_k = 0$ by showing that $\mathbf{z}' = \mathbf{z}^*$. If $\delta_k = 0$ then $\mathbf{z}' = \mathbf{z}^*$ and $\mathbf{z}$ is in the domain of q_n. If $\delta = 1$ then $z_{k+1}' > 0$ and showing that $z_k' \geq 0$ suffices to prove that $\mathbf{z}'$ is in the domain of q_n. Indeed, Equations (57) and (58) and $r_{k-1} = 0$ lead to

$$z_k' := \beta^{d_k}(\beta^\mu + r_k + 1/2) - \beta^{d_{k-1}}(\beta^\mu + 1) - \beta^{d_k}\delta_k$$

$$= \beta^{d_k}(\beta^\mu + (r_k - \delta_k) + 1/2) - \beta^{d_{k-1}}(\beta^\mu + 1).$$

Equations (57), (58) and $z_k^* \geq 0$ imply that

$$\hat{s}_k^* = \mathrm{fl}(\hat{s}_{k-1}^* + z_k) \geq \hat{s}_{k-1}^* + \beta^{d_k}/2 > \hat{s}_{k-1}^*,$$

Prop. 1 leads to $d_k \geq d_{k-1}$. Moreover, $\delta_k \leq r_k$ by definition and it follows that if $d_k > d_{k-1}$ then $\beta^{d_k}/2 \geq \beta^{d_{k-1}}$ and $z_k' \geq 0$. If $d_k = d_{k-1}$ then $\hat{s}_k^* > \hat{s}_{k-1}^*$ implies that $\beta^\mu + r_k + 1 > \beta^\mu + 1$, $r_k > 1$, and $r_k - \delta_k \geq 1$ and $z_k' \geq 0$. Therefore, $\mathbf{z}'$ is on the domain of q_n.

We now analyze η defined in Equation (53) and show that all parcels in $\eta(\mathbf{z}^*)$ and $\eta(\mathbf{z}')$ are equal. Since we break ties upward, Equation (58) shows that

$$\mathrm{fl}(\hat{s}_{k-1}^* + z_k') = \mathrm{fl}\left(\beta^{d_k}(\beta^\mu + (r_k - \delta_k) + 1/2)\right) = \beta^{d_k}(\beta^\mu + (r_k - \delta_k) + 1)$$

$$= \beta^{d_k}(\beta^\mu + r_k + 1) - \beta^{d_k}\delta_k = \mathrm{fl}(\hat{s}_{k-1}^* + z_k^*) - \beta^{d_k}\delta_k = \hat{s}_k^* - \beta^{d_k}\delta_k. \qquad (59)$$

It follows that

$$S_k(\mathbf{z}', \mathrm{Fl}) + z_{k+1}' = \mathrm{fl}(\hat{s}_{k-1}^* + z_k') + z_{k+1}' =$$

$$\left(\hat{s}_k^* - \beta^{d_k}\delta_k\right) + \left(z_{k+1}^* + \beta^{d_k}\delta_k\right) = \hat{s}_k^* + z_{k+1}^* = S_k(\mathbf{z}^*, \mathrm{Fl}) + z_{k+1}^*.$$

This equation leads to

$$S_{k+1}(\mathbf{z}', \mathrm{Fl}) = \mathrm{fl}(S_k(\mathbf{z}', \mathrm{Fl}) + z_{k+1}') = \mathrm{fl}(\hat{s}_k^* + z_{k+1}^*) = S_{k+1}(\mathbf{z}^*, \mathrm{Fl}),$$

and

$$S_{k+1}(\mathbf{z}', \mathrm{Fl}) - (S_k(\mathbf{z}', \mathrm{Fl}) + z_{k+1}') = S_{k+1}(\mathbf{z}^*, \mathrm{Fl}) - (S_k(\mathbf{z}^*, \mathrm{Fl}) + z_{k+1}^*).$$

Therefore, $S_i(\mathbf{z}', \mathrm{Fl}) = S_i(\mathbf{z}^*, \mathrm{Fl})$ for $i < k$ and $i \geq k+1$. It follows that

$$S_i(\mathbf{z}', \mathrm{Fl}) - (S_{i-1}(\mathbf{z}', \mathrm{Fl}) + z_i') = S_i(\mathbf{z}^*, \mathrm{Fl}) - (S_{i-1}(\mathbf{z}^*, \mathrm{Fl}) + z_i^*)$$

for $i < k$ and $i \geq k+1$. For $i = k$, the definition $z_k' := z_k^* - \beta^{d_k}\delta_k$ and Equation (59) yield

$$S_k(\mathbf{z}', \mathrm{Fl}) - (S_{k-1}(\mathbf{z}', \mathrm{Fl}) + z_k') = \mathrm{fl}(\hat{s}_{k-1}^* + z_k') - (\hat{s}_{k-1}^* + z_k') =$$

$$= \left(\mathrm{fl}(\hat{s}_{k-1}^* + z_k^*) - \beta^{d_k}\delta_k\right) - \left(\hat{s}_{k-1}^* + z_k^* - \beta^{d_k}\delta_k\right)$$

$$= S_k(\mathbf{z}^*, \mathrm{Fl}) - (S_{k-1}(\mathbf{z}^*, \mathrm{Fl}) + z_k^*),$$

Therefore, all parcels in the numerators η in Equation (53) are equal for $\mathbf{z}^*$ and $\mathbf{z}'$.

Let us now analyze the denominator D_n of q_n. Note that

$$(n-k+1)z'_k + (n-k)z'_{k+1} =$$

$$(n-k+1)\left(z_k^* - \beta^{d_k}\delta_k\right) + (n-k)\left(z_{k+1}^* + z_k^* + \beta^{d_k}\delta_k\right)$$

$$= (n-k+1)z_k^* + (n-k)z_{k+1}^* - \beta^{d_k}\delta_k.$$

Moreover, $z'_i = z_i^*$ for $i \notin \{k, k+1\}$ and

$$D_n(\mathbf{z}') - D_n(\mathbf{z}^*) = \left(\sum_{i=1}^n (n-i-1)z'_i\right) - \left(\sum_{i=1}^n (n-i-1)z_i^*\right) =$$

$$\left((n-k-1)z'_k + (n-k)z'_{k+1}\right) - \left((n-k-1)z_k^* + (n-k)z_{k+1}^*\right) = -\beta^{d_k}\delta_k.$$

Since the numerators of $q_n(\mathbf{z}')$ and $q_n(\mathbf{z}^*)$ are equal and $\mathbf{z}^*$ is maximal this equation implies that $\beta^{d_k}\delta_k \leq 0$. Therefore, $\delta_k = \min\{1, r_k\} = 0$, and $r_k = 0$.

Finally, for $k = n$, define $\mathbf{z}'$ with $z'_k = z_k^*$ for $k < n$ and $z'_n = z_n^* - \beta^{d_n}r_n$. As before, $\mathbf{z}'$ is in the domain of q_n and $S_k(\mathbf{z}', \text{Fl}) = S_k(\mathbf{z}^*, \text{Fl})$ for $k < n$. For $k = n$, Equation (57) leads to

$$S_{n-1}(\mathbf{z}', \text{Fl}) + z'_n = \hat{s}_{n-1}^* + z_n^* - \beta^{d_n}r_n = \beta^{d_n}(\beta^\mu + 1/2).$$

We break ties upward, $S_n(\mathbf{z}', \text{Fl}) = \text{fl}(S_{n-1}(\mathbf{z}', \text{Fl}) + z'_n) = \beta^{d_n}(\beta^\mu + 1)$ and

$$S_n(\mathbf{z}', \text{Fl}) - \left(S_{n-1}(\mathbf{z}', \text{Fl}) + z'_n\right) = \beta^{d_n}(\beta^\mu + 1) - \beta^{d_n}(\beta^\mu + 1/2) =$$

$$\beta^{d_n}/2 = \beta^{d_n}(\beta^\mu + r_n + 1) - \beta^{d_n}(\beta^\mu + r_n + 1/2)$$

$$= S_n(\mathbf{z}^*, \text{Fl}) - (S_{n-1}(\mathbf{z}^*, \text{Fl}) + z_n^*),$$

and the numerator of q_n in (53) would not change if were to replace $\mathbf{z}^*$ by $\mathbf{z}'$. However, the denominator would be reduced by $\beta^{d_n}r_n$, and this would contradict the maximality of $\mathbf{z}^*$. Therefore $r_n = 0$. In summary, $r_k = 0$ for all k, the z_k^* are as the x_k in Example 4 and we are done. $\qquad\square$

Proof of Lemma 8 Let us write $z_1 := y_0 + y_1$, $z_k := y_k$ for $k > 1$, $s_k := \sum_{i=1}^k z_i$ and $\hat{s}_k = S_k(\mathbf{z}, \text{Fl})$ for $k = 0, \ldots, n$. We prove by induction that

$$|\hat{s}_n - s_n| \leq \frac{u}{1-(n-2)u}\sum_{k=1}^n\left|\sum_{i=1}^n z_i\right|, \tag{60}$$

which is equivalent to Equation (24). For $n = 1$, Equation (60) follows from Lemma 1. We now prove Equation (60) for $n \geq 2$, assuming that it holds for $n-1$. For $\mathbf{w} \in \mathbb{R}^{n-1}$ with $w_1 = \hat{s}_1 + z_2$ and $w_k = y_{k+1}$ for $k > 1$, we obtain by induction that $S_k(\mathbf{w}, \tilde{\text{Fl}}) = \hat{s}_{k+1}$ for $\tilde{\text{Fl}} = \{\text{fl}_2, \ldots, \text{fl}_n\}$,

$$\left|\hat{s}_n - (\hat{s}_1 + z_2) - \sum_{k=3}^n z_k\right| \leq \frac{u}{1-(n-3)u}\left(\sum_{k=2}^n\left|(\hat{s}_1 + z_2) + \sum_{i=3}^k z_i\right|\right)$$

and

$$|\hat{s}_n - s_n| - |\hat{s}_1 - z_1| \leq \frac{u}{1-(n-3)u}\left((n-1)|\hat{s}_1 - z_1| + \sum_{k=2}^n\left|\sum_{i=1}^k z_i\right|\right).$$

26

Since $\hat{s}_1 = \mathrm{fl}_1(z_1)$, Lemma 1 leads to

$$|\hat{s}_n - s_n| \le \frac{u}{1+u}|z_1| + \frac{u}{1-(n-3)u}\left((n-1)\frac{u}{1+u}|z_1| + \sum_{k=2}^{n}\left|\sum_{i=1}^{k}z_i\right|\right).$$

$$= \frac{u}{1+u}\left(1 + \frac{(n-1)u}{1-(n-3)u}\right)|z_1| + \frac{u}{1-(n-3)u}\sum_{k=2}^{n}\left|\sum_{i=1}^{k}z_i\right|.$$

$$\le \frac{u}{1-(n-2)u}\sum_{k=1}^{n}\left|\sum_{i=1}^{k}z_i\right|$$

$$+ \left(\frac{1}{1+u}\left(1 + \frac{(n-1)u}{1-(n-3)u}\right) - \frac{1}{1-(n-2)u}\right)u|z_1| \qquad (61)$$

The software Mathematica shows that

$$\frac{1}{1+u}\left(1 + \frac{(n-1)u}{1-(n-3)u}\right) - \frac{1}{1-(n-2)u} = -\frac{(n-1)u^2}{(1+u)(1-(n-2)u)(1-(n-3)u)},$$

and this number is negative for $n \ge 2$ because $nu < 1$. As a result, Equation (61) implies Equation (60) and we are done. $\qquad\square$

Proof of Lemma 9 Let us start with $z > 0$ and define $m := (\lfloor w \rfloor + \lceil w \rceil)/2$. By Prop. 11, there are three possibilities :

- If $w < m$ then $r = \lfloor w \rfloor$ satisfies Equation (32).

- If $w > m$ then $r = \lceil w \rceil$ satisfies Equation (32).

- If $w = m$ then $r_1 := \lfloor w \rfloor$ and $r_2 := \lceil w \rceil$ satisfy $r_i \in [0,(\beta-1)\beta^\mu)$, $|r_i - w| \le 1/2$ and $\mathrm{fl}(z) = \beta^e(\beta^\mu + r)$ for $r \in \{r_1, r_2\}$. Therefore, Equation (32) is also satisfied.

According to Definition 2, $2u \times \beta^\mu = 1$ and Equation (32) yields

$$\left|\frac{\mathrm{fl}(z) - z}{z}\right| = \frac{|r - w|}{\beta^\mu + w} = \frac{2u|r-w|}{1+2wu} \le \frac{u}{1+2wu}. \qquad (62)$$

When $w \ge 1/2$, this equation implies that

$$\left|\frac{\mathrm{fl}(z) - z}{z}\right| \le \frac{u}{1 + \max\{1, 2w\}u},$$

and when $w < 1/2$, Equation (32) and the fact that r is integer imply that $r = 0$ and

$$\left|\frac{\mathrm{fl}(z) - z}{z}\right| = \frac{w}{\beta^\mu + w} = \frac{2wu}{1+2wu} < \frac{u}{1+u} = \frac{u}{1 + \max\{1, 2w\}u},$$

and we have verified Equation (33). Equation (62) also leads to

$$\left|\frac{\mathrm{fl}(z) - z}{z}\right| \le \frac{u}{1+2wu} \le \frac{u}{1+2(r-1/2)u} = \frac{u}{1+(2r-1)u}$$

and

$$\left|\frac{\mathrm{fl}(z) - z}{\mathrm{fl}(z)}\right| = \frac{|r - w|}{\beta^\mu + r} = \frac{2u|r-w|}{1+2ru} \le \frac{u}{1+2ru}.$$

27

This proves the last equation in Lemma 9 and we are done with $z > 0$. To prove Lemma 9 for $z < 0$, use the argument above for $z' = -z$ and the function m in Prop. 10. $\qquad\square$

Proof of Lemma 10 Let us define $\psi := \sup_{(\mathbf{z},r)\in\mathcal{Z}\times\mathcal{R}} g(\mathbf{z},r)$. If $\psi \leq \varphi$ then $g(\mathbf{z},r) \leq \varphi$ for all $(\mathbf{z},r) \in \mathcal{Z} \times \mathcal{R}$ and we are done. Let us then assume that $\varphi < \psi$ and let $\{(\mathbf{z}_k,r_k), k \in \mathbb{N}\} \subset \mathcal{Z} \times \mathcal{R}$ be a sequence such that $\lim_{k\to\infty} g(\mathbf{z}_k,r_k) = \psi$ and $g(\mathbf{z}_k,r_k) > \varphi$. It follows that $\mathbf{z}_k \in \mathcal{A}$ for each k and there exists $\lambda_k \in \mathcal{L}$ and $r'_k \in \mathcal{R}$ for which $\mathbf{z}'_k := \lambda_k \mathbf{z}_k \in \mathcal{K}$ satisfies $h(\mathbf{z}'_k, r'_k) = \lambda_k h(\mathbf{z}_k, r_k)$. Since the sequence $\mathbf{z}'_k$ is contained in the compact set $\mathcal{K}$, it has a subsequence which converges to $\mathbf{z}^* \in \mathcal{K}$, and we may assume that this subsequence is $\mathbf{z}'_k$ itself. The scaling properties of f lead to

$$f(\mathbf{z}'_k, h(\mathbf{z}'_k, r'_k)) = f(\lambda_k \mathbf{z}_k, \lambda_k h(\mathbf{z}_k, r_k)) \geq f(\mathbf{z}_k, h(\mathbf{z}_k, r_k)) = g(\mathbf{x}_k, r_k)$$

and

$$\liminf_{k\to\infty} f(\mathbf{z}'_k,\, h(\mathbf{z}'_k, r'_k)) \geq \liminf_{k\to\infty} g(\mathbf{z}_k, r_k) = \lim_{k\to\infty} g(\mathbf{z}_k, r_k) = \psi.$$

Since h is tight, there exists $r^* \in \mathcal{R}$ and a subsequence $\mathbf{z}'_{n_k}$ such that $\lim_{k\to\infty} h(\mathbf{z}'_{n_k}, r'_{n_k}) = h(\mathbf{z}^*, r^*)$. By the upper semi-continuity of f and the maximality of ψ we have

$$\psi \geq g(\mathbf{z}^*, r^*) = f(\mathbf{z}^*, h(\mathbf{z}^*, r^*))$$

$$\geq \limsup_{k\to\infty} f(\mathbf{z}'_{n_k},\, h(\mathbf{z}'_{n_k}, r'_{n_k})) \geq \liminf_{k\to\infty} f(\mathbf{z}'_k,\, h(\mathbf{z}'_k, r'_k)) \geq \psi.$$

Therefore, $g(\mathbf{z}^*, r^*) = \psi$ and we are done. $\qquad\square$

4.5 Corollaries

In this section we prove some of the corollaries stated in the article. The remaining corollaries are proved in the extended version.

