Contiguous Digit Sets and Local Roundings

Marko Petkovšek*

Department of Mathematics, University of Ljubljana
Ljubljana, Yugoslavia

Abstract: We determine for which roundings addition in the floating-point screen has representable rounding error, which roundings are implied by truncation of digit strings in different radix systems with contiguous digits, and how many additional digits (including possibly a sticky bit) have to be kept in such systems in order to perform a given rounding correctly. Throughout the paper, we emphasize clean separation of approximation from representation.

1 Introduction

The aim of the paper is to show that certain interesting properties of roundings and representations hold at a fairly general level. We restrict our attention to roundings of real numbers.

In Section 2 we define roundings by means of two parameters for each basic interval of the screen. These parameters determine the dividing point of the interval and the direction in which the dividing point moves under rounding. We characterize roundings into the floating-point screen for which the rounding error of the sum of two representable numbers is itself representable.

In Section 3 we show that radix systems in which truncation and lexicographic ordering obey natural monotonicity conditions can be characterized as systems with contiguous digit sets. Furthermore, these systems with slight modifications yield unambiguous and complete representation systems. Our treatment includes negative-base, signed-digit, and sign-magnitude systems.

In Section 4 we examine the interplay of roundings and representations. First we ask what kind of roundings are induced by truncation in different systems. Then we introduce a notion of locality which defines how many digits after the truncation point have to be kept to determine the result of rounding. We also allow for the possibility of a residual (or sticky) bit, calling a rounding which requires it quasi-local. We show that the locality of a rounding equals the number of digits preceding the extremal tail in the digit string representing the dividing point of an interval.

As usual, we shall denote the set of integers by $\mathbb{Z}$, and the set of reals by $\mathbb{R}$. The function $\mathrm{round}(x) : \mathbb{R} \to \mathbb{Z}$ maps $x \in \mathbb{R}$ to the largest integer nearest to x.

2 Roundings

Following Kulisch and Miranker [4], a subset S of a partially ordered set $\{M, \leq\}$ is a *screen* if for every $a \in M$ the set $\{x \in S\,;\, x \leq a\}$ contains a largest element, and $\{x \in S\,;\, x \geq a\}$ contains a least element.

Proposition 1 *A set S of reals is a screen in $\mathbb{R}$ if and only if S is closed, and unbounded in both directions.*

We omit the easy proof.

Definition 1 Let S be a real screen. A function

$$f : \mathbb{R} \longrightarrow S$$

*Currently at School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA.

will be called a *rounding* into S if

(A1) $x \in S$ implies $f(x) = x$,

(A2) $x, y \in \mathbb{R}$ and $x < y$ imply $f(x) \leq f(y)$.

An interval $[a, b]$ such that $a, b \in S$ and $a < x < b$ imply $x \notin S$ will be called a *basic interval* of S.

We note that functions which satisfy (A1) and (A2) are called *monotone roundings* in [4], and *optimal roundings* in [7] and [1].

Let $x \in \mathbb{R}$. Define

$$\begin{aligned} \nabla(x) &:= \max\{r \in S\,;\, r \leq x\}\,, \\ \Delta(x) &:= \min\{r \in S\,;\, r \geq x\}\,, \\ h(x) &:= \Delta(x) - \nabla(x)\,. \end{aligned}$$

For every $x \in \mathbb{R}$, $[\nabla(x), \Delta(x)]$ is a basic interval of S, and $\nabla(x) \leq x \leq \Delta(x)$. If $x \notin S$, then $\nabla(x) < x < \Delta(x)$, while for $x \in S$, $\nabla(x) = x = \Delta(x)$. Both $\nabla(x)$ and $\Delta(x)$ satisfy (A1) and (A2), which shows that there exist roundings into any real screen.

Proposition 2 *Let f be a rounding into S, and $x \in \mathbb{R}$. Then*

(P1) *f maps $\mathbb{R}$ onto S,*

(P2) $f(f(x)) = f(x)$,

(P3) $f(x) \in \{\nabla(x), \Delta(x)\}$,

(P4) *if y is between x and $f(x)$, then $f(y) = f(x)$,*

(P5) $|f(x) - x| \leq h(x)$.

