韓國數學教育學會誌<數學教育>
1983. 6. Vol. XXI, No. 3

Journal of the Korea Society of Mathematical Education
Jun. 1983. Vol. XXI, No. 3

A Study on some Properties of Floating-point Systems

By Myoung Lyoul Kim & Byung Chul Kim
Cheonbug National University, Cheonju, Korea

§1. Introduction

An agent which performs the algorithm of the given problem is called a *processor*. The processor may be a man or a machine. Let M be the set of machine representable numbers. Each number in set M can have only finite number of digits in it.

It is natural to postulate that the approximation of any number $x \notin M$ by a machine number $rd(x) \in M$ should satisfy $|x - rd(x)| \leq |x - g|$ for all $g \in M$.

Such a machine number approximation $rd(x)$ can be obtained in most case by *rounding*. [1] Generally, machine arithmetic number systems can be divided into the five categories: [2]

① Conventional Radix Number System ② Signed-Digit Number System

③ Residue Number System ④ Rational Number System ⑤ Logarithmic Number System

This paper shows the some properties of the number system on general-purpose digital computers by referencing [3] mainly.

However, the computer arithmetic is not dealt in this paper.

§2. Floating-point Systems

On computer memory, a given number may be stored in of two modes: *fixed-point and floating-point*. A floating-point hardware capabilities are now used for the vast majority of all numerical computation on computers because it has solved the problems in fixed-point hardware. Floating-point can represent numbers whose range is very large but it is handled with difficulty.

By normalizing the b-adic representation, we employ the symbol R_b to denote all real numbers, including zero:

$$R_b := \{O\} \cup \{x = Sm \cdot b^e \mid S \in \{+, -\},\ b \in N,\ b > 1,\ e \in Z,\ m = \sum_{i=1}^{\infty} x[i] b^{-i},$$

$$x[i] \in \{0, 1, \cdots, b-1\},\ x[1] \neq 0,\ x[i] \leq b-2 \text{ for infinitely many } i\}.$$

In general the element of R_b cannot be represented on computer. Only truncated version of these elements can be so represented.

Definition. A real number is called a ***normalized floating-point number*** if it is an element of the following sets $S_{b,l}$ or $S = S\ (b, l, e1, e2)$. These sets are called ***floating-point systems:***

(*i*) $S_{b,l} := \{x \in R_b \mid m = \sum_{i=1}^{l} x[i] b^{-i}\}$.

(*ii*) $S = S(b, l, e1, e2) := x \in Z_{b,l} \mid e1 \leq e \leq e2,\ \ e1, e, e2 \in Z\}$.

In order to have a unique representation of zero available in S, we put additionally that sgn $(0) = +$, mant$(0) = 0.00 \cdots\cdots 0$ (l zeros after the b-ary point), and $\exp(0) = e1$.

§3. Properties.

The following properties are easily shown from the above definition.

Property 1 S is symmetric, i.e., $0, 1 \in S \wedge -x \in S$ for all $x \in S$.

Property 2 $S_{b,l}$ and S are bounded, and we have $S \subset S_{b,l} \subset R_b = R$.

Especially, sup $S = B$ and inf $S = -B$, where $B = +0.(b-1)(b-1)\cdots\cdots(b-1)\cdot b^{e_2}$.

Property 3 (i) $S_{b,l}$ is countable, i.e., $\|S_{b,l}\| = \aleph_0$.

(ii) S is finite, and its element represents a rational number.

The representation is unique. $\|S\| = 2(b-1)b^{l-1}(e2-e1+1)+1$.

Property 4 The floating-point systems are not dense. The floating-point number in S are n uniformly distributed between $[-B, -L]$ and $[L, B]$, where $L = 0.10\cdots 0b^{e_1}$. Of course th floating-point systems are not continuum. They are discrete.

To continue, we introduce the following notation:

$R^* := R \cup \{-\infty\} \cup \{+\infty\}$, $S^* := S \cup \{-\infty\} \cup \{+\infty\}$.

Let $\bigcirc$ denote the rounding to the nearest floating-point numbers of S^*, i.e., $\bigcirc: R^* \to S^*$. An we define the floating-point operations as: For all $x, y \in S^*$ $x * y := \bigcirc(x * y)$, where $*$ in S^* is a operation corresponding to a usual operation$* \in \{+, -, \times, \div\}$ in $R*$.

We see easily that S^* is not closed under its operation $*$.

Property 5 $\bigcirc$ is not homomorphic.

For example, $b := 10$, $m := -0.9(0.1)\ 0.9$, $e \in \{-1, 0, 1\}$ and $x := 0.34$, $y := 0.54 \in R^*$. Then we get $\bigcirc x = 0.3$, $\bigcirc y = 0.5$; $(\bigcirc x) \oplus (\bigcirc y) = \bigcirc(\bigcirc x + \bigcirc y) = \bigcirc(0.8) = 0.8$, $\bigcirc(x+y) = \bigcirc(0.88) = 0.9$, i.e., $(\bigcirc x) \oplus (\bigcirc y) \neq \bigcirc(x+y)$.

Property 6 S^* is not associative for $\oplus$ and $\otimes$.

With $x := 0.7$, $y := 0.7$, $z ;= 0.9$, we obtain

$(x \oplus y) \oplus z = \bigcirc(1.4) \oplus 0.9 = 0.1 \cdot 10 \oplus 0.9 = \bigcirc 0.19 \cdot 10 = 0.2 \cdot 10$,

$x \oplus (y \oplus z) = 0.7 \oplus (\bigcirc 1.6) = 0.7 \oplus 0.2 \cdot 10 = \bigcirc 0.27 \cdot 10 = 0.3 \cdot 10$, i.e., $(x \oplus y) \oplus z \neq x \oplus (y \oplus z)$;

and $(x \otimes y) \otimes z = (\bigcirc 0.49) \otimes 0.9 = 0.5 \otimes 0.9 = \bigcirc 0.45 = 0.5 \times 10^0$,

$x \otimes (y \otimes z) = 0.7 \otimes (\bigcirc 0.63) = 0.7 \otimes 0.6 = \bigcirc 0.42 = 0.4 \times 10^0$, i.e., $(x \otimes y) \otimes z \neq x \otimes (y \otimes z)$.

Property 7 S^* is not distributive for $\otimes$ over $\oplus$.

For $x := 0.3$, $y := 0.7$, $z := 0.9$, we obtain $x \otimes (y \oplus z) = 0.3 \otimes (\bigcirc 1.6) = 0.3 \otimes 0.2 \cdot 10 = 0.6$, $x \otimes y \oplus x \otimes z = (\bigcirc 0.21) \oplus (\bigcirc 0.27) = 0.2 \oplus 0.3 = 0.5$, i.e., $x \otimes (y \oplus z) \neq x \otimes y \oplus x \oplus z$.

References

1. J. Stoer, R. Bulirsch: *Introdnction to Numerical Analysis;* Springer-Verlag New York, Inc., 198(
   p. 5.
2. Kai Hwang: *Computer Arithmetic (Principles, Architecture, and Design)*; John Wiley & Son: Inc., 1979, p. 4.
3. Ulrich W. Kulisch, Williard L. Miranker: *Computer Arithmetic in Theory and Practice;* Academ: Press, Inc., 1981, pp. 149-157.