Proof of Corollary 2 $\mathrm{fl}(x^2) \geq v$ by Monotonicity (Prop. 9), and Lemma 1 yield

$$|z - |x|| \, (z + |x|) = |z^2 - x^2| = |\mathrm{fl}(x^2) - x^2| \leq \frac{|x|^2 u}{1 + u} \tag{63}$$

for $z := \sqrt{\mathrm{fl}(x^2)} > 0$. It follows that $\delta := |z - |x|| / |x|$ satisfies

$$\delta \leq \frac{u}{1+u} \frac{|x|}{|x|+z} \leq \frac{u}{1+u} < u = \beta^{-\mu}/2 \leq \frac{1}{4} \quad \Rightarrow \quad 1 - \delta > 0.$$

Equation (63) leads to

$$\frac{u}{1+u} \geq \delta \frac{z+|x|}{|x|} \geq \delta \frac{2|x| - |z-|x||}{|x|} = \delta(2-\delta) > 0,$$

and

$$1 - \delta = \sqrt{(1-\delta)^2} = \sqrt{1 - \delta(2-\delta)} \geq \sqrt{1 - \frac{u}{1+u}} = \frac{1}{\sqrt{1+u}},$$

and

$$\delta \leq 1 - \frac{1}{\sqrt{1+u}} = \frac{u}{2}\psi \quad \text{for} \quad \psi := \frac{2}{u} \frac{\sqrt{1+u}-1}{\sqrt{1+u}} = \frac{2}{1+u+\sqrt{1+u}} < 1. \tag{64}$$

Let $\mathcal{P}$ be the complete system with the same β and μ as $\mathcal{F}$. By Prop. 14 there exists $\tilde{\mathrm{fl}}$ which rounds to nearest in $\mathcal{P}$ and is such that $\tilde{\mathrm{fl}}(w) = \mathrm{fl}(w)$ for w with $|w| \geq v_{\mathcal{F}}$. In particular, $\mathrm{fl}(x^2) = \tilde{\mathrm{fl}}(x^2)$. Since $v < 1$ and $x^2 \geq v$ we have that $|x| \geq v$ and by Prop. 5 there exists an exponent e for $\mathcal{F}$ and $r \in [0, (\beta-1)\beta^\mu) \cap \mathbb{Z}$ such that $|x| = \beta^e (\beta^\mu + r)$. This implies that $\beta^{e+\mu} \leq |x| < \beta^{e+\mu+1}$. The numbers $\beta^{2e+2\mu}$ and $\beta^{2e+2\mu+2}$ are in $\mathcal{P}$ (although $\beta^{2e+2\mu}$ may not be in $\mathcal{F}$) and, by the monotonicity of $\tilde{\mathrm{fl}}$,

$$\beta^{2e+2\mu} \leq \mathrm{fl}(x^2) = \tilde{\mathrm{fl}}(x^2) \leq \beta^{2e+2\mu+2},$$

and $\beta^{e+\mu} \leq \sqrt{\mathrm{fl}(x^2)} = z = \sqrt{\mathrm{fl}(x^2)} \leq \beta^{e+\mu+1}$. By Prop. 1 and Prop. 2, $z = \beta^e (\beta^\mu + w)$ with $0 \leq w \leq (\beta - 1)\beta^\mu$. As a result,

$$\delta = \frac{|\beta^e (\beta^\mu + w) - \beta^e (\beta^\mu + r)|}{\beta^e (\beta^\mu + r)} = \frac{|w - r|}{\beta^\mu + r},$$

and recalling that $2u\beta^\mu = 1$ and using Equation (64) we obtain

$$|w - r| \leq \frac{1}{4} \psi \left(1 + \beta^{-\mu} r\right) = \frac{1}{4} \psi (1 + 2ru). \tag{65}$$

There are two possibilities: either

$$\frac{1}{4} \psi (1 + 2ru) < \frac{1}{2} \tag{66}$$

or

$$\frac{1}{4} \psi (1 + 2ru) \geq \frac{1}{2}. \tag{67}$$

In case (66) $|w - r| < 1/2$ by Equation (65), Prop. 11 shows that $\mathrm{fl}(z) = |x|$ and Corollary 2 holds for x. For instance, if $\beta = 2$ then $2ru < 2(2-1)2^\mu u = 1$ and r satisfies Equation (66) because $\psi < 1$. Therefore, we have proved Corollary 2 for $\beta = 2$.

In order to complete the proof for the cases in which Equation (67) is valid, it suffices to show that

$$\frac{|x|}{\mathrm{fl}(z)} = \frac{|x|}{\mathrm{fl}\left(\sqrt{\mathrm{fl}(x^2)}\right)} < 1 + u = \beta^{-\mu} (\beta^\mu + 1/2), \tag{68}$$

because this equation implies that $\mathrm{fl}\left(|x| / \sqrt{\mathrm{fl}(z)}\right) \leq 1$ by Prop. 11 and monotonicity.

We first show that Equation (68) is valid when

$$\zeta := 1 + 2ru > (1 + u)^{3/2} + 1 + u. \tag{69}$$

In fact, for ψ in Equation (64), Equation (69) is equivalent to

$$\zeta > \frac{1 + u}{1 - \frac{\psi}{2} (1 + u)}, \qquad \frac{\zeta}{1 + u} - \frac{\psi}{2} \zeta - 1 > 0 \qquad \text{and} \qquad \zeta - \frac{\psi}{2} \zeta u - u > \frac{\zeta}{1 + u},$$

and can also be written as

$$1 + u > \frac{\zeta}{\zeta - \frac{\psi}{2} \zeta u - u} \qquad \text{or} \qquad \frac{1 + 2ru}{1 + 2ru - \frac{\psi}{2}(1 + 2ru)u - u} < 1 + u. \tag{70}$$

Since $w \in [0, (\beta - 1)\beta^\mu]$, Prop. 11 implies that $\mathrm{fl}(z) \geq \beta^e (\beta^\mu + w - 1/2)$ and

$$\frac{|x|}{\mathrm{fl}(z)} \leq \frac{\beta^e (\beta^\mu + r)}{\beta^e (\beta^\mu + w - 1/2)} = \frac{1 + 2ru}{1 + 2wu - u},$$

29

because $2u\beta^\mu = 1$. Equations (65) shows that $w \geq r - \psi(1 + 2ru)/4$ and

$$\frac{|x|}{\mathrm{fl}(z)} \leq \frac{1 + 2ru}{1 + 2ru - \frac{\psi}{2}(1 + 2ru)u - u}. \tag{71}$$

Equations (70) and (71) lead to Equation (68). Therefore, Equation (69) implies Equation (68) and Corollary 2 is valid when Equation (69) is satisfied.

In the case opposite to Equation (69) we have that

$$2ru \leq (1 + u)^{3/2} + u = 1 + \frac{5}{2}u + \frac{3}{8\sqrt{1 + \xi_1}}u^2 \tag{72}$$

for some $\xi_1 \in [0, u]$. Since r is integer and $2u = \beta^{-\mu}$, Equation (72) implies that

$$r < \beta^\mu + \frac{5}{4} + \frac{3}{16}u < \beta^\mu + 2 \Rightarrow r \leq \beta^\mu + 1. \tag{73}$$

Moreover, Equation (67) leads to

$$r \geq \beta^\mu \frac{2 - \psi}{\psi} = \beta^\mu \left(u + \sqrt{1 + u}\right) = \beta^\mu \left(1 + \frac{3}{2}u - \frac{1}{8(1 + \xi_2)^{3/2}}u^2\right)$$

for some $\xi_2 \in [0, u]$, and since r is integer and $2u = \beta^{-\mu}$, we have that

$$r \geq \beta^\mu + \frac{3}{4} - \frac{1}{16(1 + \xi_2)^{3/2}}u \Rightarrow r \geq \beta^\mu + 1. \tag{74}$$

Equations (73) and (74) show that there is just one r left: $r = \beta^\mu + 1$, which corresponds to $|x| = \beta^e(2\beta^\mu + 1)$. It follows that

$$x^2 = \beta^{2e}\left(4\beta^{2\mu} + 4\beta^\mu + 1\right) = \beta^{2e+\mu}\left(\beta^\mu + \left(3\beta^\mu + 4 + \beta^{-\mu}\right)\right).$$

If $\beta \geq 5$ then $3\beta^\mu + 4 + \beta^{-\mu} < (\beta - 1)\beta^\mu$ and Prop. 11 implies that

$$\mathrm{fl}(x^2) = 4\beta^{2e+\mu}\left(\beta^\mu + 1\right) \Rightarrow z = \sqrt{\mathrm{fl}(x^2)} = 2\beta^{e+\mu}\sqrt{1 + \beta^{-\mu}}$$

$$= 2\beta^{e+\mu}\left(1 + \frac{1}{2}\beta^{-\mu} - \frac{\theta_5}{2}\beta^{-\mu}\right),$$

where, for some $\xi_5 \in [0, \beta^{-\mu}]$,

$$0 \leq \theta_5 := \frac{1}{4(1 + \xi_5)^{3/2}}\beta^{-\mu} \leq \frac{1}{4} \times \frac{1}{5} = \frac{1}{20}.$$

Therefore, $z := \sqrt{\mathrm{fl}(x^2)} = \beta^e(2\beta^\mu + 1 - \theta_5)$ and the bound $|\theta_5| \leq 1/20$ and Prop. 11 imply that $\mathrm{fl}(z) = \beta^e(2\beta^\mu + 1) = |x|$ and we are done with the case $\beta \geq 5$.

For $\beta = 3$, the critical x is $3^e(2 \times 3^\mu + 1)$ and

$$x^2 = 3^{2e}\left(4 \times 3^{2\mu} + 4 \times 3^\mu + 1\right) = 3^{2e+\mu+1}\left(3^\mu + 3^{\mu-1} + 1 + \left(\frac{1}{3} + 3^{-\mu-1}\right)\right)$$

The bound

$$\frac{1}{3} + 3^{-\mu-1} \leq \frac{1}{3} + \frac{1}{9} = \frac{4}{9} < 1/2$$

30

and Prop. 11 lead to

$$\text{fl}(x^2) = 3^{2e+\mu+1}\left(3^\mu + 3^{\mu-1} + 1\right) = 4 \times 3^{2e+2\mu}\left(1 + \frac{3}{4} \times 3^{-\mu}\right)$$

and

$$z := \sqrt{\text{fl}(x^2)} = 2 \times 3^{e+\mu}\left(1 + \frac{3}{8} \times 3^{-\mu} - \frac{\theta_3}{2} \times 3^{-\mu}\right) = 3^e\left(2 \times 3^\mu + \frac{3}{4} - \theta_3\right)$$

where, for some $\xi_3 \in [0, 1/3]$,

$$0 \le \theta_3 := \frac{1}{4(1+\xi_3)^{3/2}} \times \frac{9}{16} \times 3^{-\mu} \le \frac{3}{64}.$$

Since $3/4 - 3/64 = 45/64 > 1/2$, Prop. 11 shows that $\text{fl}(z) = |x|$ when $\beta = 3$.

Finally, for $\beta = 4$, we care about $x = 4^e\left(2 \times 4^\mu + 1\right)$ and

$$x^2 = 4^{2e}\left(4 \times 4^{2\mu} + 4 \times 4^\mu + 1\right) = 4^{2e+1+\mu}\left(4^\mu + 1 + 4^{-\mu-1}\right),$$

$4^{-\mu-1} < 1/2$ and Prop. 11 yields

$$\text{fl}(x^2) = 4^{2e+1+\mu}\left(4^\mu + 1\right) = 4^{2e+1+2\mu}\left(1 + 4^{-\mu}\right).$$

It follows that

$$z := \sqrt{\text{fl}(x^2)} = 2 \times 4^{e+\mu}\sqrt{1 + 4^{-\mu}} = 2 \times 4^{e+\mu}\left(1 + \frac{1}{2} \times 4^{-\mu} - \frac{\theta_4}{2} \times 4^{-\mu}\right)$$

where, for some $\xi_4 \in [0, 1/4]$,

$$0 < \theta_4 := \frac{1}{4\sqrt{1+\xi_4}}4^{-\mu} < \frac{1}{16}.$$

Therefore, $z = 4^{e+1}\left(2 \times 4^\mu + 1 - \theta_4\right)$, $\text{fl}(z) = |x|$ and we are done. $\qquad\square$

Proof of Corollary 5 Let $\mathcal{P}$ be the perfect system corresponding to β and μ and $\tilde{\text{Fl}}$ the rounding tuple in Prop. 15 or 16, depending on whether $\mathcal{F}$ is an IEEE system or a MPFR system. As in the proof of Lemma 5, we define $z_1 := y_0 + y_1$, $z_k := y_k$ for $2 \le k \le n$, $s_k := \sum_{i=1}^{k} z_i$ and $\hat{s}_k := S_k(\mathbf{x}, \text{Fl})$ for $k = 0, \ldots, n$. We also use the set $\mathcal{T}$ of indexes k in $[1, n]$ such that $|S_{k-1}(\mathbf{z}, \text{Fl}) + z_k| < \tau$ for

$$\tau := \beta^{e_\alpha}(\beta^\mu + r) \qquad \text{and} \qquad r := \beta^\mu \frac{\beta - 1}{2}.$$

Note that $\tau \in \mathcal{E}_{e_\alpha} \subset \mathcal{F}$ because r is integer and $r < (\beta - 1)\beta^\mu$. The threshold τ was chosen because $v = \beta^{e_\alpha + \mu}$,

$$\tau = \frac{\beta + 1}{2}v < \beta v \tag{75}$$

and Prop. 13 shows that

$$|z| \le \beta v \Rightarrow |\text{fl}(z) - z| \le \alpha/2, \tag{76}$$

where $\alpha = \beta^{e_\alpha}$ for IEEE systems and $\alpha = v = \beta^{e_\alpha + \mu}$ for MPFR systems.

Let $m \in [0, n]$ be the size of $\mathcal{T}$. We prove by induction that

$$\eta(\mathbf{z}, \mathrm{Fl}) := \sum_{k=1}^{n} |S_k(\mathbf{z}, \mathrm{Fl}) - (S_{k-1}(\mathbf{z}, \mathrm{Fl}) + z_k)|$$

satisfies

$$\eta(\mathbf{z}, \mathrm{Fl}) \leq \frac{m\alpha}{2} + \frac{(n-m)u}{1+(n-m)u}\left(\frac{m\alpha}{2} + \sum_{k=1}^{n} |z_k|\right). \tag{77}$$

If $m = 0$ then $S_n(\mathbf{z}, \mathrm{Fl}) = S_n(\mathbf{z}, \tilde{\mathrm{Fl}})$ and Equation (77) follows from Lemma 5. Assuming that Equation (77) holds for $m - 1$, let us show that it holds for m. If $|s_1| < \tau$ then the sum $(\hat{s}_1 + z_2) + \sum_{k=3}^{n} z_k$ has $n - 1$ parcels and there are $m - 1$ indices in $[2, n] \cap \mathcal{T}$. As a result $(n-1) - (m-1) = n - m$, Equation (76), the identity $s_1 = z_1$ and induction yield

$$\eta(\mathbf{z}, \mathrm{Fl}) = |\hat{s}_1 - s_1| + \left(\sum_{k=2}^{n} |\hat{s}_k - (\hat{s}_{k-1} + z_k)|\right)$$

$$\leq \frac{\alpha}{2} + \left(\frac{(m-1)\alpha}{2} + \frac{(n-m)u}{1+(n-m)u}\left(\frac{(m-1)\alpha}{2} + |\hat{s}_1 + z_2| + \sum_{k=3}^{n} |z_k|\right)\right)$$

$$\leq \frac{m\alpha}{2} + \frac{(n-m)u}{1+(n-m)u}\left(\left(|\hat{s}_1 - s_1| - \frac{\alpha}{2}\right) + \frac{m\alpha}{2} + |s_1| + \sum_{k=2}^{n} |z_k|\right)$$

$$\leq \frac{m\alpha}{2} + \frac{(n-m)u}{1+(n-m)u}\left(\frac{m\alpha}{2} + \sum_{k=1}^{n} |z_k|\right).$$

Therefore, Equation (77) holds when $|s_1| < \tau$. Let us then assume that $|s_1| \geq \tau$ and define $\ell \in [2, n]$ as the first index such that $|\hat{s}_{\ell-1} + z_\ell| < \tau$,

$$S := \sum_{k=1}^{\ell-1} |z_k|, \qquad p := \ell - 1 \qquad \text{and} \qquad q := n - m - \ell + 1. \tag{78}$$

Monotonicity and $\tau \in \mathcal{F}$ implies that $|\hat{s}_\ell| = |\mathrm{fl}_\ell(\hat{s}_{\ell-1} + z_\ell)| \leq \tau$ and the proof of Lemma 5, Equation (76) and induction yield

$$\eta(\mathbf{z}, \mathrm{Fl}) = \sum_{k=1}^{\ell-1} |\hat{s}_k - (s_{k-1} + z_k)| + |\hat{s}_\ell - \hat{s}_{\ell-1} - z_\ell| + \sum_{k=\ell+1}^{n} |\hat{s}_k - (\hat{s}_{k-1} + z_k)|$$

$$\leq \frac{pu}{1+pu}S + \frac{\alpha}{2} + \left(\frac{(m-1)\alpha}{2} + \frac{qu}{1+qu}\left(\frac{(m-1)\alpha}{2} + |\hat{s}_\ell + z_{\ell+1}| + \sum_{k=\ell+2}^{n} |z_k|\right)\right)$$

$$\leq \frac{pu}{1+pu}S + \frac{m\alpha}{2} + \frac{qu}{1+qu}\left(\frac{(m-1)\alpha}{2} + \tau + \sum_{k=\ell+1}^{n} |z_k|\right). \tag{79}$$