Proof: These properties are easy consequences of (A1) and (A2). □

Theorem 1 *Let S be a real screen. Then a map $f : \mathbb{R} \to S$ is a rounding into S if and only if in every basic interval $[a, b]$ of S there exists a number d such that $f(d) \in \{\nabla(d), \Delta(d)\}$, and for every $x \in \mathbb{R}$,*

(a) *if $a \leq x < d$ then $f(x) = a$,*

(b) *if $d < x \leq b$ then $f(x) = b$.*

This is just a restatement of Theorem 1.28, p. 33 of [4], so we do not give a proof.

We shall call d the *dividing point* of the interval $[a, b]$ under f.

Definition 2 Let f be a rounding into S, and $x \in \mathbb{R}$. We shall denote the dividing point of $[\nabla(x), \Delta(x)]$ under f by $d(x)$. If $h(x) > 0$, let

$$\rho(x) := \frac{d(x) - (\nabla(x) + \Delta(x))/2}{h(x)/2}\,,$$

and

$$\tau(x) := \begin{cases} +1, & \text{if } f(d(x)) = \Delta(x); \\ -1, & \text{if } f(d(x)) = \nabla(x). \end{cases}$$

For short, we shall call $\rho(x)$ and $\tau(x)$ the *parameters* of f on the interval $[\nabla(x), \Delta(x)]$.

Clearly, $-1 \leq \rho(x) \leq 1$. Because of (A1), $\rho(x) = \pm 1$ implies $\tau(x) = \rho(x)$. It is easy to show that

$$|f(x) - x| \;\leq\; \frac{1 + |\rho(x)|}{2}\, h(x)\,, \tag{1}$$

an improvement upon (P5).

According to Theorem 1, f is completely determined by S and by the values of its parameters on all the basic intervals of S.

Theorem 2 *Let f be a rounding into S, and $x \notin S$. Then*

$$\begin{aligned} & f(x) \\ &= \nabla(x) + \operatorname{round}\left(\tau(x)\left(\frac{x - \nabla(x)}{h(x)} - \frac{\rho(x)}{2}\right)\right)\tau(x)h(x) \\ &= \Delta(x) + \operatorname{round}\left(\tau(x)\left(\frac{x - \Delta(x)}{h(x)} - \frac{\rho(x)}{2}\right)\right)\tau(x)h(x)\,. \end{aligned}$$

If, furthermore, $\nabla(x)/h(x)$ is an integer, then

$$f(x) = \operatorname{round}\left(\tau(x)\left(\frac{x}{h(x)} - \frac{\rho(x)}{2}\right)\right)\tau(x)h(x)\,.$$

Proof: These formulas are straightforward consequences of Definition 2 and the properties of round(x). □

Definition 3 A rounding will be called *uniform* if there exist constants ρ, τ such that $\rho(x) = \rho$ and $\tau(x) = \tau$ for all $x \notin S$. We shall denote such a rounding by $U_{\rho,\tau}(x)$.

A rounding will be called *semi-uniform* if $0 \in S$ and there exist constants ρ, τ such that $\rho(x) = \rho\,\mathrm{sgn}(x)$ and $\tau(x) = \tau\,\mathrm{sgn}(x)$ for all $x \notin S$. We shall denote such a rounding by $S_{\rho,\tau}(x)$.

Example 1 $\nabla(x) = U_{1,1}(x)$ and $\Delta(x) = U_{-1,-1}(x)$. $S_{1,1}$ is usually called *chopping* or *truncation*. The name *rounding* is often reserved for roundings with $\rho(x) = 0$. In the IEEE Floating-Point Standard [2], $U_{1,1}$ is called *rounding toward* $-\infty$, $U_{-1,-1}$ *rounding toward* $+\infty$, and $S_{1,1}$ *rounding toward* 0. The fourth rounding defined in the Standard, *rounding to nearest*, has $\rho(x) = 0$ but $\tau(x)$ alternates on successive basic intervals since the last digit kept is always even; this definition assumes standard radix system representation of S. In [4], $U_{1,1}$ is called *downwardly directed rounding* and $U_{-1,-1}$ *upwardly directed rounding.*

Theorem 3 *Let S_1, S_2 be two real screens, and f a rounding into S_1. For each basic interval J of S_2, let $\varphi_J : J \to I$ be a linear function which maps J onto some basic interval I of S_1. Define*

$$g(x) := \begin{cases} x, & \text{if } x \in S_2; \\ \varphi_J^{-1}(f(\varphi_J(x))), & \text{if } x \notin S_2 \text{ and } x \in J. \end{cases}$$

Then g is a rounding into S_2, and the parameters of g on J agree with those of f on I if φ_J is increasing, and differ in sign only if φ_J is decreasing.

