

Approximating roots and reciprocal roots of binary floating-point numbers

Robin Leroy (eggrobin)

2018-03-30

In the following, $\mathbb{N} := [0, \infty[\cap \mathbb{Z}$. We define $\lfloor x \rfloor := x - \lceil x \rceil$, so that $\forall x \in \mathbb{R}, \lfloor x \rfloor \in [0, 1[$. $B \in \mathbb{N}$ is arbitrary.

Let $x > 0$. There are unique $F \in [0, 1[$, $K \in \mathbb{Z}$, such that $x = 2^K(1 + F)$; define

$$\text{定}x := B + K + F.$$

Let $X \in \mathbb{R}$; define

$$\text{浮}X := 2^{\lfloor X - B \rfloor}(1 + \lfloor X \rfloor).$$

Then $\text{定浮}X = X$, $\text{浮定}x = x$, $1 + \text{定}x = \text{定}(2x)$.

Let $n \in \mathbb{Z} \setminus \{0, 1\}$, $\gamma \in \mathbb{R}$. For $x > 0$, define

$${}^nr(x) := \text{浮}\left(C_{n,\gamma} + \frac{\text{定}x}{n}\right),$$

where

$$C_{n,\gamma} := \frac{(n-1)B - \gamma}{n}.$$

Consider the signed relative error $\epsilon(x)$ of ${}^nr(x)$ as an approximation of $\sqrt[n]{x}$. For $x = 2^K(1 + F)$, we have

$$\begin{aligned} \epsilon(x) &= \frac{{}^nr(x)}{\sqrt[n]{x}} - 1 \\ &= \frac{2^{\lfloor \frac{K+F-\gamma}{n} \rfloor} \left(1 + \left\lfloor \frac{K+F-\gamma}{n} \right\rfloor\right)}{2^{\frac{K}{n}} \sqrt[n]{1+F}} - 1 \\ &= 2^{\lfloor \frac{K+F-\gamma}{n} \rfloor - \frac{K}{n}} \frac{1 + \left\lfloor \frac{K+F-\gamma}{n} \right\rfloor}{\sqrt[n]{1+F}} - 1, \end{aligned}$$

which is invariant under addition of n to K , so that

$$\epsilon(x) = \epsilon(2^n x).$$

in other words,

$$\epsilon^{\text{浮}} : X \mapsto \epsilon(\text{浮}X)$$

is periodic with period n .

Consider the interval

$$I_{n,\gamma} := \begin{cases} [2^{\lfloor \gamma \rfloor}(1 + \lfloor \gamma \rfloor), 2^{\lfloor \gamma \rfloor+n}(1 + \lfloor \gamma \rfloor)[& n > 0, \\ [2^{\lfloor \gamma \rfloor+n}(1 + \lfloor \gamma \rfloor), 2^{\lfloor \gamma \rfloor}(1 + \lfloor \gamma \rfloor)[& \text{otherwise.} \end{cases}$$

Note that

$$\text{定}I_{n,\gamma} = \begin{cases} [B + \gamma, B + n + \gamma[& n > 0, \\ [B + n + \gamma, B + \gamma[& \text{otherwise,} \end{cases}$$

so that it covers one period of the relative error.

For $n = -2$, which has received particular attention, γ here corresponds to $2t - 1$ in [cite Robertson here], $2r_0 - 1$ in [cite Lomont here], -3σ in [cite McEniry here]. For $n = 3$, $C_{n,\gamma}$ corresponds to Kahan's C [cite Kahan here].

Let $x \in I_{n,\gamma}$. Then, with $F \in [0, 1[$, $K \in \mathbb{Z}$, such that $x = 2^K(1 + F)$,

$$nr(x) = 1 + \frac{K + F - \gamma}{n} = 1 + \frac{K + 2^{-K}x - 1 - \gamma}{n},$$

and $K \in [\lfloor \gamma \rfloor, \lfloor \gamma \rfloor + n - 1] \cap \mathbb{Z}$ if $n > 0$, $K \in [\lfloor \gamma \rfloor + n, \lfloor \gamma \rfloor] \cap \mathbb{Z}$ otherwise. For fixed K , i.e., for $x \in [2^K, 2^{K+1}[$, $\epsilon'(x) = 0$ at

$$x = 2^K \left(1 + \frac{K - \gamma}{n - 1} \right),$$

which is in $[2^K, 2^{K+1}[$ unless $K = \lfloor \gamma \rfloor$ and $n > 0$, or $K = \lfloor \gamma \rfloor + n$ and $n > 0$.

It follows that the maximum for $x > 0$ of $|\epsilon(x)|$ is the maximum of the absolute values of the following:

- the value $\epsilon(2^{\lfloor \gamma \rfloor}(1 + \lfloor \gamma \rfloor)) = \frac{1}{\sqrt[n]{2^{\lfloor \gamma \rfloor}(1 + \lfloor \gamma \rfloor)}} - 1$ at the endpoint of $I_{n,\gamma}$;
- the values at powers of two within $I_{n,\gamma}$, $\epsilon(2^K) = 2^{-\frac{K}{n}} \left(1 + \frac{K - \gamma}{n} \right) - 1$ for $K \in [\lfloor \gamma \rfloor + 1, \lfloor \gamma \rfloor + n] \cap \mathbb{Z}$ if $n > 0$, $K \in [\lfloor \gamma \rfloor + n + 1, \lfloor \gamma \rfloor] \cap \mathbb{Z}$ otherwise;
- the smooth extrema, $\epsilon \left(2^K \left(1 + \frac{K - \gamma}{n - 1} \right) \right)$ where $K \in [\lfloor \gamma \rfloor + 1, \lfloor \gamma \rfloor + n - 2] \cap \mathbb{Z}$ if $n > 0$ and $K \in [\lfloor \gamma \rfloor + n, \lfloor \gamma \rfloor] \cap \mathbb{Z}$ otherwise.