If $S \geq 7\tau/6$ then

$$\frac{p}{1+pu}S + \frac{q}{1+qu}\tau \leq \left(\frac{p}{1+pu} + \frac{6}{7}\frac{q}{1+qu}\right)S \leq \frac{(p+q)u}{1+(p+q)u}S - \Delta S$$

for

$$\Delta := \frac{p+q}{1+(p+q)u} - \left(\frac{p}{1+pu} + \frac{6}{7}\frac{q}{1+qu}\right).$$

32

The software Mathematica shows that

$$\Delta = q\frac{1+qu-6(2+qu+pu)pu}{(1+pu)(1+qu)(1+(p+q)u)}$$

and the hypothesis $20nu \leq 1$ implies that $\Delta \geq 0$. Therefore, if $S \geq 7\tau/6$ then Equation (79) leads to

$$\eta(\mathbf{z},\mathrm{Fl}) \leq \frac{m\alpha}{2} + \frac{(p+q)u}{1+(p+q)u}\left(\frac{m\alpha}{2}+S+\sum_{k=\ell+1}^{n}|z_k|\right),$$

and Equation (77) follows from Equation (78). We can then assume that $S < 7\tau/6$ and, for $1 \leq k < \ell$, Lemma 5 leads to

$$|\hat{s}_k| \leq |s_k|+|\hat{s}_k-s_k| \leq \left(1+\frac{ku}{1+ku}\right)S < 1.05 \times \frac{7}{6}\frac{\beta+1}{2}v < \frac{2}{3}(\beta+1)v \leq \beta v,$$

and Equation (76) implies that $|\hat{s}_k - (\hat{s}_k+z_k)| \leq \alpha/2$ for $1 \leq k < \ell$. It follows that

$$\sum_{k=1}^{\ell-1}|\hat{s}_k-(s_{k-1}-z_k)| \leq (\ell-1)\alpha/2 = p\alpha/2. \tag{80}$$

The identity $uv = \alpha/2$ for IEEE systems and the inequality $uv = u\alpha \leq \alpha/4$ for MPFR systems, the hypothesis $20nu \leq 1$ and the fact that

$$\sum_{k=1}^{\ell-1}|z_k| \geq |z_1| = |s_1| \geq \tau = (\beta+1)v/2 \geq \frac{3}{2}v$$

imply that

$$\frac{p\alpha}{2} \leq \frac{pvu}{2} \leq \frac{2}{3}(1+pu)\frac{pu}{1+pu}\sum_{k=1}^{\ell-1}|z_k|$$

$$\leq \frac{2}{3} \times \frac{21}{20} \times \frac{pu}{1+pu}\sum_{k=1}^{\ell-1}|z_k| = \frac{7}{10}\frac{pu}{1+pu}\sum_{k=1}^{\ell-1}|z_k|.$$

Using induction as in Equation (79) and the bounds in the previous equation and in Equation (80), and recalling that $|z_1| \geq \tau$, we obtain

$$\eta(\mathbf{z},\mathrm{Fl}) \leq \frac{p\alpha}{2} + \frac{\alpha}{2} + \frac{(m-1)\alpha}{2} + \frac{qu}{1+qu}\left(\frac{(m-1)\alpha}{2}+\tau+\sum_{k=\ell+1}^{n}|z_k|\right)$$

$$\leq \frac{7}{10}\frac{pu}{1+pu}\sum_{k=1}^{n}|z_k| + \frac{m\alpha}{2} + \frac{qu}{1+qu}\left(\frac{m\alpha}{2}+\sum_{k=1}^{n}|z_k|\right). \tag{81}$$

According to the software Mathematica,

$$\frac{p+q}{1+(p+q)u} - \left(\frac{q}{1+qu}+\frac{7}{10}\frac{p}{1+pu}\right) = p\frac{3+3pu-7(2+qu+pu)qu}{10(1+pu)(1+qu)(1+(p+q)u)},$$

and this number is positive due to the hypothesis $20nu \leq 1$. As a result, Equation (81) implies Equation (77) and this concludes the inductive proof of Equation (77). This equation leads to

$$\left|\mathrm{Fl}\left(\sum_{k=0}^{n}y_k\right)-\sum_{k=0}^{n}y_k\right| \leq \frac{m\alpha}{2} + \frac{(n-m)u}{1+(n-m)u}\left(\frac{m\alpha}{2}+\sum_{k=0}^{n}|y_k|\right), \tag{82}$$

and implies Equation (18) because $0 \leq m \leq n$.

Finally, when the additional condition in Corollary 5 holds we have that

$$\sum_{k=0}^{n} |y_k| \geq \theta \frac{(1+nu)^2}{u} \alpha$$

for $1 > \theta := \frac{1}{(1+nu)^2} \geq (20/21)^2 > 0.9$ and the software Mathematica shows that

$$\frac{nu}{1+nu} \theta \frac{(1+nu)^2}{u} \alpha - \left(\frac{m\alpha}{2} + \frac{(n-m)u}{1+(n-mu)u} \left(\frac{m\alpha}{2} + \theta \frac{(1+nu)^2}{u} \alpha \right) \right)$$

$$= \alpha m \frac{2\theta - 1 + 2u((\theta-1)n+m)}{1+(n-m)u} \geq \alpha m \frac{0.8 - 2u(0.1n-m)}{1+(n-m)u} > 0,$$

and Equation (17) follows from Equation (82). $\square$

Proof of Corollary 7 Define $z_1 := y_0 + y_1$ and $z_k := y_k$ for $2 \leq k \leq n$, $s_k := \sum_{i=1}^{k} z_i$ and $\hat{s}_k := S_k(\mathbf{x}, \mathrm{Fl})$ for $k = 0, \ldots, n$. Let $\mathcal{P}$ be the perfect system corresponding to β and μ and $\tilde{\mathrm{Fl}}$ the rounding tuple in Props. 15 or 16, depending on whether $\mathcal{F}$ is an IEEE system or a MPFR system. By definition of $\tilde{\mathrm{Fl}}$, we have that $\mathrm{fl}_k(s_{k-1}+z_k) = \tilde{\mathrm{fl}}_k(s_{k-1}+z_k)$ when $|s_{k-1}+z_k| \geq v$. Let $\mathcal{T}$ be the set of indexes k in $[1,n]$ such that $|S_{k-1}(\mathbf{z}, \mathrm{Fl}) + z_k| < v$ and $m \in [0,n]$ its size. We prove by induction that

$$\left| \mathrm{Fl}\left(\sum_{k=1}^{n} z_k \right) - \sum_{k=1}^{n} z_k \right| \leq (1 + 2(n-m)u) m \frac{\alpha}{2} + \frac{1 - mu/2}{1-(n-2)u} u \sum_{k=1}^{n} \left| \sum_{i=1}^{k} z_i \right|. \quad (83)$$

When $m = 0$ we have that $\hat{s}_n = S_n(\mathbf{z}, \tilde{\mathrm{Fl}})$ and Equation (83) follows from Lemma 8. Assuming that Equation (83) holds for $m-1$, let us prove it for m. Let ℓ be the last element of $\mathcal{T}$ (Note that $\ell \geq m$.) It follows that $|\hat{s}_{k-1} - z_k| \geq v$ for $k > \ell$ and $\hat{s}_k = \tilde{\mathrm{fl}}_k(\hat{s}_\ell + \sum_{i=\ell+1}^{k} z_i)$ for $k > \ell$. The proof of Lemma 8 shows that

$$\left| \hat{s}_n - \left(\hat{s}_\ell + \sum_{k=\ell+1}^{n} z_k \right) \right| \leq \frac{u}{1-((n-\ell)-2)u} \sum_{k=\ell+1}^{n} \left| \hat{s}_\ell + \sum_{i=\ell+1}^{k} z_i \right|$$

$$\leq \frac{u}{1-(n-\ell-2)u} \sum_{k=\ell+1}^{n} \left(|\hat{s}_\ell - s_\ell| + \left| \sum_{i=1}^{k} z_i \right| \right)$$

$$= Au \left((n-\ell) |\hat{s}_\ell - s_\ell| + \sum_{k=\ell+1}^{n} \left| \sum_{i=1}^{k} z_i \right| \right), \quad (84)$$

for

$$A := \frac{1}{1-(n-\ell-2)u}.$$

Moreover, $|\hat{s}_{\ell-1} + z_\ell| < v$ and, by induction and Prop. 13,

$$|\hat{s}_\ell - s_\ell| \leq |\hat{s}_\ell - \hat{s}_{\ell-1} - z_\ell| + |\hat{s}_{\ell-1} - s_{\ell-1}|$$

$$\leq \frac{\alpha}{2} + (1 + 2(\ell-m)u)(m-1)\frac{\alpha}{2} + \frac{(1-(m-1)u/2)u}{1-(\ell-3)u} \sum_{k=1}^{\ell-1} \left| \sum_{i=1}^{k} z_i \right|$$

$$= (m + 2(\ell - m)(m - 1)u)\frac{\alpha}{2} + Cu\sum_{k=1}^{\ell-1}\left|\sum_{j=1}^{k}z_j\right|. \tag{85}$$

for

$$C := \frac{1 - (m-1)u/2}{1 - (\ell - 3)u}.$$

Combining Equations (84) and (85) we obtain

$$|\hat{s}_n - s_n| \le \left|\hat{s}_n - \left(\hat{s}_\ell + \sum_{k=\ell+1}^{n}z_k\right)\right| + |\hat{s}_\ell - s_\ell| \le$$

$$\le (1 + A(n - \ell)u)|\hat{s}_\ell - s_\ell| + Au\sum_{k=\ell+1}^{n}\left|\sum_{i=1}^{k}z_i\right|$$

$$\le D(m + 2(\ell - m)(m - 1)u)\frac{\alpha}{2} + DCu\sum_{k=1}^{\ell-1}\left|\sum_{i=1}^{k}z_i\right| + Au\sum_{k=\ell+1}^{n}\left|\sum_{i=1}^{k}z_i\right|, \tag{86}$$

for

$$D := 1 + A(n - \ell)u = \frac{1 + 2u}{1 - (n - \ell - 2)u}.$$

We now show that $Q < 1$ for

$$Q := \frac{D(m + 2(\ell - m)(m - 1)u)}{(1 + 2(n - m)u)m} = \frac{(1 + 2u)(1 + 2(\ell - m)(1 - 1/m)u)}{(1 - (n - \ell - 2)u)(1 + 2(n - m)u)}.$$

It easy to see that $Q < 1$ when $\ell = n$. Since $20nu \le 1$, when $\ell < n$ we have

$$Q < \frac{(1 + 2u)(1 + 2(\ell - m)u)}{(1 - (n - \ell - 2)u)(1 + 2(n - m)u)}$$

$$= \frac{1 + (2\ell - 2m + 2)u + 4(\ell - m)u^2}{1 + (n + \ell - 2m + 2)u - 2(n - \ell - 2)(n - m)u^2}$$

$$= \frac{1 + (2\ell - 2m + 2)u + 4(\ell - m)u^2}{1 + (2\ell - 2m + 2)u + (n - \ell)\left(1 - 2\frac{(n - \ell - 2)}{n - \ell}(n - m)u\right)u}$$

$$\le \frac{1 + (2\ell - 2m + 2)u + 0.2u}{1 + (2\ell - 2m + 2)u + (1 - 0.1)u} < 1.$$

Therefore, $Q < 1$ and, equivalently,

$$D(m + 2(\ell - m)(m - 1)u) \le (1 + 2(n - m)u)m. \tag{87}$$

Moreover,

$$DC = \frac{1 + 2u}{1 - (n - \ell - 2)u}\frac{1 - (m - 1)u/2}{1 - (\ell - 3)u} = \frac{(1 + 2u)(1 - (m - 1)u/2)}{1 - (n - 5)u + (\ell - 3)(n - \ell - 2)u^2}.$$

Note that the function $h(\ell) := (\ell - 3)(n - \ell - 2)$ is concave. Therefore its minimum in the interval $[1, n]$ is at the endpoints. Since $h(1) = h(n) = -2(n - 3)$, we have

$$DC \le \frac{(1 + 2u)(1 - (m - 1)u/2)}{1 - (n - 5)u - 2(n - 3)u^2},$$

35

and the software Mathematica shows that

$$\frac{(1+2u)(1-(m-1)u/2)}{1-(n-5)u-2(n-3)u^2} - \frac{1-mu/2}{1-(n-2)u} = -u\frac{1-2u-mu+nu}{2(1+3u-nu)(1+2u-nu)} < 0,$$

where the last inequality follows from the hypothesis $20nu \leq 1$. Therefore,

$$DC \leq \frac{1-mu/2}{1-(n-2)u}.$$

Not also that, since $\ell \geq m$ and $20nu \leq 1$,

$$A - \frac{1-mu/2}{1-(n-2)u} = \frac{1-(n-2)u-(1-mu/2)(1-(n-\ell-2)u)}{(1-(n-2)u)(1-(n-\ell-2)u)}$$

$$= -\frac{\ell - m/2\,(1-(n-\ell-2)u)}{(1-(n-2)u)(1-(n-\ell-2)u)}u < 0,$$

and

$$A \leq \frac{1-mu/2}{1-(n-2)u}.$$

The bounds on DC and A above, combined with Equations (86) and (87) imply Equation (83), and we completed the inductive proof of this equation.

Finally, when $u\sum_{k=1}^{n}|\sum_{i=0}^{n}y_i| \geq n\alpha$ Equation (83) leads to

$$\left| \mathrm{Fl}\left(\sum_{k=0}^{n} y_k\right) - \sum_{k=0}^{n} y_k \right| \leq \theta_m u \sum_{k=1}^{n}\left|\sum_{i=0}^{k} y_i\right|,$$

for

$$\theta_m := (1+2(n-m)u)\frac{m}{2n} + \frac{1-mu/2}{1-(n-2)u}.$$

The derivative of θ_m with respect to m is

$$\frac{1-u(4m-2)-2u^2(n^2-2mn+4m-2n)}{2n(1+2u-nu)},$$

and it is positive because $20mu \leq 20nu \leq 1$. Thus, θ_m is maximized for $m = n$ and

$$\theta_m \leq \frac{m}{2} + \frac{1-nu/2}{1-(n-2)u} = \frac{3-2(n-1)u}{2(1-(n-2)u)},$$

and Equation (26) holds because

$$\frac{3-2(n-1)u}{2(1-(n-2)u)} - \frac{3}{2}\left(1+\frac{nu}{2}\right) = -u\frac{8+n(1-3(n-2)u)}{1-(n-2)u} < 0$$

when $20nu \leq 1$. □

5 Extended version

In this part of the article we prove Lemmas 4 and 6, the corollaries which were not proved in the previous sections, and the propositions. We try to prove every assertion we make, no matter how trivial it may sound. In all propositions $\mathcal{F}$ is a floating point system, $z \in \mathbb{R}$, $x \in \mathcal{F}$, fl rounds to nearest in $\mathcal{F}$, and u, e_α, μ α and ν are the numbers related to this system in Definitions 2, 6, 7, 9, 11 and 12.

5.1 Proofs of Lemmas 4 and 6

In this section we prove Lemmas 4 and 6.

Proof of Lemma 4   If $b - a < \beta v$ then Lemma 4 follows from Lemma 2. Therefore, we can assume that $b - a \geq \beta v$. Prop. 2 implies that $a = \beta^d (\beta^\mu + r)$ and $b = \beta^e (\beta^\mu + s)$ with $d, e \in \mathbb{Z}$ and $r, s \in [0, (\beta - 1) \beta^\mu)$. Since $a \leq b \leq 2a$ and $\beta \geq 2$,

$$\beta^d (\beta^\mu + r) \leq \beta^e (\beta^\mu + s) \leq 2\beta^d (\beta^\mu + r) \leq \beta^{d+1} (\beta^\mu + r).$$

Prop. 1 shows that $d \leq e \leq d + 1$ and either (i) $e = d$ or (ii) $e = d + 1$. In case (i) $b - a = \beta^e (s - r) \geq \beta v$. Since $0 \leq s - r < (\beta - 1) \beta^\mu$ and $b - a \geq v$, Prop. 5 implies that $b - a \in \mathcal{F}$. In case (ii) $0 < b - a = \beta^d t$ for $t := ((\beta - 1) \beta^\mu + \beta s - r) > 0$ and

$$b - a \leq a \Rightarrow t \leq \beta^\mu + r < \beta^{1+\mu}.$$

This bound, the assumption $b - a \geq \beta v$ and Prop. 5 imply that $z \in \mathcal{F}$. $\square$

Proof of Lemma 6   The function g_k has first derivative

$$g_k{}'(u) = -g_k(u) \sum_{i=1}^{k} \frac{n_i}{1 + n_i u}$$

and second derivative

$$g_k{}''(u) = g_k(u) \left(\left(\sum_{i=1}^{k} \frac{n_i}{1 + n_i u} \right)^2 + \sum_{i=1}^{k} \frac{n_i^2}{(1 + n_i u)^2} \right) > 0,$$

and, therefore, it is convex. Similarly, the function f_k has first derivative

$$f_k{}'(u) = f_k(u) \sum_{i=1}^{k} \frac{n_i}{(1 + n_i u)(1 + 2n_i u)}$$

and second derivative

$$f_k{}''(u) = f_k(u) \left(\left(\sum_{i=1}^{k} \frac{n_i}{(1 + n_i u)(1 + 2n_i u)} \right)^2 - \sum_{i=1}^{k} \frac{n_i^2 (3 + 4n_i u)}{((1 + n_i u)(1 + 2n_i u))^2} \right).$$

It follows that

$$f_k{}''(u) = -f_k(u) \mathbf{v}^{\mathrm{T}} \left(3\mathbf{I} - \mathbb{1}\mathbb{1}^{\mathrm{T}} \right) \mathbf{v} - 4 f_k(u) u \sum_{i=1}^{k} \frac{n_i^3}{((1 + n_i u)(1 + 2n_i u))^2}, \tag{88}$$

where $\mathbf{I}$ is the $k \times k$ identity matrix, $\mathbb{1} \in \mathbb{R}^k$ is the vector with all entries equal to 1 and $\mathbf{v} \in \mathbb{R}^k$ has entries

$$v_i := \frac{n_i}{(1 + n_i u)(1 + 2n_i u)}.$$

The $k \times k$ symmetric matrix $\mathbf{M} = 3\mathbf{I} - \mathbb{1}\mathbb{1}^{\mathrm{T}}$ has a $(k - 1)$ dimensional eigenspace associated to the eigenvalue 3 which is orthogonal to $\mathbb{1}$, and $\mathbb{1}$ is an eigenvector with eigenvalue $3 - k$. Therefore, $\mathbf{M}$ is positive semidefinite for $k \leq 3$, Equation (88) implies that

$$f_k{}''(u) \leq -4 f_k(u) u \sum_{i=1}^{k} \frac{n_i^3}{((1 + n_i u)(1 + 2n_i u))^2} < 0$$

and we are done. $\square$

5.2 Proofs of the remaining corollaries

In this section we prove the corollaries which were not proved in the previous sections.