Proof: Let $I = [a, b]$, $J = [c, d]$, and $c < x < d$. Then $\varphi_J(x) \notin S_1$, and by Theorem 2,

$$f(\varphi_J(x)) = a + \mathrm{round}\left(\tau\left(\frac{\varphi_J(x) - a}{b - a} - \frac{\rho}{2}\right)\right)\tau(b - a),$$

where ρ and τ are the parameters of f on I.

Let $\varphi_J(x) = Ax + B$. Then $\varphi_J^{-1}(y) = (y - B)/A$. If φ_J is increasing then $a = Ac + B$ and $b = Ad + B$, so that

$$\varphi_J^{-1}(f(\varphi_J(x))) = c + \mathrm{round}\left(\tau\left(\frac{x - c}{d - c} - \frac{\rho}{2}\right)\right)\tau(d - c).$$

By Theorem 2, this agrees with the rounding into S_2 which has parameters ρ and τ on J.

If φ_J is decreasing then $a = Ad + B$ and $b = Ac + B$, so that

$$\varphi_J^{-1}(f(\varphi_J(x))) = d + \mathrm{round}\left((-\tau)\left(\frac{x - d}{d - c} - \frac{(-\rho)}{2}\right)\right)(-\tau)(d - c).$$

By Theorem 2, this agrees with the rounding into S_2 with parameters $-\rho$ and $-\tau$ on J. □

Example 2 Let $\beta, s \in \mathbb{Z}$ and $\beta \geq 2$. The set

$$S(\beta, s) := \{k\beta^{-s};\ k \in \mathbb{Z}\}$$

is the *fixed-point screen*. It is equidistant, and $S(\beta, 0) = \mathbb{Z}$. If f is a rounding into $\mathbb{Z}$, then

$$f_s(x) := f(x\beta^s)\beta^{-s}$$

is a rounding into $S(\beta, s)$. If f is uniform or semi-uniform, then f_s is too, and has the same parameters. This can be proved by means of Theorem 3.

Example 3 Let $\beta, s \in \mathbb{Z}$ and $\beta \geq 2$, $t \geq 1$. The set

$$P(\beta, t) := \{k\beta^e;\ k, e \in \mathbb{Z},\ \beta^{t-1} \leq |k| < \beta^t\} \cup \{0\}$$

is the *floating-point screen*. If f is a rounding into $\mathbb{Z}$, then

$$f_t(x) := \begin{cases} 0, & \text{if } x = 0; \\ f(x\beta^n)\beta^{-n}, & \text{otherwise}, \end{cases}$$

where n is the unique integer satisfying

$$\beta^{t-1} \leq |x|\beta^n < \beta^t,$$

is a rounding into $P(\beta, t)$. If f is uniform or semi-uniform, then f_s is too, and has the same parameters. This again can be proved by means of Theorem 3.

It is well known that the rounding error of a sum of two floating-point numbers is itself a floating-point number when $\rho = 0$ (cf. [3], Section 4.2.2, Theorem B, p. 220). This fact is useful for construction of floating-point algorithms (cf. [4], Section 6.10, p. 192 ff.). Therefore it is of interest to generalize this result to other roundings.

Theorem 4 *Let f be a rounding into the floating-point screen $P(\beta,t)$. Then $f(a+b)-(a+b) \in P(\beta,t)$ for all $a,b \in P(\beta,t)$, if and only if either $|\rho(x)| < 1-2(\beta^{-1}-\beta^{-1-t})$, or $|\rho(x)| = 1-2(\beta^{-1}-\beta^{-1-t})$ and $\operatorname{sgn}\tau(x) = \operatorname{sgn}\rho(x)$, for all $x \notin P(\beta,t)$.*

Proof: Denote $c := \beta^{-1}-\beta^{-1-t}$. We shall only prove sufficiency of $|\rho| < 1-2c$ for representability of error.

Assume that a and b are nonnegative, but consider both addition and subtraction. Also assume that $a \geq b$. As long as $a+b$ has no more than $2t$ digits the error will clearly be representable. For $a+b$ to have more than $2t$ digits, the mantissas of a and b have to be separated by at least one position, so that they together occupy at least $2t+1$ positions. Wlg. assume that a is in the range $[\beta^{t-1},\beta^t)$. Then, even if all of its digits are maximal, b cannot exceed $(\beta-1)\sum_{k=2}^{t+1}\beta^{-k} = \beta^{-1}-\beta^{-1-t}$, so that $0 \leq b \leq c$. We claim that in this case $f(a+b)=a$. Then the error is $-b$, which is a floating-point number.

