Approximating roots and reciprocal roots of binary floating-point numbers

Robin Leroy (eggrobin)

2018-03-29

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

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

$$x\sharp \coloneqq B + K + F.$$

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

$$X\flat \coloneqq 2^{\lfloor X - B \rfloor}(1 +]X[).$$

Then $X\flat\sharp = X$, $x\sharp\flat = x$, $x\sharp + 1 = (2x)\sharp$.

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

$$^n r(x) \coloneqq \left(C_{n,\gamma} + \frac{x\sharp}{n}\right)\flat,$$

where

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

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

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

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

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

in other words,

$$\epsilon_\flat : X \mapsto \epsilon(X\flat)$$

is periodic with period n.

Consider the interval

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

Note that

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

so that it covers one period of the relative error.

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

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

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