Proof of Corollary 1 Corollary 1 is a consequence of the convexity of $(1+u)^{-k}$ and the concavity of f^k for $k \le 3$ and f in (15), which yield

$$1 - ku \le \frac{1}{(1+u)^k} \le \left(\frac{1+2u}{1+u}\right)^k \le 1 + ku$$

for $k = 1$, 2 and 3. □

Proof of Corollary 3 Let $\tilde{\mathrm{Fl}}$ be the rounding tuple in Prop. 15. If the y_k are floating point numbers then $S_k(\mathbf{y}, \mathrm{Fl}) = S_k(\mathbf{y}, \tilde{\mathrm{Fl}})$ for all k and Corollary 3 follows from Lemma 5. □

Proof of Corollary 4 Let $\tilde{\mathrm{Fl}}$ be the rounding tuple in Prop. 16. If all y_k are non negative floating point numbers then $S_k(\mathbf{y}, \mathrm{Fl}) = S_k(\mathbf{y}, \tilde{\mathrm{Fl}})$ for all k and Corollary 4 follows from Lemma 5. □

Proof of Corollary 6 If $\mathcal{F}$ is a MPFR system, let $\tilde{\mathrm{Fl}}$ be the rounding tuple in Prop. 16. Since all y_k belong to $\mathcal{M}$ and are non negative we have that $S_k(\mathbf{y}, \mathrm{Fl}) = S_k(\mathbf{y}, \tilde{\mathrm{Fl}})$ for all k and Corollary 6 follows from Lemma 7. If $\mathcal{F}$ is an IEEE system, let $\tilde{\mathrm{Fl}}$ be rounding tuple in Prop. 15. Since all y_k are floating point numbers, $S_k(\mathbf{y}, \mathrm{Fl}) = S_k(\mathbf{y}, \tilde{\mathrm{Fl}})$ for all k and Corollary 6 follows from Lemma 7. □

Proof of Corollary 8 In a perfect system, the dot product of $n+1$ numbers evaluated using a fma, as in Definition 20, is the floating point sum of the $(n+2)$ real numbers $p_0 := 0$ and $p_k := x_{k-1}y_{k-1}$ for $k > 0$, and Equation (27) follows from Lemma 5 applied to the p_k. □

Proof of Corollary 9 In an unperfect systems, the dot product of $n+1$ numbers evaluated using a fma, as in Definition 20, is the floating point sum of the $(n+2)$ real numbers $p_0 := 0$ and $p_k := x_{k-1}y_{k-1}$ for $k > 0$, and Corollary 9 follows from Corollary 5 applied to the p_k. □

Proof of Corollary 10 The dot product is the floating point sum of the floating point numbers $p_k := \mathrm{r}_k(x_k y_k)$. In a perfect system, Lemma 1 shows that

$$p_k = x_k y_k + \theta_k \frac{u}{1+u} x_k y_k \qquad \text{with} \qquad |\theta_k| \le 1,$$

and Lemma 5 implies that

$$\left| \mathrm{Fl}\left(\sum_{k=0}^{n} p_k\right) - \sum_{k=0}^{n} p_k \right| = \frac{nu}{1+nu} \sum_{k=0}^{n} |p_k|.$$

It follows that

$$\left| \mathrm{Fl}\left(\sum_{k=0}^{n} x_k y_k\right) - \sum_{k=0}^{n} x_k y_k \right| \le \sum_{k=0}^{n} |p_k - x_k| + \left| \mathrm{Fl}\left(\sum_{k=0}^{n} x_k y_k\right) - \sum_{k=0}^{n} p_k \right| \le \beta_n u \sum_{k=0}^{n} |x_k y_k|$$

for
$$\beta_n := \frac{1}{1+u}\left(1 + \frac{n}{1+nu}(1+2u)\right) = \frac{n+1+3nu}{1+(n+1)u+nu^2}.$$

Finally, note that for $n \geq 1$ and $20nu \leq 1$,

$$\beta_n - \frac{n+1}{1+nu/2} = -u\frac{(n-2)(n-1-nu)}{(1+nu/2)(1+(n+1)u+nu^2)} \leq 0,$$

and

$$\beta_n - \frac{n+1}{1+(n-3)u} = -u\frac{n+4+10nu-2n^2u}{(1+(n+1)u+nu^2)(1+(n-3)u)} < 0.$$

$\square$

Proof of Corollary 11   The dot product is the sum of the $n+1$ floating point numbers $p_k := r_k(x_k y_k)$, and Corollary 3 shows that

$$\left| \mathrm{Fl}\left(\sum_{k=0}^n p_k\right) - \sum_{k=0}^n p_k \right| \leq \frac{nu}{1+nu}\sum_{k=0}^n |p_k|.$$

We also have

$$|p_k - x_k y_k| \leq \frac{u}{1+u}|x_k y_k| + \frac{\alpha}{2}$$

and

$$\left| \mathrm{Fl}\left(\sum_{k=0}^n x_k y_k\right) - \sum_{k=0}^n x_k y_k \right| \leq \left| \mathrm{Fl}\left(\sum_{k=0}^n p_k\right) - \sum_{k=0}^n p_k \right| + \sum_{k=0}^n |p_k - x_k y_k|$$

$$\leq \frac{nu}{1+nu}\sum_{k=0}^n |p_k| + \sum_{k=0}^n |p_k - x_k y_k|$$

$$\leq \frac{nu}{1+nu}\left(\frac{1+2u}{1+u}\sum_{k=0}^n |x_k y_k| + (n+1)\frac{\alpha}{2}\right) + \frac{(n+1)\alpha}{2} + \frac{u}{1+u}\sum_{k=0}^n |x_k y_k|$$

$$= \beta_n u \sum_{k=0}^n |x_k y_k| + b\frac{\alpha}{2}$$

for β_n in Corollary 10 and

$$b := (n+1)\left(1 + \frac{nu}{1+nu}\right) = (n+1)\frac{1+2nu}{1+nu} < 1.05(n+1),$$

because $20nu \leq 1$. Finally, if $u\sum_{k=0}^n |x_k y_k| \geq \alpha$ then

$$\beta_n u \sum_{k=0}^n |x_k y_k| + b\frac{\alpha}{2} \leq \theta_n u \sum_{k=0}^n |x_k y_k| \qquad \text{for} \qquad \theta_n := \beta_n + \frac{n+1}{2}\frac{1+2nu}{1+nu},$$

and the software Mathematica shows that

$$\theta_n - 3\frac{n+1}{2} = -u\frac{n^2 - 3n + 2 + nu(1+n)}{2(1+u)(1+nu)}$$

which is negative for $n \geq 1$. This proves the last equation in Corollary 11. $\square$

Proof of Corollary 12   The dot product is the sum of the $n+1$ floating point numbers $p_k := \mathrm{r}_k(x_k y_k)$, and the proof of Corollary 5 shows that

$$\left| \mathrm{Fl}\left(\sum_{k=0}^{n} p_k \right) - \sum_{k=0}^{n} p_k \right| \leq \frac{m\alpha}{2} + \frac{(n-m)u}{1+(n-m)u} \left(\frac{m\alpha}{2} + \sum_{k=0}^{n} |p_k| \right),$$

for some $m \in [0, n]$. We also have that

$$|p_k - x_k y_k| \leq \frac{u}{1+u} |x_k y_k| + \frac{\alpha}{2}$$

and

$$\left| \mathrm{Fl}\left(\sum_{k=0}^{n} x_k y_k \right) - \sum_{k=0}^{n} x_k y_k \right| \leq \left| \mathrm{Fl}\left(\sum_{k=0}^{n} p_k \right) - \sum_{k=0}^{n} p_k \right| + \sum_{k=0}^{n} |p_k - x_k y_k|$$

$$\leq \frac{m\alpha}{2} + \frac{(n-m)u}{1+(n-m)u} \left(\frac{m\alpha}{2} + \sum_{k=0}^{n} |p_k| \right) + \sum_{k=0}^{n} |p_k - x_k y_k|$$

$$\leq \frac{(n-m)u}{1+(n-m)u} \left(\frac{1+2u}{1+u} \sum_{k=0}^{n} |x_k y_k| + (m+n+1)\frac{\alpha}{2} \right) + \qquad (89)$$

$$\frac{(m+n+1)\alpha}{2} + \frac{u}{1+u} \sum_{k=0}^{n} |x_k y_k| \leq \beta_n u \sum_{k=0}^{n} |x_k y_k| + b\frac{\alpha}{2}$$

for β_n in Corollary 10 and

$$b := \frac{n^2 + n - m^2 - m}{1+(n-m)u} u + (m+n+1) \leq \frac{n(n+1)u}{1+nu} + 2n + 1 \leq 2.05n + 1.05.$$

Finally, if $u \sum_{k=0}^{n} |x_k y_k| \geq \alpha$, then

$$\left| \mathrm{Fl}\left(\sum_{k=0}^{n} x_k y_k \right) - \sum_{k=0}^{n} x_k y_k \right| \leq \gamma_n u \sum_{k=0}^{n} |x_k y_k|$$

for

$$\gamma_n := \beta_n + \frac{1}{2} \left(\frac{n^2 + n - m^2 - m}{1+(n-m)u} u + (m+n+1) \right).$$

The derivative of γ_n with respect to m is

$$-\frac{1+2nu+u}{(1+(n-mu)u)^2} < 0$$

and γ_n is maximized for $m = 0$, in which case it is equal to the θ_n in the proof of Corollary 11. This proves the last statement in Corollary 12. $\qquad \square$

5.3   Numbers

This section contains new propositions about real and integer numbers, and the proofs of propositions related to these numbers stated in the main part of the article.

5.3.1 Propositions

This sections presents more propositions regarding real and integer numbers.

Proposition 21 (Continuity of the normal form) *If e is integer, $|z| = \beta^e (\beta^\mu + w)$ with $0 < w < (\beta - 1)\beta^\mu$ and*

$$|y - z| < \beta^e \min\{w, (\beta - 1)\beta^\mu - w\}$$

then $y = \mathrm{sign}(z)\beta^e (\beta^\mu + v)$ with $0 < v < (\beta - 1)\beta^\mu$ and $|v - w| = \beta^{-e}|y - z|$. ▲

Proposition 22 (Discontinuity of the normal form) *If e is integer and $|z| = \beta^{e+\mu}$ with $|y - z| < \beta^{e+\mu-1}(\beta - 1)$ then we have three possibilities:*

(i) *$|y| < |z|$ and $y = \mathrm{sign}(z)\beta^{e-1}(\beta^\mu + w)$ with*

$$0 < v = (\beta - 1)\beta^\mu - \beta^{1-e}|y - z| < (\beta - 1)\beta^\mu.$$

(ii) *$|y| = |z|$ and $y = z$.*

(iii) *$|y| > |z|$ and $y = \mathrm{sign}(z)\beta^e (\beta^\mu + w)$ with $0 < w = \beta^{-e}|y - z| < \beta^{\mu-1}(\beta - 1)$.*

▲

5.3.2 Proofs

In this section we prove the propositions regarding integer and real numbers.

Proof of Proposition 1 Since $d, e \in \mathbb{Z}$ and $d < e$ we have that $e - d \geq 1$ and

$$\beta^e (\beta^\mu + w) - \beta^d (\beta^\mu + v) \geq \beta^d \left(\left(\beta^{e-d} - 1 \right) \beta^\mu - v \right) \geq \beta^d ((\beta - 1)\beta^\mu - v) > 0,$$

and this shows that $\beta^e (\beta^\mu + w) > \beta^d (\beta^\mu + v)$. □

Proof of Proposition 2 The integer exponent $e := \lfloor \log_\beta (|z|) \rfloor - \mu$ satisfies

$$\log_\beta (|z|) - \mu - 1 < e \leq \log_\beta (|z|) - \mu \qquad \text{and} \qquad \beta^{-\mu-1}|z| < \beta^e \leq |z|\beta^{-\mu}.$$

The equation above shows that $w := \beta^{-e}|z| - \beta^\mu$ satisfies $0 \leq w < (\beta - 1)\beta^\mu$ and $z = \mathrm{sign}(z)\beta^e (\beta^\mu + w)$. If $z = \mathrm{sign}(z)\beta^d (\beta^\mu + v)$ with $d \in \mathbb{Z}$ and $0 \leq v < (\beta - 1)\beta^\mu$ then

$$\beta^e (\beta^\mu + w) = |z| = \beta^d (\beta^\mu + v),$$

and Prop. 1 implies that $d = e$, and the equation above implies that $v = w$. □

Proof of Proposition 21 We have that

$$\left| 1 - \frac{y}{z} \right| < \frac{\beta^e w}{|z|} = \frac{w}{\beta^\mu + w} < 1 \ \Rightarrow \frac{y}{z} > 0 \Rightarrow y \neq 0,$$

and Prop. 2 yield d and $v \in [0, (\beta - 1)\beta^\mu)$ such that $y = \mathrm{sign}(y)\beta^d (\beta^\mu + v)$. The inequality

$$\frac{\mathrm{sign}(y)\beta^d (\beta^\mu + v)}{\mathrm{sign}(z)\beta^e (\beta^\mu + w)} = \frac{y}{z} > 0$$

41

implies that $\mathrm{sign}(y) = \mathrm{sign}(z)$. Moreover,

$$\beta^d \left(\beta^\mu + v\right) = |y| \le |z| + |y - z| < |z| + \beta^e \left((\beta - 1)\beta^\mu - w\right) = \beta^{e+1+\mu}$$

and Prop. 1 implies that $d \le e$. Similarly,

$$\beta^d \left(\beta^\mu + v\right) = |y| \ge |z| - |y - z| > |z| - \beta^e w = \beta^{e+\mu},$$

and $d \ge e$. Therefore $d = e$, $y = \mathrm{sign}(y)\beta^e\left(\beta^\mu + v\right)$ and $|y - z| = \beta^e |w - z|$. $\qquad\square$

Proof of Proposition 22 We have that

$$\left|1 - \frac{y}{z}\right| < \frac{\beta - 1}{\beta} < 1 \;\; \Rightarrow \frac{y}{z} > 0 \Rightarrow y \ne 0,$$

and Prop 2 yields $d \in \mathbb{Z}$ and $w \in [0, (\beta - 1)\beta^\mu)$ such that $y = \mathrm{sign}(y)\beta^d\left(\beta^\mu + w\right)$. The inequality

$$\frac{\mathrm{sign}(y)\beta^d\left(\beta^\mu + w\right)}{\mathrm{sign}(z)\beta^{e+\mu}} = \frac{y}{z} > 0$$

implies that $\mathrm{sign}(y) = \mathrm{sign}(z)$. We also have that

$$\beta^d\left(\beta^\mu + w\right) = |y| \le |z| + |y - z| < |z| + \beta^{e+\mu-1}(\beta - 1) = \beta^e\left(\beta^\mu + \beta^{\mu-1}(\beta - 1)\right)$$

and Prop. 1 implies that $d \le e$. It $|y| \ge |z|$ then Prop. 1 implies that $d \ge e$. It follows that $d = e$ and the conditions in items (ii) and (iii) in Prop. 22 are satisfied. If $|y| < |z|$ then Prop. 1 implies that $d < e$ and

$$\beta^d\left(\beta^\mu + w\right) = |y| \ge |z| - \beta^{e+\mu-1}(\beta - 1) = \beta^{e-1}\left(\beta^{\mu+1} - \left(\beta^{\mu+1} - \beta^\mu\right)\right) = \beta^{e-1+\mu}$$

and Prop. 1 imply that $d \ge e - 1$. Therefore $d = e - 1$ and the conditions in item (i) in Prop. 22 are satisfied. $\qquad\square$

5.4 Floating point systems

In this section we present more definitions related to floating point systems and more propositions about them. We prove the propositions regarding floating point systems stated in the previous sections and the propositions stated here. In most definitions, propositions and proofs in this section $\mathcal{F}$ is a floating point system, fl rounds to nearest in $\mathcal{F}$, $z, w \in \mathbb{R}$ and $x, y \in \mathcal{F}$, and the numbers α and ν are as in Definitions 11 and 12, and the exceptions are stated explicitly.