The entire interval $(a+\frac{\rho-1}{2}, a+\frac{\rho+1}{2})$ rounds into a, so it suffices to show that both b and $-b$ lie between $\frac{\rho-1}{2}$ and $\frac{\rho+1}{2}$. From $|\rho| < 1-2c$ it follows that $\frac{\rho-1}{2} < -c$ and $c < \frac{\rho+1}{2}$, so $\frac{\rho-1}{2} < -b$ and $b < \frac{\rho+1}{2}$. It also follows that $|\rho| < 1$, hence $\frac{\rho-1}{2} < 0 < \frac{\rho+1}{2}$, hence $\frac{\rho-1}{2} < b$ and $-b < \frac{\rho+1}{2}$ proving the claim. □

Example 4 Let $\beta = 10$, $t = 4$, $a = 1000$ and $b = 0.09999$. Then $c = 10^{-1}-10^{-5} = 0.09999$. If $\frac{\rho+1}{2} < c$ then $f(a+b) = 1001$, and the error 0.90001 is not representable. If $\frac{\rho+1}{2} > c$ then $f(a+b) = 1000$, and the error 0.09999 is representable.

3 Representation with contiguous digit sets

Definition 4 A *representation system* for reals is a pair (R,v) where R is a set and v a bijection from R into $\mathbb{R}$.

This simple definition requires that representation be complete nad nonambiguous. To construct useful representation systems, we use integer radix systems with special properties. We study these systems in the main part of this section, and return to representation systems at the end of the section.

An *integer radix system* $\mathcal{S}(\beta,\mathcal{D})$ is given by *base* β and *digit set* D where β is an integer, $|\beta| \geq 2$, and D is a finite set of integers containing 0. Let

$$R_n := \prod_{k>n}\{0\} \times \prod_{k\leq n} D, \qquad \text{for } n \in \mathbb{Z},$$

and

$$R := \bigcup_{n\in\mathbb{Z}} R_n.$$

Thus R is the set of all two way sequences of digits in which all positions from some place on (in direction of increasing subscripts) contain zeros.

Define the *valuation function* $v : R \to \mathbb{R}$ by setting

$$v(r) := \sum_{k\leq n} r_k\beta^k$$

where $r = \ldots 00r_nr_{n-1}r_{n-2}\ldots \in R_n$ and $r_k \in D$ for $k \leq n$. The number $v(r)$ is the *value* of r, and r is a *representation* of $v(r)$. A representation r is *nonzero* if $r_n \neq 0$ for some $n \in \mathbb{Z}$. Let

$$F_n := R \bigcap \left(\prod_{k\geq n} D \times \prod_{k<n}\{0\} \right), \qquad \text{for } n \in \mathbb{Z},$$

and

$$F := \bigcup_{n\in\mathbb{Z}} F_n.$$

F_0 is the set of *integral* representations, and F the set of *finite* representations.

A number is *representable in* $\mathcal{S}$ if it belongs to $v(R)$, and *ambiguous in* $\mathcal{S}$ if it has more than one representation in $\mathcal{S}$. The system $\mathcal{S}$ is *complete over* $\mathbb{R}$ if $v(R) = \mathbb{R}$, and *unambiguous* if v is one-to-one.

In [5], a digit set D is called *basic for* β if every integer has a unique integral representation in $\mathcal{S}(\beta, D)$. If $\beta > 0$, $d \geq 0$ for all $d \in D$ and every nonnegative integer has a unique integral representation in $\mathcal{S}(\beta, D)$ then D is *positive semi-basic* for β. Analogously, we shall call D *negative semi-basic* for β if $-D$ is positive semi-basic for β. The following theorem summarizes some results about basic and semi-basic digit sets:

Theorem 5 ([5], [6])

1. *If D is basic for β then $\mathcal{S}(\beta, D)$ is complete over $\mathbb{R}$.*
2. *If D is basic for β then D is a complete residue system mod $|\beta|$.*
3. *If D consists of $|\beta|$ contiguous integers which include 0, and either D contains positive and negative values, or β is negative, or both, then D is basic for β.*
4. *There are, however, many noncontiguous basic digit sets. In fact, if $|\beta| \geq 3$ then there are infinitely many basic digit sets for β; for instance, the sets $\{-1, 0, 3^n - 2\}$ are basic for $\beta = 3$, for all positive integers n.*
5. *Let $\beta > 0$. Then the set $D_\beta := \{0, 1, \ldots, \beta - 1\}$ is the only positive semi-basic digit set for β.*

In view of our definitions, the standard integer radix systems with base $\beta > 0$ and digit set D_β enhanced with a sign can be seen as consisting of the two systems with semi-basic digit sets for β, namely $\mathcal{S}(\beta, D_\beta)$ and $\mathcal{S}(\beta, -D_\beta)$.

Let $\mathcal{S}(\beta, D)$ be an integer radix system. Define

$$\begin{aligned} d_{\max} &:= \max D \\ d_{\min} &:= \min D \\ \mu_k &:= \begin{cases} d_{\max}, & \text{if } \beta^k > 0 \\ d_{\min}, & \text{if } \beta^k < 0 \end{cases} \\ \nu_k &:= \begin{cases} d_{\min}, & \text{if } \beta^k > 0 \\ d_{\max}, & \text{if } \beta^k < 0 \end{cases} \\ m &:= \min v(R_{-1}) \\ M &:= \max v(R_{-1}) . \end{aligned}$$

Then one can see easily that

$$\begin{aligned} m &= \frac{d_{\min}}{\beta - 1} \\ M &= \frac{d_{\max}}{\beta - 1} \end{aligned}$$

when $\beta > 0$, and

$$\begin{aligned} m &= \frac{\beta d_{\max} + d_{\min}}{\beta^2 - 1} \\ M &= \frac{\beta d_{\min} + d_{\max}}{\beta^2 - 1} \end{aligned}$$

when $\beta < 0$. In both cases,

$$M - m = \frac{d_{\max} - d_{\min}}{|\beta| - 1} .$$

For $r \in R$, let the mappings $S, T, T_n : R \to R$, for $n \in \mathbb{Z}$, be defined by

$$\begin{aligned} S(r)_i &:= r_{i-1} \quad \text{(left shift)} \\ T(r)_i &:= \begin{cases} r_i, & \text{if } i \geq 0 \\ 0, & \text{otherwise} \end{cases} \quad \text{(truncation)} \\ T_n(r) &:= S^n T S^{-n} \quad \text{(truncation after } n\text{-th digit)} . \end{aligned}$$

Then obviously

$$\begin{aligned} v(S(r)) &= \beta v(r) \qquad (2) \\ T_0 &= T \\ T_m T_n &= T_{\max\{m,n\}} \end{aligned}$$

Let g denote the smallest gap in the set of integers which have integral representations in $\mathcal{S}(\beta, D)$; that is, $g := \min\{|v(s) - v(r)|;\ s, r \text{ integral}\}$.

Proposition 3 *In any integer radix system $\mathcal{S}(\beta, D)$, the following conditions are equivalent:*

1. *For all $r, s \in R$, if $v(T(r)) < v(T(s))$ then $v(r) \leq v(s)$.*
2. *For all $r, s \in R$, if $v(r) < v(s)$ then $v(T(r)) \leq v(T(s))$.*
3. *For all $r, s \in R$, if $n = \max\{k;\ r_k \neq s_k\}$ and $r_n \beta^n < s_n \beta^n$ then $v(r) \leq v(s)$.*
4. *For all $r, s \in R$, if $n = \max\{k;\ r_k \neq s_k\}$ and $v(r) < v(s)$ then $r_n \beta^n < s_n \beta^n$.*

5. $d_{\max} - d_{\min} \le g(|\beta| - 1)$.

Sketch of proof: It is rather straightforward to prove the equivalence of 1, 3 and 5. Notice that 2 and 4 are just the contrapositives of 1 and 3, respectively. □

By Theorem 5, every contiguous digit set with $|\beta|$ digits is basic or semi-basic for β. Then $d_{\max} - d_{\min} = |b| - 1$, $d_{\max} \le |b| - 1$, and $d_{\min} \ge 1 - |b|$; hence $M - m = 1$, $0 \le M \le 1$ and $-1 \le m \le 0$. Also, $g = 1$, and so condition 5 of Proposition 3 is fulfilled. Therefore the other four conditions are fulfilled, too. Conversely, one can see easily that a system $\mathcal{S}(\beta, D)$ with $|D| \ge |\beta|$, $g = 1$, and which satisfies any one of the monotonicity conditions of Proposition 3, has contiguous digits.