5.4.1 Propositions

This section presents more propositions regarding floating point systems.

Proposition 23 (Minimality of alpha) $\alpha \in \mathcal{F}$ and if $x \in \mathcal{F} \setminus \{0\}$ then $|x| \ge \alpha$. ▲

Proposition 24 (Empty normal range) *If e is an exponent for $\mathcal{F}$ and r is an integer with $r \in [0, (\beta - 1)\beta^\mu)$ then $\mathcal{F} \cap \left(\beta^e\left(\beta^\mu + r\right), \beta^e\left(\beta^\mu + r + 1\right)\right) = \emptyset$.* ▲

Proposition 25 (Empty subnormal range) *Let $\mathcal{I}_{e_\alpha}$ be an IEEE system. If $r \in \mathbb{Z}$ and $-\beta^\mu \le r < \beta^\mu$ then $\left(\beta^{e_\alpha} r, \beta^{e_\alpha}(r+1)\right) \cap \mathcal{I}_{e_\alpha} = \emptyset$.* ▲

Proposition 26 (Scale invariance) *If $\mathcal{F}$ is perfect then $x \in \mathcal{F}$ if and only if $\beta x \in \mathcal{F}$. If $\mathcal{F}$ is unperfect and $x \in \mathcal{F}$ then $\beta x \in \mathcal{F}$.* ▲

5.4.2 Proofs

In this section we prove the propositions regarding floating point systems.

Proof of Proposition 3 According to Definitions 7 and 9 of MPFR system and IEEE system, we have three possibilities: (i) $x = 0$, in which case $x = \beta^{e_\alpha} r$ for $r = 0$, (ii) x is subnormal, and $\mathcal{F}$ is an IEEE system and $x = \beta^{e_\alpha} r$ with $|r| \in [1, \beta^\mu) \cap \mathbb{Z}$ and (iii) $x \in \mathcal{E}_e$ for some $e \geq e_\alpha$, and $x = \beta^{e_\alpha + e} r$ with $|r| \in [\beta^\mu, \beta^{1+\mu}) \cap \mathbb{Z}$. $\qquad\square$

Proof of Proposition 4 In the three possible cases, Definitions 6, 7 and 9, the floating point systems are clearly symmetric. $\qquad\square$

Proof of Proposition 5 If $\mathcal{F}$ is a perfect system or a MPFR system and $x \in \mathcal{F} \setminus \{0\}$ then $|x| \in \mathcal{E}_e$ for some exponent e for $\mathcal{F}$ by Definitions 6 and 7. If $\mathcal{F}$ is an IEEE system $\mathcal{I}_{e_\alpha}$ then $v = \beta^{e_\alpha + \mu}$ and x with $|x| \geq v$ is not subnormal. As a result, by definition of IEEE system, $|x| \in \mathcal{E}_e$ for some exponent e for $\mathcal{F}$. If $0 < |x| < v$ then $\mathcal{F}$ is not perfect, because $v = 0$ for perfect systems. Moreover, $|x| \notin \mathcal{E}_e$ for $e \geq e_\alpha$ and, by Definition 7, $\mathcal{F}$ is not a MPFR system. Therefore, $\mathcal{F}$ is an IEEE system and x is subnormal.

Regarding the converse part, if r is a multiple of β then we can replace e by $e + 1$ and r by r/β and z stays the same. Therefore, we can assume that r is not a multiple of β. In particular, $|r| < \beta^{1+\mu}$. By symmetry (Prop. 4), it suffices to show that $|z| \in \mathcal{F}$ when $|z| \geq v$. If $\mathcal{F}$ is perfect then $|z| \in \mathcal{E}_e$ and Prop. 5 holds. Therefore, we can assume that $\mathcal{F}$ is unperfect. In this case $v = \beta^{e_\alpha + \mu}$ by Definition 12 and $\beta^e |r| \geq \beta^{e_\alpha + \mu}$; actually $\beta^e |r| > \beta^{e_\alpha + \mu}$ because r is not a multiple of β. Since $0 < |r| < \beta^{1+\mu}$, there exists a first integer $d > 1$ such that $\beta^d |r| \geq \beta^{1+\mu}$. Dividing $\beta^{d-1} |r|$ by β^μ we obtain that $\beta^{d-1} |r| = \beta^\mu q + p$ for $p, q \in \mathbb{Z}$ with $q \geq 0$ and $0 \leq p < \beta^\mu$. The definition of d yields $\beta^{1+\mu} > \beta^{d-1} |r| = \beta^\mu q + p$ and

$$s := (q - 1)\beta^\mu + p < (\beta - 1)\beta^\mu.$$

Moreover,

$$\beta^{1+\mu} q + \beta p = \beta^d |r| \geq \beta^{1+\mu} \Rightarrow q \geq 1 - p/\beta^\mu > 0 \ \Rightarrow q \geq 1 \ \Rightarrow s \geq 0.$$

As a result, $|r| = \beta^{1-d} (\beta^\mu + s)$ with $s \in [0, (\beta - 1)\beta^\mu)$ and Prop. 1 leads to

$$|z| \geq v \Rightarrow \beta^{e+1-d} (\beta^\mu + s) \geq \beta^{e_\alpha} (\beta^\mu + 0)$$

$$\Rightarrow e + 1 - d \geq e_\alpha \Rightarrow |z| = \beta^{e+1-d} (\beta^\mu + s) \in \mathcal{E}_{e+1-d} \subset \mathcal{F}.$$

$\qquad\square$

Proof of Proposition 6 Prop. 4 states that $x + y \in \mathcal{I} \Leftrightarrow -(x + y) \in \mathcal{I}$, and it suffices to show that $|x + y| \in \mathcal{I}$. Since x and y are subnormal, $x = \text{sign}(x)\beta^{e_\alpha} r_x$ with $r_x \in [1, \beta^\mu) \cap \mathbb{Z}$ and $y = \text{sign}(y)\beta^{e_\alpha} r_y$ with $r_y \in [1, \beta^\mu) \cap \mathbb{Z}$. If $\text{sign}(x) = -\text{sign}(y)$ then $|x + y| = \beta^{e_\alpha} |r_x - r_y|$ and $|x + y|$ is either 0 or subnormal, because

$$|r_x - r_y| < \max\{r_x, r_y\} < \beta^\mu.$$

If $\text{sign}(x) = \text{sign}(y)$ then $|x + y| = \beta^{e_\alpha} (r_x + r_y)$ with $1 < r_x + r_y < 2\beta^\mu \leq \beta^{1+\mu}$. If $r_x + r_y < \beta^\mu$ then $x + y$ is subnormal, otherwise $|x + y| \geq \beta^{e_\alpha + \mu} = v$ and Prop. 5 implies that $|x + y| \in \mathcal{I}$. $\qquad\square$

Proof of Proposition 7   Let us start with $s := x + z > 0$. If $x \leq \beta^{e+\mu}$ then Prop. 7 holds because $z = s - x \geq \beta^e (r + 1/2) \geq \beta^e/2$. If $x \geq \beta^{e+\mu+1}$ then

$$z := s - x \leq \beta^e \left(r + 1/2 + \beta^\mu - \beta^{\mu+1} \right) = -\beta^e/2 - ((\beta - 1)\beta^\mu - (r+1)) \leq -\beta^e/2,$$

because $r \in [0, (\beta-1)\beta^\mu) \cap \mathbb{Z}$, and again $|z| \geq \beta^e/2$. Therefore, we only need to analyze the case $\beta^{e+\mu} < x < \beta^{e+\mu+1}$. In this case, by Prop. 2, $x = \beta^d (\beta^\mu + t)$ for $d \in \mathbb{Z}$ and $t \in [0, (\beta-1)\beta^\mu) \cap \mathbb{Z}$. Prop. 1 implies that $d = e$ and

$$z = s - x = \beta^e (r - t + 1/2) \;\Rightarrow\; |z| \geq \beta^e |r - t + 1/2| \geq \beta^e/2,$$

because $r - t \in \mathbb{Z}$, and we are done with the case $x + z > 0$. Finally, when $x + z < 0$ the argument above for $-x$ and $-z$ and leads to $|z| = |-z| \geq \beta^e/2$. □

Proof of Proposition 23   If $\mathcal{F}$ is perfect then $\alpha = 0 \in \mathcal{F}$ and Prop. 23 is trivial. If $\mathcal{F}$ is the MPFR system $\mathcal{M}_{e_\alpha, \beta, \mu}$ then $\alpha = \beta^{e_\alpha + \mu} \in \mathcal{E}_{e_\alpha} \subset \mathcal{F}$ and if $x \in \mathcal{F} \setminus \{0\}$ then $|x| \in \mathcal{E}_e$ for some $e \geq e_\alpha$. By definition of $\mathcal{E}_e$, $|x| = \beta^e (\beta^\mu + r)$ with $r \geq 0$ and $|x| \geq \alpha$. Finally, if $\mathcal{F}$ is the IEEE system $\mathcal{I}_{e_\alpha, \beta, \mu}$ then $\alpha = \beta^{e_\alpha} \in \mathcal{E}_{e_\alpha} \subset \mathcal{F}$ and if $x \in \mathcal{F} \setminus \{0\}$ then $|x| \in \mathcal{F} \setminus \{0\}$ and either (i) $|x| \in \mathcal{S}_{e_\alpha}$ or (ii) $|x| \in \mathcal{E}_e$ with $e \geq e_\alpha$. In case (i), $|x| = \beta^{e_\alpha} r$ for $r \in \mathbb{Z} \setminus 0$ and $|x| \geq \alpha$. As for the MPFR system, in case (ii) $|x| \geq \alpha$. □

Proof of Proposition 24   We show that if $z \in (\beta^e (\beta^\mu + r), \; \beta^e (\beta^\mu + r + 1))$ then $z \notin \mathcal{F}$. By Prop. 1 and 2, there exists w with $r < w < r + 1$ such that $z = \beta^e (\beta^\mu + w)$. $z \notin -\mathcal{E}_d$ because $z > 0$. Prop. 1 implies that if $d > e$ then $z < y$ for $y \in \mathcal{E}_d$ and if $d < e$ then $z > y$ for $y \in \mathcal{E}_d$. Therefore, $z \notin \bigcup_{d \neq e} \mathcal{E}_d$. Moreover, if $y \in \mathcal{E}_e$ then $y = \beta^e (\beta^\mu + s)$ with $s \in \mathbb{Z}$ and $y \neq z$ because $w \notin \mathbb{Z}$. As a result, $z \notin \bigcup_{d \in \mathbb{Z}} (\mathcal{E}_d \cup -\mathcal{E}_d)$. This proves that $z \notin \mathcal{F}$ when $\mathcal{F}$ is a perfect or MPFR system. Finally, if $\mathcal{F}$ is an IEEE system $\mathcal{I}_{e_\alpha}$ then $e \geq e_\alpha$ because e is an exponent for $\mathcal{F}$, and $z > y$ for all $y \in \mathcal{S}_{e_\alpha} \cup -\mathcal{S}_{e_\alpha}$. This shows that $z \notin \mathcal{S}_{e_\alpha} \cup -\mathcal{S}_{e_\alpha}$ and $z \notin \mathcal{F}$. □

Proof of Proposition 25   We show that if $z \in (\beta^{e_\alpha} r, \; \beta^{e_\alpha} (r + 1))$ then $z \notin \mathcal{I}$. We have that $|z| < \beta^{e_\alpha} \max \{|r|, |r+1|\} \leq \beta^{e_\alpha + \mu}$ and $|z| \notin \bigcup_{e=e_\alpha}^{\infty} \mathcal{E}_e$. Moreover, $w := \beta^{-e_\alpha} z$ is such that $r < w < r + 1$ and $z = \beta^{e_\alpha} w$. It follows that $w \notin \mathbb{Z}$ and $|z| \notin \mathcal{S}_{e_\alpha}$, and combining the arguments above and symmetry (Prop. 4) we conclude that $z \notin \mathcal{I}$. □

Proof of Proposition 26   Since $\mathcal{A}_e := \{\beta x, x \in \mathcal{E}_e\} = \mathcal{E}_{e+1}$, the set $\mathcal{P}$ in Definition 6 is such that $x \in \mathcal{P}$ if and only if $\beta x \in \mathcal{P}$. For the MPFR system $\mathcal{M}_{e_\alpha, \beta, \mu}$, if $x \in \mathcal{M}$ then $|x| \in \mathcal{E}_e$ for some $e \geq e_\alpha$, $|\beta x| \in \mathcal{E}_{e+1} \subset \mathcal{M}$ and $\beta x \in \mathcal{M}$ by symmetry (Prop. 4.) For the IEEE system $\mathcal{I}_{e_\alpha, \beta, \mu}$, if $x \in \mathcal{I}$ then either $x \in \mathcal{E}_e$ for some $e \geq e_\alpha$, and the argument used in the MFPR case applies to x, or $x = \text{sign}(x) \beta^{e_\alpha} r$ with $r \in [0, \beta^\mu) \cap \mathbb{Z}$. If $\beta r < \mu$ then $|\beta x| = \beta^{e_\alpha} (\beta r) \in \mathcal{S}_{e_\alpha}$ and $\beta x \in \mathcal{I}$ by symmetry. If $\beta r \geq \beta^\mu$ then $s = \beta r - \beta^\mu \in [0, (\beta-1)\beta^\mu) \cap \mathbb{Z}$ and $|\beta x| = \beta^{e_\alpha} (\beta^\mu + s) \in \mathcal{E}_{e_\alpha} \subset \mathcal{I}$ and $s \in \mathcal{I}$ by symmetry. □

5.5   Rounding

This section proves the propositions about rounding to nearest stated previously, and states and proves more propositions about rounding.

5.5.1 Propositions

In this section we state more propositions regarding rounding to nearest.

Proposition 27 (Propagation of the sign) *If* $\mathrm{fl}(z) \neq 0$ *then* $\mathrm{sign}(\mathrm{fl}(z)) = \mathrm{sign}(z)$. *For a general* $z \in \mathbb{R}$, $\mathrm{fl}(z) = \mathrm{sign}(z)\,|\mathrm{fl}(z)|$. ▲

Proposition 28 (Rounding after scaling) *Let m be an integer. If $\mathcal{F}$ is perfect then the function* $\mathrm{s}(z) := \beta^{-m}\mathrm{fl}(\beta^m z)$ *rounds to nearest in* $\mathcal{F}$. ▲

Proposition 29 (Rounding in an interval) *If* $a, b \in \mathcal{F}$ *and* $a \leq z \leq b$ *then* $\mathrm{fl}(z) \in [a, b]$ *and* $|\mathrm{fl}(z) - z| \leq (b-a)/2$. *Moreover, if* $z < m := (a+b)/2$ *then* $\mathrm{fl}(z) < b$ *and if* $z > m$ *then* $\mathrm{fl}(z) > a$. ▲

Proposition 30 (Combination) *For* $\mathcal{A}_1, \mathcal{A}_2 \subset \mathbb{R}$ *with* $\mathcal{A}_1 \cup \mathcal{A}_2 = \mathbb{R}$, *let* $f_i : \mathcal{A}_i \to \mathbb{R}$ *be such that, for* $z_i \in \mathcal{A}_i$ *and* $x \in \mathcal{F}$, $f_i(z_i) \in \mathcal{F}$ *and* $|z_i - f_i(z_i)| \leq |z_i - x|$. *The function* $\mathrm{fl} : \mathbb{R} \to \mathbb{R}$ *given by* $\mathrm{fl}(z) = f_1(z)$ *for* $z \in \mathcal{A}_1$ *and* $\mathrm{fl}(z) = f_2(z)$ *for* $z \in \mathcal{A}_2 \setminus \mathcal{A}_1$ *rounds to nearest in* $\mathcal{F}$. ▲

Proposition 31 (Extension) *If* $\mathcal{A} \subset \mathbb{R}$ *and* $f : \mathcal{A} \to \mathbb{R}$ *is such that, for* $z \in \mathcal{A}$ *and* $x \in \mathcal{F}$, $f(z) \in \mathcal{F}$ *and* $|z - f(z)| \leq |z - x|$ *then there exists a function* fl *which rounds to nearest in* $\mathcal{F}$ *and is such that* $\mathrm{fl}(z) = f(z)$ *for* $z \in \mathcal{A}$. ▲

5.5.2 Proofs

In this section we prove the propositions regarding rounding to nearest.

Proof of Proposition 8   By definition of rounding to nearest, $0 = |x - x| \geq |\mathrm{fl}(x) - x|$. Therefore, $\mathrm{fl}(x) = x$. □

Proof of Proposition 9   Let us show that if $\mathrm{fl}(z) > \mathrm{fl}(w)$ then $z > w$. Indeed, in this case we have that

$$|\mathrm{fl}(w) - z| \geq |\mathrm{fl}(z) - z| \geq \mathrm{fl}(z) - z > \mathrm{fl}(w) - z.$$

Therefore, $|\mathrm{fl}(w) - z| > \mathrm{fl}(w) - z$ and this implies that $z > \mathrm{fl}(w)$. It follows that

$$z - \mathrm{fl}(w) = |\mathrm{fl}(w) - z| \geq |\mathrm{fl}(z) - z| \geq \mathrm{fl}(z) - z \;\Rightarrow\; z \geq \frac{\mathrm{fl}(z) + \mathrm{fl}(w)}{2}.$$

Similarly,

$$|w - \mathrm{fl}(z)| \geq |w - \mathrm{fl}(w)| \geq w - \mathrm{fl}(w) > w - \mathrm{fl}(z) \;\Rightarrow\; w \leq \mathrm{fl}(z),$$

and

$$\mathrm{fl}(z) - w = |\mathrm{fl}(z) - w| \geq |\mathrm{fl}(w) - w| \geq w - \mathrm{fl}(w) \;\Rightarrow\; w \leq \frac{\mathrm{fl}(z) + \mathrm{fl}(w)}{2}.$$

As a result, $w \leq (\mathrm{fl}(z) + \mathrm{fl}(w))/2 \leq z$. Moreover, $w \neq z$ because $\mathrm{fl}(z) \neq \mathrm{fl}(w)$. Therefore, $z > w$ as we have claimed. Logically, we have proved that $z \leq w \Rightarrow \mathrm{fl}(z) \leq \mathrm{fl}(w)$.