Definition 5 A representation $r \in R$ is *maximal* if there exists an $n \in \mathbb{Z}$ such that $r = \ldots r_{n+2} r_{n+1} \mu_n \mu_{n-1} \ldots$. A representation $r \in R$ is *minimal* if there exists an $n \in \mathbb{Z}$ such that $r = \ldots r_{n+2} r_{n+1} \nu_n \nu_{n-1} \ldots$. A representation $r \in R$ is *extremal* if it is either maximal or minimal. The set of extremal representations will be denoted by E. The sets of maximal and minimal nonzero representations will be denoted by E_M and E_m, respectively.

The left shift mapping S preserves extremal representations; if $\beta > 0$ then it preserves both maximal and minimal representations, and if $\beta < 0$ it swaps them. The set of values of extremal representations can be written as

$$v(E) = \{(k + M)\beta^n \,;\, k, n \in \mathbb{Z}\} = \{(k + m)\beta^n \,;\, k, n \in \mathbb{Z}\}.$$

Define

$$\mathrm{depth}(r) := \begin{cases} \min\{n \,;\, v(r)\beta^n - M \in \mathbb{Z}\}, & \text{if } r \in E; \\ +\infty, & \text{otherwise.} \end{cases}$$

Proposition 4 *In a system with $|\beta|$ contiguous digits, a real number is ambiguous if and only if it belongs to $v(E)\backslash\{0\}$, and every ambiguous number has exactly two distinct representations, one of which is maximal and the other minimal.*

Sketch of proof: If we have two distinct representations of the same value, we can shift them if necessary so that the first position where they differ is 0. Then the values of the integral parts differ by at least 1, and so do the values of the fractional parts. But this is only possible if their values are M and m, respectively. It follows that the two representations are extremal. □

Now we return to representation systems. To obtain a representation system from a system $\mathcal{S}(\beta, D)$, we have to enforce completeness and nonambiguity. If D is basic then $\mathcal{S}(\beta, D)$ is complete, and if furthermore the digits are contiguous we can achieve nonambiguity, according to Proposition 4, by removing all maximal or all minimal nonzero representations. This gives us the basic representation systems.

If D is semi-basic then we can achieve completeness by introducing the sign, that is, by taking the union of representations yielded by the two systems $\mathcal{S}(\beta, D)$ and $\mathcal{S}(\beta, -D)$. To achieve nonambiguity, we remove maximal nonzero representations from the system with positive semi-basic digit set, and minimal nonzero representations from the system with negative semi-basic digit set. This gives us the standard representation system with integer base $\beta \ge 2$ and digits $0, 1, \ldots, \beta - 1$.

Definition 6 Let $\mathcal{S}(\beta, D)$ be an integer radix system with basic or semi-basic contiguous digit set D. If D is basic the representation systems $(R\backslash E_M, v)$ and $(R\backslash E_m, v)$ will be called *basic representation systems* and will be denoted by $\mathcal{S}_\beta^{+1}(D)$ and $\mathcal{S}_\beta^{-1}(D)$, respectively. If D is positive semi-basic the representation system $((R\backslash E_M) \cup (R^-\backslash E_m^-), v \cup v^-)$ where R^-, E_m^-, v^- refer to the system $\mathcal{S}(\beta, -D)$, will be called the *standard representation system*, and will be denoted by $\mathcal{S}_\beta^s$.

4 Truncation and locality

In representation systems the evaluation function v is invertible. If f maps R into F_n, then vfv^{-1} maps $\mathbb{R}$ into $v(F_n)$. We shall call vfv^{-1} the *map induced by* f. Of special interest are the maps induced by truncations after the n-th digit, T_n. In basic and standard representation systems, the set $v(F_n)$ is equal to the fixed-point screen $S(\beta, -n)$, so we can ask: When is the induced map a rounding?

Proposition 5 *Let $\mathcal{S}$ be a basic representation system. Then $v(R_{-1}) = I$, where*

$$I := \begin{cases} [m, M), & \text{if } \mathcal{S} = \mathcal{S}_\beta^{+1}(D) \\ (m, M], & \text{if } \mathcal{S} = \mathcal{S}_\beta^{-1}(D) \end{cases}$$

Proof: Let $x \in I$. Recursively define the sequences $(x_k)_{k=0}^{-\infty}$ and $(d_k)_{k=0}^{-\infty}$ by

$$\begin{aligned} d_0 &= 0 \\ x_0 &= x \end{aligned}$$

and, for $k \leq -1$,

$$\begin{aligned} d_k &= \begin{cases} \lfloor x_{k+1}\beta - m \rfloor, & \text{if } \mathcal{S} = \mathcal{S}_\beta^{\mathrm{sgn}\beta^k}(D) \\ \lceil x_{k+1}\beta - M \rceil, & \text{if } \mathcal{S} = \mathcal{S}_\beta^{-\mathrm{sgn}\beta^k}(D) \end{cases} \\ x_k &= x_{k+1}\beta - d_k \,. \end{aligned}$$

One can prove by induction on k that $d_k \in D$, $x_k \in I$, and $x_k\beta^k + \sum_{j=0}^{k} d_j\beta^j = x$, for $k = 0, -1, \ldots$. From this it follows that $\ldots 00 d_{-1} d_{-2} \ldots$ is a representation of x in the system $\mathcal{S}_\beta^\sigma(D)$, with $\sigma = +1$ when $I = [m, M)$ and $\sigma = -1$ when $I = (m, M]$. □

Theorem 6 *Let $\sigma \in \{-1, +1\}$. In $\mathcal{S}_\beta^\sigma(D)$,*

$$vTv^{-1}(x) = U_{2M-1,\sigma}(x)\,,$$

and in $\mathcal{S}_\beta^s$

$$vTv^{-1}(x) = S_{1,1}(x)\,.$$

Proof: From Proposition 5 it follows that in the systems $\mathcal{S}_\beta^\sigma(D)$, every real number x has a unique partition in the form

$$x = k + z \tag{3}$$

where $k \in \mathbb{Z}$ and $z \in v(R_{-1})$. If r is a representation of x in $\mathcal{S}_\beta^\sigma(D)$, then $x = v(T(r)) + v(s)$ where $v(T(r)) \in \mathbb{Z}$ and $s \in R_{-1}$. Therefore, by uniqueness of (3),

$$vTv^{-1}(x) = v(T(r)) = k\,.$$

It follows that vTv^{-1} maps the entire interval $k + [m, M)$, when $\sigma = +1$ (the entire interval $k + (m, M]$, when $\sigma = -1$) into k. Hence vTv^{-1} is a rounding into the screen $\mathbb{Z}$ with parameters $\rho = 2M - 1$ and σ (unless $M = 0$ and $\sigma = +1$, or $M = 1$ and $\sigma = -1$). Since D is basic, $0 < M < 1$, hence vTv^{-1} is always a rounding.

In $\mathcal{S}_\beta^s$, $M = 1$ and $\sigma = +1$ for $x > 0$, while $M = -1$ and $\sigma = -1$ for $x < 0$, so that in this case $vTv^{-1}(x) = S_{1,1}(x)$. □

This can be easily generalized to mappings induced by truncation after n-th digit, by using Theorem 3.

In floating-point arithmetic it is important to know how many digits after the truncation point affect the result of rounding. Even if this number is not finite it may still be the case that very little information is required about the removed digits. We formalize these considerations by introducing locality of roundings.

Definition 7 Let $\mathcal{S}$ be a basic or a standard representation system, and f a rounding into the screen $\mathbb{Z}$. Denote $t_n := vT_nv^{-1}$. We shall call f *local* in $\mathcal{S}$ if there exists an integer n such that $f(x) = f(t_n(x))$, and *quasi-local* if $f(x)$ additionally depends on the value of $\mathrm{depth}(v^{-1}(x))$. If f is either local or quasi-local, let *loc(f)* denote minimum n such that

$$\begin{aligned} &t_{-n}(x) = t_{-n}(y) \ \wedge\ \mathrm{depth}(v^{-1}(x)) = \mathrm{depth}(v^{-1}(y)) \\ &\Rightarrow f(x) = f(y)\,. \end{aligned}$$

Theorem 7 *In $\mathcal{S}_\beta^\sigma(D)$, the rounding $U_{\rho,\sigma}$ is local, the rounding $U_{\rho,-\sigma}$ is quasi-local, and*

$$\mathrm{loc}(U_{\rho,\sigma}) = \mathrm{loc}(U_{\rho,-\sigma}) = \mathrm{depth}(v^{-1}\left(\frac{\rho+1}{2}\right))\,.$$

In $\mathcal{S}^s_\beta$, the rounding $S_{\rho,1}$ is local, the rounding $S_{\rho,-1}$ is quasi-local, and

$$\mathrm{loc}(S_{\rho,1}) = \mathrm{loc}(S_{\rho,-1}) = \mathrm{depth}(v^{-1}\left(\frac{\rho+1}{2}\right)).$$

Sketch of proof: Let $f := U_{\rho,\tau}$. Denote $\zeta := \frac{\rho+1}{2}$, and $\delta := \mathrm{depth}(v^{-1}(\zeta))$. Assume that δ is finite.