When $x \in \mathcal{F}$ we have that $\mathrm{fl}(x) = x$ (Prop. 8) and the argument above shows that $\mathrm{fl}(z) > x \Rightarrow z > x$ and $x > \mathrm{fl}(z) \Rightarrow x > z$. Moreover,

$$|x| > |\mathrm{fl}(z)| \Rightarrow |x| > \mathrm{fl}(z) \Rightarrow |x| > z$$

and using the function m in Prop. 10 we obtain

$$|x| > |\mathrm{fl}(z)| \Rightarrow |x| > -\mathrm{fl}(z) \Rightarrow |x| > \mathrm{m}(-z) \Rightarrow |x| > -z.$$

Therefore, $|x| > |\mathrm{fl}(z)| \Rightarrow |x| > \max\{z, -z\} = |z|$.

Finally, if $|x| < |\mathrm{fl}(z)|$ then either (i) $\mathrm{fl}(z) < 0$ or (ii) $\mathrm{fl}(z) > 0$. In both cases Prop. 27 shows that $\mathrm{sign}(z) = \mathrm{sign}(\mathrm{fl}(z))$. In case (i) z is positive and

$$|x| < |\mathrm{fl}(z)| \Rightarrow |x| < \mathrm{fl}(z) \Rightarrow |x| < z = |z|.$$

and in case (ii) z is negative and

$$|x| < |\mathrm{fl}(z)| \Rightarrow |x| < -\mathrm{fl}(z) \Rightarrow |x| < \mathrm{m}(-z) \Rightarrow |x| < -z = |z|.$$

Therefore, $|x| < |\mathrm{fl}(z)| \Rightarrow |x| < |z|$ in both cases and we are done. □

Proof of Proposition 10   If $x \in \mathcal{F}$ and $z \in \mathbb{R}$ then $-x \in \mathcal{F}$ by symmetry and

$$|\mathrm{m}(z) - z| = |(-\mathrm{fl}(-z)) - z| = |\mathrm{fl}(-z) - (-z)| \le |(-x) - (-z)| = |x - z|.$$

Therefore, $|\mathrm{m}(z) - z| \le |x - z|$ and m rounds to nearest. □

Proof of Proposition 11   Let us start with $z > 0$. $w \le (\beta - 1)\beta^\mu$ and if $\lfloor w \rfloor = (\beta - 1)\beta^\mu$ then $a = b = \beta^{e+1+\mu} \in \mathcal{E}_{e+1}$, and this implies that $a, b \in \mathcal{F}$ because $e + 1$ is also an exponent for $\mathcal{F}$. Similarly, if $\lceil w \rceil = (\beta - 1)\beta^\mu$ then $b \in \mathcal{E}_{e+1} \subset \mathcal{F}$. If $\lceil w \rceil < (\beta - 1)\beta^\mu$ then $0 \le \lfloor w \rfloor \le \lceil w \rceil < (\beta - 1)\beta^\mu$ and $a, b \in \mathcal{E}_e \subset \mathcal{F}$. Therefore, in all cases, $a, b \in \mathcal{F}$. If $w \in \mathbb{Z}$ then $\lfloor w \rfloor = \lceil w \rceil$ and $z = a = b \in \mathcal{F}$ and $\mathrm{fl}(z) = a = b = m$ because $\mathrm{fl}(x) = x$ when $x \in \mathcal{F}$ by Prop. 8. If $w \notin \mathbb{Z}$ then $\lceil w \rceil = \lfloor w \rfloor + 1$, Prop. 24 shows that $(a, b) \cap \mathcal{F} = \emptyset$, Equation (34) follows from Prop. 29, and we also have that $(b - a)/2 \le \beta^e/2$.

For the last paragraph in Prop. 11, we either have (i) $r \le w$ or (ii) $r > w$. In case (i)

$$r \le w \le r + |w - r| < r + 1/2 \Rightarrow \lfloor w \rfloor = r, \ \lceil w \rceil = r + 1$$

and

$$\frac{1}{2}(\lfloor w \rfloor + \lceil w \rceil) = r + 1/2 > w.$$

This implies that $a = \beta^e(\beta^\mu + r)$, $b = \beta^e(\beta^\mu + r + 1)$ and $z < (a + b)/2$, and the results in the previous paragraph show that $\beta^e(\beta^\mu + r) = a = \mathrm{fl}(z)$.

In case (ii), $r > w \ge 0 \Rightarrow r \ge 1$ and

$$r - 1/2 \le w < r \Rightarrow \lfloor w \rfloor = r - 1, \ \lceil w \rceil = r \ \text{and} \ \frac{1}{2}(\lfloor w \rfloor + \lceil w \rceil) = r - 1/2 < w.$$

This implies that $a = \beta^e(\beta^\mu + r - 1)$, $b = \beta^e(\beta^u + r)$ and $z > (a + b)/2$, and the results in the first paragraph of this proof show that $\beta^e(\beta^u + r) = b = \mathrm{fl}(z)$.

Finally, for $z < 0$ the arguments above for $\tilde{z} = -z$ and $\tilde{\mathrm{fl}}$ equal to the function m in Prop. 10 and symmetry (Prop. 4) prove Prop. 11 for z. □

Proof of Proposition 12   Recall that $v = \beta^{e_\alpha + \mu} \in \mathcal{E}_{e_\alpha} \subset \mathcal{I}$, and by symmetry $-v \in \mathcal{I}$. Let us write $w := \beta^{-e_\alpha} z$ and $r := \lfloor w \rfloor$. We have that $a = \beta^{e_\alpha} r$ and if $r = w$ then $a = b = z$ and $\mathrm{fl}(z) = z$ by Prop. 8 and Prop. 12 is valid. Let us then assume that

46

$r \neq w$. This implies that $w \notin \mathbb{Z}$, $r < \beta^\mu$, $r+1 = \lceil w \rceil$ and $b = \beta^{e_\alpha}(r+1)$. We have that $a \in \mathcal{F}$ because

$$
\begin{aligned}
w < 1 - \beta^\mu &\Rightarrow r = -\beta^\mu \Rightarrow a = -\nu \in -\mathcal{E}_{e_\alpha} \subset \mathcal{F}, \\
1 - \beta^\mu < w < 0 &\Rightarrow 1 - \beta^\mu \leq r < 0 \Rightarrow a \in -\mathcal{S}_{e_\alpha} \subset \mathcal{F}, \\
0 < w < 1 &\Rightarrow r = 0 \Rightarrow a = 0 \in \mathcal{F}, \\
1 < w < \beta^\mu &\Rightarrow 1 \leq r < \beta^\mu \Rightarrow a \in \mathcal{S}_{e_\alpha} \subset \mathcal{F},
\end{aligned}
$$

and $b \in \mathcal{F}$ because

$$
\begin{aligned}
-\beta^\mu < w < -1 &\Rightarrow 1 - \beta^\mu < r+1 \leq -1 \Rightarrow b \in -\mathcal{S}_{e_\alpha} \subset \mathcal{F}, \\
-1 < w < 0 &\Rightarrow r+1 = 0 \Rightarrow b = 0 \in \mathcal{F}, \\
0 < w < \beta^\mu - 1 &\Rightarrow 1 \leq r+1 < \beta^\mu \Rightarrow b \in \mathcal{S}_{e_\alpha} \subset \mathcal{F}, \\
\beta^\mu - 1 < w < \beta^\mu &\Rightarrow r+1 = \beta^\mu \Rightarrow b = \nu \in \mathcal{E}_{e_\alpha} \in \mathcal{F}.
\end{aligned}
$$

Therefore, by monotonicity $\mathrm{fl}(z) \in [a,b] \cap \mathcal{F}$ and Prop. 25 implies that $\mathrm{fl}(z) \in \{a,b\}$. It follows that

$$
|\mathrm{fl}(z) - z| = \min\{z - a, b - z\} \leq \frac{b-a}{2} = \frac{\beta^{e_\alpha}(r+1) - \beta^{e_\alpha}(r)}{2} = \alpha/2.
$$

Finally, if $z < m$ then $|b - z| > |a - z| \Rightarrow \mathrm{fl}(z) = a$ and if $z > m$ then $|a - z| > |b - z| \Rightarrow \mathrm{fl}(z) = b$. $\qquad\square$

Proof of Proposition 13   Note that, by Prop. 23, if $x \in \mathcal{F} \setminus \{0, \pm\alpha\}$ then $|x| > \alpha$. When $|z| < \alpha/2$, if $x \in \mathcal{F} \setminus \{0\}$ then Prop. 23 implies that $|x| \geq \alpha$ and

$$
|x - z| \geq |x| - |z| \geq \alpha - |z| > \alpha/2 > |z - 0|,
$$

and $\mathrm{fl}(z) = 0$ because $0 \in \mathcal{F}$.

When $|z| = \alpha/2$, $|z - 0| = |z - \mathrm{sign}(z)\alpha| = \alpha/2$ and $|z - (-\mathrm{sign}(z))| = 3\alpha/2$. As a result, if $x \in \mathcal{F} \setminus \{0, \pm\alpha\}$ then

$$
|x - z| \geq |x| - |z| > \alpha - \alpha/2 = \alpha/2 = |z - 0|,
$$

and the bounds above imply that $\mathrm{fl}(z) \in \{0, \mathrm{sign}(z)\alpha\}$.

When $\alpha/2 < |z| < \alpha$, $|z - \mathrm{sign}(z)\alpha| = \alpha - |z| < \alpha/2$, $|z - 0| = |z| > \alpha/2$ and

$$
|z - (-\mathrm{sign}(z))\alpha| = |z| + \alpha > \alpha/2.
$$

Moreover, if $x \in \mathcal{F} \setminus \{0, \pm\alpha\}$ has the same sign as z then $x > \alpha$ and

$$
|x - z| = x - z = (x - \alpha) + (\alpha - z) > |\mathrm{sign}(z)\alpha - z|.
$$

and if x has the opposite sign of z then $|x - z| \geq |x| > \alpha > |\mathrm{sign}(z)\alpha - z|$, and the bounds in this paragraph imply that $\mathrm{fl}(z) = \mathrm{sign}(z)\alpha$. Finally, if $|z| = \alpha$ then $\mathrm{fl}(z) = z = \mathrm{sign}(z)\alpha$ by Prop. 8. $\qquad\square$

Proof of Proposition 14   Let $\mathcal{A}$ be the set $\{z \in \mathbb{R} \text{ with } |z| \geq \nu_\mathcal{F}\}$ and $f : \mathcal{A} \to \mathbb{R}$ the function $f(z) = \mathrm{fl}(z)$. We claim that if $x \in \mathcal{P}$ and $z \in \mathcal{A}$ then $|z - f(z)| \leq |x - f(z)|$.

In fact, if $x \in \mathcal{F}$ then $|x - z| \geq |\mathrm{fl}(z) - z| = |f(z) - z|$, because fl rounds to nearest in $\mathcal{F}$. If $x \notin \mathcal{F}$ then

$$x \in \mathcal{P} \setminus \mathcal{F} \subset \left(\bigcup_{e=-\infty}^{+\infty} \left(\mathcal{E}_e \bigcup -\mathcal{E}_e \right) \setminus \bigcup_{e=e_\alpha}^{+\infty} \left(\mathcal{E}_e \bigcup -\mathcal{E}_e \right) \right) = \bigcup_{e=-\infty}^{e_\alpha - 1} \left(\mathcal{E}_e \bigcup -\mathcal{E}_e \right)$$

and

$$|x| < \beta^{e_\alpha - 1} \left(\beta^\mu + (\beta - 1) \beta^\mu \right) = \beta^{e_\alpha + \mu - 1} = \nu_\mathcal{F}.$$

since $\nu_\mathcal{F} \in \mathcal{F}$, if $z \geq \nu_\mathcal{F}$ then $z \geq |x| \geq x$ and

$$|x - z| = z - x = |\nu_\mathcal{F} - z| + |\nu_\mathcal{F} - x| \geq |\mathrm{fl}(z) - z| + |\nu_\mathcal{F} - x| > |f(z) - z|.$$

Similarly, $\nu_\mathcal{F} \in \mathcal{F}$ and if $z \leq -\nu_\mathcal{F}$ then $z \leq -|x| \leq x$ and

$$|x - z| = x - z = |-\nu_\mathcal{F} - z| + |-\nu_\mathcal{F} - x| \geq |\mathrm{fl}(z) - z| + |-\nu_\mathcal{F} - x| > |f(z) - z|,$$

Therefore, $|x - z| \geq |f(z) - z|$ in all cases. To complete the proof it suffices to take the extension of f to $\mathbb{R}$ given by Prop. 31. $\qquad\square$

Proof of Proposition 15 For $k = 1, \ldots, n$ let $\tilde{\mathrm{fl}}_k$ be the adapter of fl_k in Prop. 14. On the one hand, by the definition of $\tilde{\mathrm{fl}}_k$, we have that if $x, y \in \mathcal{F}$ and $|x + y| \geq \nu_\mathcal{I}$ then

$$\mathrm{fl}_k(x + y) = \tilde{\mathrm{fl}}_k(x + y). \tag{90}$$

On the other hand, Lemma 3 shows that Equation (90) holds when $|x + y| \leq \nu_\mathcal{I}$. Therefore, Equation (90) holds for all $x, y \in \mathcal{I}$. For $\mathbf{x} \in \mathcal{I}^{n+1}$ define $\mathbf{z} \in \mathbb{R}^n$ as $z_1 := x_0 + x_1$ and $z_k := x_k$ for $2 \leq k < n$. We now prove by induction that $S_k(\mathbf{z}, \mathrm{Fl}) = S_k(\mathbf{z}, \tilde{\mathrm{Fl}})$. By definition, $S_0(\mathbf{z}, \mathrm{Fl}) = 0 = S_0(\mathbf{z}, \tilde{\mathrm{Fl}})$. Let us then analyze $k > 0$ assuming that $S_{k-1}(\mathbf{z}, \tilde{\mathrm{Fl}}) = S_{k-1}(\mathbf{z}, \mathrm{Fl}) \in \mathcal{I}$. Using Equation (90) we deduce that

$$S_k(\mathbf{z}, \tilde{\mathrm{Fl}}) = \tilde{\mathrm{fl}}_k(S_{k-1}(\mathbf{z}, \mathrm{Fl}) + z_k) = \mathrm{fl}_k(S_{k-1}(\mathbf{x}, \mathrm{Fl}) + z_k) = S_k(\mathbf{z}, \mathrm{Fl}) \in \mathcal{I}$$

and we are done. $\qquad\square$

Proof of Proposition 16 For $k = 1, \ldots, n$, let $\tilde{\mathrm{fl}}_k$ be the adapter of fl_k in Prop. 14. By the definition of $\tilde{\mathrm{fl}}_k$ we have that if $x, y \in \mathcal{F}$ and $|x + y| \geq \alpha_\mathcal{I} = \nu_\mathcal{I}$ then

$$\mathrm{fl}_k(x + y) = \tilde{\mathrm{fl}}_k(x + y), \tag{91}$$

and, of course, this equation is also satisfied when $x + y = 0$. For $\mathbf{x} \in \mathcal{I}^{n+1}$ define $\mathbf{z} \in \mathbb{R}^n$ as $z_1 := x_0 + x_1$ and $z_k := x_k$ for $2 \leq k < n$. We now prove by induction that if $y_k := S_{k-1}(\mathbf{z}, \mathrm{Fl}) + z_k \geq 0$ for $k = 0, \ldots, n$ then $S_k(\mathbf{z}, \mathrm{Fl}) = S_k(\mathbf{z}, \tilde{\mathrm{Fl}})$. By definition, $S_0(\mathbf{z}, \mathrm{Fl}) = 0 = S_0(\mathbf{z}, \tilde{\mathrm{Fl}})$. Let us then analyze $k > 0$ assuming that $S_{k-1}(\mathbf{z}, \tilde{\mathrm{Fl}}) = S_{k-1}(\mathbf{z}, \mathrm{Fl}) \in \mathcal{M}$. The assumption that $y_k \geq 0$ and Prop. 23 implies that either $y_k = 0$ or $y_k \geq \alpha_\mathcal{M} = \nu_\mathcal{M}$, and in both cases Equation (91) holds for $x + y = y_k$. It follows that

$$S_k(\mathbf{z}, \tilde{\mathrm{Fl}}) = \tilde{\mathrm{fl}}_k(S_{k-1}(\mathbf{z}, \mathrm{Fl}) + z_k) = \tilde{\mathrm{fl}}_k(y_k) = \mathrm{fl}_k(y_k) = S_k(\mathbf{z}, \mathrm{Fl}) \in \mathcal{M},$$

and we are done. $\qquad\square$

Proof of Proposition 17 If $w = 0$ then we can take $\delta = \beta^{e-1}/2$, because in this case $z \in \mathcal{E}_e \subset \mathcal{F}$ and $\mathrm{fl}_1(z) = z$ by Prop. 8 and, according to Prop. 22, if $|y - z| < \delta$ then either

(i) $y = \text{sign}(z)\beta^e(\beta^\mu + v)$ with

$$0 \le v = \beta^{-e}|y-z| < \beta^{-e}\delta < 1/2 \Rightarrow \lfloor v \rfloor = 0$$

and $\text{fl}_2(y) = \text{fl}_1(z) = z$ by Prop. 11, or

(ii) $y = \text{sign}(z)\beta^{e-1}(\beta^\mu + v)$ for

$$(\beta - 1)\beta^\mu - \beta^{1-e}|y - z| = v < (\beta - 1)\beta^\mu \Rightarrow$$

$$(\beta - 1)\beta^\mu - 1/2 < v < (\beta - 1)\beta^\mu \Rightarrow \lceil v \rceil = (\beta - 1)\beta^\mu$$

and, by Prop. 11,

$$\text{fl}_2(y) = \text{sign}(z)\beta^{e-1}(\beta^\mu + (\beta - 1)\beta^\mu) = \text{sign}(z)\beta^{e+\mu} = z = \text{fl}_1(z)$$

Let us then assume that $w > 0$ and write $m := \lfloor w \rfloor + 1/2$ and show that

$$\delta = \beta^e \min\{w, (\beta - 1)\beta^\mu - w, 1/2 - |m - w|, |m - w|\}$$

is a valid choice. Note that $\delta > 0$, because $|m - w| \le 1/2$ for a general w and $w \ne 1/2$ for the particular w we discuss here. If $|y - z| < \delta$ then Prop. 21 implies that $y = \text{sign}(z)\beta^e(\beta^\mu + v)$ with

$$|v - w| = \beta^{-e}|y - z| < \beta^{-e}\delta \le \min\{1/2 - |m - w|, |m - w|\}.$$

On the one hand, if $w < m$ then $|m - w| = m - w$,

$$\lfloor w \rfloor = m - 1/2 < m - (|w - m| + |v - w|) \le v \le w + |w - v| < w + |m - w| = m,$$

$\lfloor v \rfloor = \lfloor w \rfloor$ and Prop. 11 implies that $\text{fl}_2(y) = \text{fl}_1(z) = \text{sign}(z)\beta^e(\beta^\mu + \lfloor w \rfloor)$. On the other hand, if $w > m$ then $|m - w| = w - m$,

$$m = w - |w - m| < w - |w - v| \le v \le m + (|w - m| + |v - w|) < m + 1/2 = \lceil w \rceil,$$

$\lceil v \rceil = \lceil w \rceil$ and Prop. 11 implies that $\text{fl}_2(y) = \text{fl}_1(z) = \text{sign}(z)\beta^e(\beta^\mu + \lceil w \rceil)$. $\qquad \square$

Proof of Proposition 18 For $k = 1, \dots, n$ Props. 10 and 28 show that the function $\tilde{\text{fl}}_k(z) := \sigma\beta^{-m}\text{fl}_k(\sigma\beta^m z)$ rounds to nearest in $\mathcal{P}$, and we define $\tilde{\text{Fl}} := \{\tilde{\text{fl}}_1, \dots, \tilde{\text{fl}}_n\}$. We now prove by induction in $k = 0, \dots, n$ that

$$S_k(\sigma\beta^m \mathbf{z}, \text{Fl}) = \sigma\beta^m S_k(\mathbf{z}, \tilde{\text{Fl}}), \tag{92}$$

For $k = 0$, $S_0(\sigma\beta^m \mathbf{z}, \text{Fl}) = 0 = \sigma\beta^m S_k(\mathbf{z}, \tilde{\text{Fl}})$ by definition. Assuming that (92) holds for $k \ge 0$ we have that

$$S_{k+1}(\sigma\beta^m \mathbf{z}, \text{Fl}) = \text{fl}_{k+1}(S_k(\sigma\beta^m \mathbf{z}, \text{Fl}) + \sigma\beta^m z_k)$$

$$= \text{fl}_{k+1}(\sigma\beta^m(S_k(\mathbf{z}, \tilde{\text{Fl}}) + z_k)) = \sigma\beta^m \tilde{\text{fl}}_{k+1}(S_k(\mathbf{z}, \tilde{\text{Fl}}) + z_k) = S_{k+1}(\mathbf{z}, \tilde{\text{Fl}}),$$

and we are done. $\qquad \square$

Proof of Proposition 27 Prop. 8 shows that $\text{fl}(0) = 0$. Therefore, if $\text{fl}(z) \ne 0$ then either (i) $z > 0$ or (ii) $z < 0$. In case (i)

$$|\text{fl}(z) - z| \le |0 - z| \Rightarrow z - \text{fl}(z) \le z \Rightarrow \text{fl}(z) \ge 0 \Rightarrow \text{sign}(\text{fl}(z)) = 1 = \text{sign}(z).$$

In case (ii)
$$|\mathrm{fl}(z) - z| \leq |0 - z| \Rightarrow \mathrm{fl}(z) - z \leq -z \Rightarrow \mathrm{fl}(z) \leq 0.$$

Since $\mathrm{fl}(z) \neq 0$ this implies that $\mathrm{sign}(\mathrm{fl}(z)) = -1 = \mathrm{sign}(z)$. It follows that if $\mathrm{fl}(z) \neq 0$ then $\mathrm{fl}(z) = \mathrm{sign}(\mathrm{fl}(z)) |\mathrm{fl}(z)| = \mathrm{sign}(z) |\mathrm{fl}(z)|$ and it is clear that this equality also holds when $\mathrm{fl}(z) = 0$. $\qquad\square$

Proof of Proposition 28 Suppose $x \in \mathcal{F}$ and $z \in \mathbb{R}$. When $\mathcal{F}$ is perfect we have that $\beta^m x \in \mathcal{F}$ by Prop. 26 and since fl rounds to nearest we have

$$|\mathrm{s}(z) - z| = \left| \left(\beta^{-m} \mathrm{fl}(\beta^m z)\right) - z \right| = \beta^{-m} |\mathrm{fl}(\beta^m z) - (\beta^m z)|$$

$$\leq \beta^{-m} |\mathrm{fl}(\beta^m z) - (\beta^m x)| = \left| \left(\beta^{-m} \mathrm{fl}(\beta^m z)\right) - x \right| = |\mathrm{s}(z) - x|.$$

Therefore, s rounds to nearest in $\mathcal{F}$. $\qquad\square$

Proof of Proposition 29 Since $x = a, b \in \mathcal{F}$, the definition of rounding to nearest yields $|z - a| \geq |z - \mathrm{fl}(z)|$ and $|z - b| \geq |z - \mathrm{fl}(z)|$. If $y < a$ then $y < z$ and

$$|z - y| = z - y > z - a = |z - a| \geq |z - \mathrm{fl}(z)| \;\Rightarrow\; |z - y| > |z - \mathrm{fl}(z)|$$

Therefore, $\mathrm{fl}(z) \neq y$. Similarly, if $y > b$ then $y > z$ and

$$|z - y| = y - z > b - z = |z - b| \geq |z - \mathrm{fl}(z)| \;\Rightarrow\; |z - y| > |z - \mathrm{fl}(z)|$$

As a result, $\mathrm{fl}(z) \neq y$ and $\mathrm{fl}(z) \in \mathbb{R} \setminus (\{y < a\} \cup \{y > b\}) = [a, b]$. If $z \leq m$ then

$$|\mathrm{fl}(z) - z| \leq |a - z| = z - a \leq m - a = \delta := (b - a)/2.$$

and if $z \geq m$ then
$$|\mathrm{fl}(z) - z| \leq |b - z| = b - z \leq b - m = \delta.$$

Therefore, $|\mathrm{fl}(z) - z| \leq \delta$. If $z < m$ then

$$\mathrm{fl}(z) \leq |\mathrm{fl}(z) - z| + (z - a) + a \leq \delta + z < \delta + m = b,$$

and $\mathrm{fl}(z) < b$. If $z > m$ then

$$\mathrm{fl}(z) \geq b - (b - z) - |z - \mathrm{fl}(z)| \geq z - \delta > m - \delta = a,$$

and $\mathrm{fl}(z) > a$. $\qquad\square$

Proof of Proposition 30 If $z \in \mathbb{R}$ then either (i) $z \in \mathcal{A}_1$ or (ii) $z \in \mathcal{A}_2 \setminus \mathcal{A}_1$. In case (i), for $x \in \mathcal{F}$ we have that $|x - z| \geq |\mathrm{fl}_1(z) - z|$ by hypothesis. Therefore, $|x - z| \geq |\mathrm{fl}_1(z) - z| = |\mathrm{fl}(z) - z|$ in case (i). In case (ii), for $x \in \mathcal{F}$ we have that $|x - z| \geq |\mathrm{fl}_2(z) - z| = |\mathrm{fl}(z) - z|$. As a result, $|x - z| \geq |\mathrm{fl}(z) - z|$ in both cases and fl rounds to nearest in $\mathcal{F}$. $\qquad\square$

Proof of Proposition 31 We assume that there exists $f_2 : \mathbb{R} \to \mathbb{R}$ which rounds to nearest in $\mathcal{F}$. Take $\mathcal{A}_1 = \mathcal{A}$ and $\mathcal{A}_2 = \mathbb{R} \setminus \mathcal{A}$. Prop. 30 with $f_1 = f$ implies that there exists fl which rounds to nearest in $\mathcal{F}$ and is such that $\mathrm{fl}(z) = f(z)$ for $z \in \mathcal{A}$. $\qquad\square$

5.6 Tightness

In this section we prove the propositions regarding tightness, and present and prove additional propositions about this subject.

5.6.1 Propositions

In this section we present additional propositions regarding tightness.

Proposition 32 (Tightness and continuity) *Let $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ be topological spaces and $\mathcal{R}$ a set. If $g : \mathcal{A} \times \mathcal{B} \to \mathcal{C}$ is continuous and $h : \mathcal{A} \times \mathcal{R} \to \mathcal{B}$ is tight then $f : \mathcal{A} \times \mathcal{R} \to \mathcal{C}$ given by $f(a,r) = g(a,h(a,r))$ is tight. In particular, if $\mathcal{R}$ is a tight set of functions from $\mathcal{A}$ to $\mathcal{B}$ then the function $f : \mathcal{A} \times \mathcal{R} \to \mathcal{B}$ given by $f(a,r) = g(a,r(a))$ is tight.* ▲

Proposition 33 (Tight chain rule) *Let $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ be topological spaces and let $\mathcal{T}$ and $\mathcal{U}$ be sets. If the functions $h : \mathcal{A} \times \mathcal{T} \to \mathcal{B}$ and $g : \mathcal{B} \times \mathcal{U} \to \mathcal{C}$ are tight then the function $f : \mathcal{A} \times (\mathcal{T} \times \mathcal{U}) \to \mathcal{C}$ given by $f(a,(t,u)) := g(h(a,t),u)$ is tight.* ▲

5.6.2 Proofs

This section contains the proofs of the propositions regarding tightness.

Proof of Proposition 19 Let $\mathcal{R}$ be the set of all functions which round to nearest in $\mathcal{F}$ and let $\mathcal{S} = \{(z_k, \mathrm{fl}_k), k \in \mathbb{N}\} \subset \mathbb{R} \times \mathcal{R}$ be a sequence with $\lim_{k \to \infty} z_k = z$. Props. 12, 13 and 17 imply that there exist $a, b \in \mathcal{F}$ and $\delta > 0$ such that if $|y - z| < \delta$ then $\mathrm{fl}(y) \in \{a, b\}$ for $\mathrm{fl} \in \mathcal{R}$. Let $m \in \mathbb{N}$ be such that $k > m \Rightarrow |z_k - z| < \delta$ and define $\mathcal{A} := \{k \geq m \text{ with } \mathrm{fl}_k(z_k) = a\}$ and $\mathcal{B} := \{k \geq m \text{ with } \mathrm{fl}_k(z_k) = b\}$. Since $\mathcal{A} \bigcup \mathcal{B} = \{k \geq m, k \in \mathbb{N}\}$ is infinite, $\mathcal{A}$ or $\mathcal{B}$ is infinite. By exchanging a and b if necessary, we may assume that $\mathcal{A}$ is infinite, and $\{(z_{n_k}, \mathrm{fl}_{n_k}), n_k \in \mathcal{A}\}$ is a subsequence of $\mathcal{S}$. We claim that the function $\mathrm{fl} : \mathbb{R} \to \mathbb{R}$ given by $\mathrm{fl}(w) = \mathrm{fl}_m(w)$ for $w \neq z$ and $\mathrm{fl}(z) = a$ rounds to nearest in $\mathcal{F}$. Indeed, if $z' \in \mathcal{F} \setminus \{z\}$ and $w \in \mathbb{R}$ then

$$\left|w - \mathrm{fl}(z')\right| = \left|w - \mathrm{fl}_m(z')\right| \geq \left|z' - \mathrm{fl}_m(z')\right| = \left|z' - \mathrm{fl}(z')\right|$$

because fl_m rounds to nearest in $\mathcal{F}$, and

$$|w - \mathrm{fl}(z)| = |w - a| = \left|w - \mathrm{fl}_{n_k}(z_k)\right| \geq \left|z_k - \mathrm{fl}_{n_k}(z_k)\right| = |z_k - a| = |z_k - \mathrm{fl}(z)|.$$

because the fl_{n_k} round to nearest in $\mathcal{F}$. Taking the limit $k \to \infty$ in the equation above we obtain $|w - \mathrm{fl}(z)| \geq |z - \mathrm{fl}(z)|$, and fl rounds to nearest in $\mathcal{F}$. Finally,

$$\lim_{k \to \infty} \varphi(z_{n_k}, \mathrm{fl}_{n_k}) = \lim_{k \to \infty} \mathrm{fl}_{n_k}(z_{n_k}) = a = \mathrm{fl}(z)$$

and $\mathcal{R}$ is tight. □

Proof of Proposition 20 For $n = 0$, $T_0(\mathbf{z}, \mathrm{Fl}) = 0$ and Prop. 20 follows from Prop. 32, because constant functions are continuous. Assuming that Prop. 20 holds for $n \geq 0$, let us show that it holds for $n + 1$. By induction and Prop. 32 the function $h : \mathbb{R}^{n+1} \times \mathcal{R}^n \to \mathbb{R}^{n+1} \times \mathbb{R}$ given by $h(\mathbf{w}, \mathrm{Fl}) := (T_n(P_n \mathbf{w}, \mathrm{Fl}), w_{n+1})$ is tight. The function $g : (\mathbb{R}^{n+1} \times \mathbb{R}) \times \mathcal{R} \to \mathbb{R}^{n+2}$ given by $g((\mathbf{w}, z), \mathrm{fl}) := (\mathbf{w}, \mathrm{fl}(w_{n+1} + z))$ is also tight by Prop. 32 because $\mathcal{R}$ is tight. Finally, Prop. 20 follows from Prop. 33 for

$f = T_n$, g and h because $T_{n+1}(\mathbf{z}, \text{Fl}) = g(h(\mathbf{w}, P_n\text{Fl}), \text{fl}_{n+1})$. $\qquad\square$

Proof of Proposition 32   Let $\{(a_k, r_k), k \in \mathbb{N}\} \subset \mathcal{A} \times \mathcal{R}$ be a sequence with $\lim_{k\to\infty} a_k = a$. Since h is tight, there exists $r \in \mathcal{R}$ and a subsequence $\{(a_{n_k}, r_{n_k}), k \in \mathbb{N}\}$ such that $\lim_{k\to\infty} h(a_{n_k}, r_{n_k}) = h(a, r)$. By continuity of g,

$$\lim_{k\to\infty} f(a_{n_k}, r_{n_k}) = \lim_{k\to\infty} g(a_{n_k}, h(a_{n_k}, r_{n_k})) = g(a, h(a, r)) = f(a, r),$$

and f is tight. To handle the particular case, note that when $\mathcal{R}$ is set of tight functions as in the hypothesis the function $h : \mathcal{A} \times \mathcal{R} \to \mathcal{B}$ given by $h(a, r) = r(a)$ is tight. $\qquad\square$

Proof of Proposition 33   Let $\{(a_k, (t_k, u_k)), k \in \mathbb{N}\} \subset \mathcal{A} \times (\mathcal{T} \times \mathcal{U})$ be a sequence such that $\lim_{k\to\infty} a_k = a$. Since h is tight, there exists $t \in \mathcal{T}$ and a subsequence $\{(a_{n_k}, t_{n_k}), k \in \mathbb{N}\}$ of $\{(a_k, t_k), k \in \mathbb{N}\}$ such that $b_{n_k} := h(a_{n_k}, t_{n_k})$ satisfies $\lim_{k\to\infty} b_{n_k} = h(a, t) =: b$. Since b_{n_k} converges to b and g is tight, there exists $u \in \mathcal{U}$ and a subsequence $\{(b_{m_k}, u_{m_k}), k \in \mathbb{N}\}$ of $\{(b_{n_k}, u_{n_k}), k \in \mathbb{N}\}$ such that

$$g(b, u) = \lim_{k\to\infty} g(b_{m_k}, u_{m_k}) = \lim_{k\to\infty} g(h(a_{m_k}, t_{m_k}), u_{m_k}).$$

This leads to

$$f(a, (t, u)) = g(h(a, t), u) = g(b, u) =$$
$$\lim_{k\to\infty} g(h(a_{m_k}, t_{m_k}), u_{m_k}) = \lim_{k\to\infty} f(a_{m_k}, (t_{m_k}, u_{m_k})),$$

and f is tight. $\qquad\square$

5.7   Examples

In this section we verify the examples 2 to 5. Example 1 needs no verification.