Let $x \in \mathbb{R}$, $r := v^{-1}(x)$, $k := v(T(r))$, and $z := x - k$. Then $k \in \mathbb{Z}$ and $-1 \le z \le 1$, so that $f(x) \in \{k-1, k, k+1\}$. The exact value of $f(x)$ depends on the relation between x and $k+\zeta$, when $z \ge 0$, and between x and $k-1+\zeta$, when $z \le 0$. This means that we have to compare z either with ζ or with $\zeta - 1$. Observe that $\mathrm{depth}(v^{-1}(y)) = \mathrm{depth}(v^{-1}(y-1))$, for every $y \in \mathbb{R}$.

Case 1: $\sigma = +1$. If z and ζ differ in their first δ digits then knowledge of $t_{-\delta}(z)$ suffices to determine the relationship between them. Otherwise, as $v^{-1}(\zeta)$ is minimal, $z \ge \zeta$. This suffices when $\tau = +1$. If we additionally know the depth of $v^{-1}(x)$ then we can also determine whether $z = \zeta$ or $z > \zeta$, which is necessary when $\tau = -1$.

Case 2: $\sigma = -1$. If z and ζ differ in their first δ digits then knowledge of $t_{-\delta}(z)$ suffices to determine the relationship between them. Otherwise, as $v^{-1}(\zeta)$ is maximal, $z \le \zeta$. This suffices when $\tau = -1$. If we additionally know the depth of $v^{-1}(x)$ then we can also determine whether $z = \zeta$ or $z < \zeta$, which is necessary when $\tau = +1$.

It is not difficult to construct examples which show that locality of $U_{\rho,\tau}$ is at least δ.

The proof of the second part of the theorem concerning standard systems is analogous. □

Again, the notion of locality and the results concerning it can be extended to other screens using Theorem 3.

Example 5 In $\mathcal{S}^s_3(\{-1,0,1\})$, we have

$$\begin{aligned} \mathrm{loc}(U_{0,\tau}) &= 0 \\ \mathrm{loc}(U_{2c-1,\tau}) &= 1, \quad \text{for } c \in \{1/6,\ 5/6\} \end{aligned}$$

When $\tau = \sigma$, the roundings are local, otherwise quasi-local.

In $\mathcal{S}^s_{10}$ we have

$$\begin{aligned} \mathrm{loc}(S_{1,1}) &= \mathrm{loc}(S_{-1,-1}) = 0 \\ \mathrm{loc}(S_{2c-1,\tau}) &= 1, \quad \text{for } c \in \{1/10,\ 2/10 \ldots, 9/10\} \end{aligned}$$

The roundings are local if $\tau = +1$, quasi-local if $\tau = -1$.

References

[1] K. Hwang, *Computer Arithmetic: Principles, Architecture, and Design*, John Wiley & Sons, New York, 1979.

[2] P-854: A proposed radix- and word-length-independent standard for floating-point arithmetic, *IEEE Micro*, August 1984, 86–99.

[3] D. E. Knuth, *The Art of Computer Programming, Vol. 2: Seminumerical Algorithms*, Addison-Wesley, Reading, Mass., 1981.

[4] U. W. Kulisch, W. L. Miranker, *Computer Arithmetic in Theory and Practice*, Academic Press, New York, 1981.

[5] D. W. Matula, Radix arithmetic: Digital algorithms for computer architecture, in *Applied Computation Theory: Analysis, Design, Modeling*, R. T. Yeh, Ed., Prentice-Hall, Englewood Cliffs, N.J., 1976, pp. 374–448.

[6] D.W. Matula, Basic digit sets for radix representation, J.ACM 29 (1982) 1131–1143.

[7] J.M. Yohe, Roundings in floating point arithmetic, *IEEE Trans. Computers* C-22 (1973) 577–586.