Verification of Example 2   Our parcels are $y_0 := 1$ and $y_k := 1 + 2^{\lfloor \log_2(k+1) \rfloor} u$ for $k = 1, \ldots, n := 2^m - 1$ and we break ties downward. If $1 \leq 2^\ell - 1 \leq k < 2^{\ell+1} - 1$ then $y_k = 1 + 2^\ell u$ and we now show by induction that, for $k \geq 1$,

$$\sum_{i=0}^{k} y_i = k + 1 + \frac{4^\ell + 2}{3} u + \left(k + 1 - 2^\ell\right) 2^\ell u \qquad \text{and} \qquad \text{fl}\left(\sum_{i=0}^{k} y_i\right) = k + 1. \quad (93)$$

Indeed, for $k = 1$ we have $\ell = 1$ and $y_0 + y_1 = 2 + 2u$, the first equality in Equation (93) is clearly correct and the second holds because we break ties downward.

If $\text{fl}\left(\sum_{i=0}^{k} y_i\right) = k + 1$ and $2^\ell - 2 \leq k < 2^{\ell+1} - 2$ then $y_{k+1} = 1 + 2^\ell u$ and

$$\text{fl}\left(\sum_{i=0}^{k+1} y_i\right) = \text{fl}\left(k + 1 + 1 + 2^\ell u\right) = k + 2 + 2^\ell u = k + 2 = (k + 1) + 1,$$

because $k + 2 \geq 2^\ell$ and we break ties downward. Therefore, $\text{fl}\left(\sum_{i=0}^{k} y_i\right) = k + 1$.

Let us now assume that the first Equation in (93) holds for k and show that it holds for $k + 1$. When $2^\ell - 1 \leq k < 2^{\ell+1} - 2$ we have that $2^\ell - 1 \leq k + 1 < 2^{\ell+1} - 1$ and

$$\sum_{i=0}^{k+1} y_i = \left(\sum_{i=0}^{k} y_i\right) + y_{k+1} = k + 1 + \frac{4^\ell + 2}{3} u + \left(k + 1 - 2^\ell\right) 2^\ell u + 1 + 2^\ell u \quad (94)$$

52

$$= (k+1) + 1 + \frac{4^{\ell+1}+2}{3}u + \left((k+1)+1-2^{\ell+1}\right)2^{\ell+1}u$$

and the first equality in Equation (93) holds for $k+1$. For $k = 2^{\ell+1} - 2$, we have that

$$2^{\ell+1} - 1 = k+1 < 2^{(\ell+1)+1} - 1$$

and Equation (94) leads to

$$\sum_{i=0}^{k+1} y_i = k+2 + \frac{4^\ell + 2}{3}u + 4^\ell u = (k+1)+1+\frac{4^{\ell+1}+2}{3}u + \left((k+1)+1-2^\ell\right)2^{\ell+1}u$$

because $k+2-2^{\ell+1} = 0$, and the first equality in Equation (93) is satisfied for $k+1$.

Finally, for $n = 2^m - 1$ we have that $\ell = m$ and

$$\frac{1}{u}\left(\sum_{k=0}^n y_k - \mathrm{fl}\left(\sum_{k=0}^n y_k\right)\right) = \frac{4^m+2}{3} = \frac{(n+1)^2+2}{3} = \frac{n^2+2n+3}{3}.$$

The last equation in Example 2 follows from the equation above and the fact that $y_k < 2$ when $2^m u < 1$. $\qquad\square$

Verification of Example 3 Let us define $\rho := u^{-k}$. Since $x_k = \rho^k$ and we break ties downward, we have $\mathrm{fl}\left(\sum_{i=0}^k x_i\right) = \rho^k$ and

$$\sum_{i=0}^k x_i = \frac{\rho^{k+1}-1}{\rho-1} \quad \text{and} \quad \sum_{k=1}^n \sum_{i=0}^k x_i = \frac{\rho^{n+2}-\rho^2-n(\rho-1)}{(\rho-1)^2} = \frac{1}{u^n}\frac{1-u^n-nu^{n+1}(1-u)}{(1-u)^2}.$$

It follows that

$$\sum_{k=0}^n x_k - \mathrm{fl}\left(\sum_{k=0}^n x_k\right) = \frac{\rho^{n+1}-1}{\rho-1} - \rho^n = \frac{\rho^n-1}{\rho-1} = \frac{1}{u^{n-1}}\frac{1-u^n}{1-u} = \kappa_n u \sum_{k=1}^n \sum_{i=0}^k x_i$$

for

$$\kappa_n := \frac{(1-u)(1-u^n)}{1-u^n-nu^{n+1}(1-u)}.$$

If $2nu < 1$ then $1-u \le \kappa_n \le (1-u)(1+u^n)$ because

$$0 < \frac{\kappa_n}{1-u} - 1 = \frac{nu^{n+1}(1-u)}{1-u^n-nu^{n+1}(1-u)} = u^n\frac{nu(1-u)}{1-u^n-nu^{n+1}(1-u)} < u^n.$$

$\qquad\square$

Verification of Example 4 Recall that $x_0 := u$, $x_1 := 1$ and

$$x_k := \beta^{e_k}(1+u) - \beta^{e_{k-1}}(1+2u)$$

for $k \ge 2$, with $0 = e_1 < e_2 \cdots < e_n \in \mathbb{Z}$. Induction using the basic properties of rounding to nearest in Prop. 11 shows that

$$s_k := \sum_{i=0}^k x_i = \beta^{e_k}(1+u) - u\sum_{i=1}^{k-1}\beta^{e_i}, \qquad\qquad \hat{s}_k := \mathrm{fl}\left(\sum_{i=0}^k x_i\right) = \beta^{e_k}(1+2u)$$

53

for $k \geq 1$ and

$$\sum_{k=1}^{n} s_k = (1+u)\sum_{k=1}^{n} \beta^{e_k} - u\sum_{k=1}^{n}\sum_{i=1}^{k-1} \beta^{e_i} = (1+u)\sigma_n - u\sum_{k=1}^{n} \sigma_{k-1},$$

for $\sigma_k := \sum_{i=1}^{k} \beta^{e_i}$ (we assume that $\sum_{i}^{k} a_k = 0$ when $k < i$.) Therefore

$$\hat{s}_n - s_n = \left(\beta^{e_n} + \sum_{k=1}^{n-1} \beta^{e_k}\right) u = u\sum_{k=1}^{n} \beta^{e_k} = u\sigma_n,$$

and

$$\frac{\hat{s}_n - s_n}{\sum_{k=1}^{n} s_k} = \frac{\sigma_n u}{\sigma_n + u\left(\sigma_n - \sum_{k=0}^{n-1} \sigma_k\right)} = \frac{u}{1 + u\left(1 - \sum_{k=1}^{n-1} v_k\right)} \qquad \text{for} \qquad v_k := \frac{\sigma_k}{\sigma_n}. \quad (95)$$

Note that

$$\sigma_{(k+1)} - 1 = \sum_{i=1}^{k+1} \beta^{e_i} - 1 = \sum_{i=2}^{k+1} \beta^{e_i} \geq \beta \sum_{i=2}^{k+1} \beta^{e_{(i-1)}} = \beta \sum_{i=1}^{k} \beta^{e_i} = \beta\sigma_k.$$

Since $\sigma_0 = 0$ and $1/\sigma_n = v_1 = \sigma_1/\sigma_n$, dividing the last equation by σ_n we obtain

$$v_1 + \beta v_k - v_{k+1} \leq 0 \qquad \text{for } k = 1,\ldots,n-2, \qquad \text{and} \qquad v_1 + \beta v_{n-1} \leq 1. \quad (96)$$

We end the verification of Example 4 using a duality argument to prove that

$$\sum_{k=1}^{n-1} v_k \leq \frac{1}{\beta - 1} - \frac{n}{\beta^n - 1}. \quad (97)$$

This equation combined with Equation (95) shows that the value of τ_n mentioned in Example 4 is appropriate. We use basic facts about duality in linear programming [3] applied to the problem with variables v_k, objective function $\sum_{k=1}^{n-1} v_k$ and constraints given by $v_k \geq 0$ and Equation (96). This problem can be written as

$$\begin{cases} \text{maximize} & \mathbb{1}^T\mathbf{v} & = & \sum_{k=1}^{n-1} v_k \\ \text{subject to} & \mathbf{A}\mathbf{v} & \leq & \mathbf{e}, & v_k & \geq & 0. \end{cases} \quad (98)$$

where the matrix A has $a_{11} := \beta + 1$, $a_{i1} = 1$ for $1 < i < n$, $a_{ii} = \beta$ for $2 \leq i < n$, $a_{i,i+1} = -1$ for $1 \leq i < n-1$ and the remaining a_{ij} are 0. The vector $\mathbb{1}$ has all its entries equal to 1 and $e_i = 0$ for $1 \leq i < n-1$ and $e_{n-1} = 1$.

This problem has a feasible solution

$$v_k = \frac{\beta^k - 1}{\beta^n - 1}, \qquad \text{for} \qquad k = 1,\ldots,n-1$$

and

$$\sum_{k=0}^{n-1} v_k = \frac{1}{\beta^n - 1}\left(\frac{\beta^n - 1}{\beta - 1} - n\right) = \frac{1}{\beta - 1} - \frac{n}{\beta^n - 1}. \quad (99)$$

Its dual has $n - 1$ variables, which we call $y_1,\ldots,y_{n-1}$, and is

$$\begin{cases} \text{minimize} & \mathbf{e}^T\mathbf{y} & = & y_{n-1} \\ \text{subject to} & \mathbf{A}^T\mathbf{y} & \geq & \mathbb{1}, & y_k & \geq & 0. \end{cases} \quad (100)$$

54

We claim that the vector $\mathbf{y} \in \mathbb{R}^{n-1}$ with entries

$$y_{n-1} = \frac{1}{\beta - 1} - \frac{n}{\beta^n - 1} \quad \text{and} \quad y_k = \beta^{n-k-1} y_{n-1} + \frac{1}{\beta - 1} \quad \text{for } k = 1 \dots n-2,$$

is a feasible solution of the dual problem. Indeed, $y_{n-1} \geq 0$ because

$$\frac{\beta^n - 1}{\beta - 1} = \sum_{k=0}^{n-1} \beta^k \geq n,$$

and the other entries of $\mathbf{y}$ are clearly non negative because $y_{n-1} \geq 0$. The first inequality in the system $\mathbf{A}^{\mathsf{T}} \mathbf{y} \geq \mathbb{1}$ is satisfied because

$$(\beta + 1) y_1 + \sum_{k=2}^{n-1} y_k = \left((\beta + 1) \beta^{n-2} + \sum_{k=2}^{n-2} \beta^{n-k-1} + 1 \right) y_{n-1} + \frac{\beta + n - 2}{\beta - 1}$$

$$= \left(\sum_{k=0}^{n-1} \beta^k \right) y_{n-1} + \frac{\beta + n - 2}{\beta - 1}$$

$$= \frac{\beta^n - 1}{\beta - 1} \left(\frac{1}{\beta - 1} - \frac{n}{\beta^n - 1} \right) + \frac{\beta + n - 2}{\beta - 1} = \frac{\beta^n - \beta}{(\beta - 1)^2} + 1 \geq 1,$$

and the remaining inequalities are satisfied as equalities, because

$$-y_{k-1} + \beta y_k = -\beta^{n-k} y_{n-1} - \frac{1}{\beta - 1} + \beta \beta^{n-k-1} y_{n-1} + \frac{\beta}{\beta - 1} = 1.$$

The value of the objective function of the dual problem for $\mathbf{y}$, y_{n-1}, is equal to the value of the objective function of the primal problem in (99). Therefore, this is the optimal value of both problems and Equation (97) holds. The linear programming problem above also shows that the worst case in Equation (21) is achieved for $e_k = k - 1$, because these exponents lead to the v_k in the solution of the primal problem. $\quad \square$

Verification of Example 5   Recall that $x_0 := u$, $x_1 := 1$ and $x_k := -2^{1-k}(1 + 3u)$ for $k > 1$. It follows by induction that

$$\sum_{i=0}^{k} x_i = 2^{1-k}(1 + 3u) - 2u \quad \text{and} \quad \mathrm{fl}\left(\sum_{i=0}^{k} x_i \right) = 2^{1-k}(1 + 2u).$$

Since $2^n u \leq 1$, we have

$$\sum_{k=1}^{n} \left| \sum_{i=0}^{k} x_i \right| = 2\left(1 - 2^{-n}\right)(1 + 3u) - 2nu > 0,$$

and Equations (23) follows from the expressions above. Finally, since $2^{-n} \geq u$, we have that $nu < 1$ and

$$\kappa_n - (1 - u) = u \frac{(2^{-n} - u)n + 3(1 - 2^{-n})u}{(1 - 2^{-n})(1 + 3u) - nu} > 0$$

and

$$1 - \kappa_n = u \frac{1 - 2^{-n}(n + 1)}{(1 - 2^{-n})(1 + 3u) - nu} \geq 0.$$

$\quad \square$

References

[1] Boldo, S., Melquiond, G., Flocq: A unified library for proving floating-point algorithms in Coq. In: Antelo, E., Hough, D., Ienne, P. (eds.) 20th IEEE Symposium on Computer Arithmetic, 243–252. Tübingen, Germany, 2011.

[2] Boldo, S., Stupid is as Stupid Does: Taking the Square Root of the Square of a Floating-Point Number. In Sergiy Bogomolov and Matthieu Martel, editors, Proceedings of the Seventh and Eighth International Workshop on Numerical Software Verification, volume 317 of Electronic Notes in Theoretical Computer Science, 50–55, Seattle, 2015.

[3] Chvátal, V., Linear Programming. W.H. Freeman. 1983.

[4] Cody, W. and Waite, W., *Software Manual for the Elementary Functions.* Prentice-Hall, Englewood Cliffs, NJ, 1980.

[5] de Camargo, A. Pierro, Mascarenhas, W. F., The stability of extended Floater-Hormann interpolants, arXiv:1409.2808v5 [math.NA] 27 May 2015

[6] Demmel, J., Effects of Underflow on Solving Linear Systems, Technical Report, Computer Science Division, U.C. Berkeley, 1981.

[7] Demmel, J., Underflow and the Reliability of Numerical Software SIAM J. Sci. and Stat. Comput., 5(4), 887–919. 1984.

[8] Fousse, L., Hanrot, G., Lefèvre, V., Pélissier, P. and Zimmermann, P., MPFR: A Multiple-Precision Binary Floating-Point Library with Correct Rounding, ACM TOMS, 2007.

[9] Higham, N. J., *The accuracy of floating point summation*, SIAM J. Sci. Comput. 14:4, 783–799, 1993.

[10] Higham, N. J., *Accuracy and stability of numerical algorithms*, second edition, SIAM, 2002.

[11] IEEE Computer Society, *IEEE Standard for Floating-Point Arithmetic,* doi:10.1109/IEEESTD.2008.4610935. ISBN 978-0-7381-5753-5. 2008.

[12] Kahan, W., Mathematics written in sand – the HP–15C, Intel 8087, etc. http://www.cs.berkeley.edu/ wkahan/MathSand.pdf, 1983.

[13] Kulisch, U., Computer Arithmetic and Validity: Theory, Implementation, and Applications (de Gruyter Studies in Mathematics), 2008.

[14] Knuth, D., The Art of Computer Programming, vol. 2, Seminumerical Algorithms, second edition, Addison Wesley, 1981.

[15] Mascarenhas, W. F., The stability of barycentric interpolation at the Chebyshev points of the second kind, Numer. Math., 128:2, 265–300, 2014.

[16] Mascarenhas, W. F. and de Camargo, A. Pierro, On the backward stability of the second barycentric formula for interpolation, Dolomites Res. Notes Approx. 7, 1–12, 2014.

[17] Mascarenhas, W. F. and de Camargo, A. Pierro, The effects of rounding errors on barycentric interpolation (extended version, with complete proofs.), arXiv:1309.7970v3 [math.NA] 12 Jan 2016. Online version in Numerische Mathematik: DOI 10.1007/s00211-016-0798-x, 2016.

[18] C.-P. Jeannerod, S.M. Rump. Improved error bounds for inner products in floating-point arithmetic. SIAM. J. Matrix Anal. & Appl. (SIMAX), 34(2):338–344, 2013.

[19] S.M. Rump. Error estimation of floating-point summation and dot product. BIT Numerical Mathematics, 52(1):201–220, 2012.

[20] M. Lange, S. M. Rump, Sharp estimates for perturbation errors in summations, available in http://www.ti3.tuhh.de/paper/rump/LaRu2017a.pdf

[21] Neumaier, A., Inner product rounding error analysis in the presence of underflow, Computing, 34 (4), 365–373, 1985.

[22] Wilkinson, J. H., *The Algebraic Eigenvalue Problem*, Clarendon Press, 1965.

[23] Wilkinson, J.H., A priori analysis of algebraic processes. Proceedings of International Congress of Mathematicians, 1966,Moscow. 629–640, 1968.