

A nearly correctly-rounded cube root

Robin Leroy (eggrobin)

2021-03-01

This document describes the error analysis of the real cube root function `Cbrt` implemented in `numerics/cbrt.cpp`.

Overview

The general approach to compute the cube root of $y > 0$ is the same as the one described in [KB01]:

1. integer arithmetic is used to get a an initial quick approximation q of $\sqrt[3]{y}$;
2. a root finding method is used to improve that that to an approximation ξ with a third of the precision;
3. ξ is rounded to a third of the precision, resulting in the rounded approximation x whose cube x^3 can be computed exactly;
4. a single high order iterate of a root finding method is used to get the final result.

Notation

We define the fractional part as $\text{frac } a := a - \lfloor a \rfloor \in [0, 1[$, regardless of the sign of a .

The quantities $p \in \mathbb{N}$ (precision in bits) and $\text{bias} \in \mathbb{N}$ are as defined in IEEE 754-2008.

We use capital letters fixed-point numbers involved in the computation, and $A > 0$ for the normal floating-point number $a > 0$ reinterpreted as a binary fixed-point number with t bits after the binary point¹,

$$\begin{aligned} A &:= \text{bias} + \lfloor \log_2 a \rfloor + \text{frac}(2^{-\lfloor \log_2 a \rfloor} a) \\ &= \text{bias} + \lfloor \log_2 a \rfloor + 2^{-\lfloor \log_2 a \rfloor} a - 1, \end{aligned}$$

and *vice versa*,

$$a := 2^{A - \text{bias}}(1 + \text{frac } A).$$

This corresponds to [KB01]'s $B + K + F$.

For both fixed- and floating-point numbers, given $\alpha \in \mathbb{R}$, we write $\llbracket \alpha \rrbracket$ for the nearest representable number (rounding ties to even). For fixed-point numbers, we write $\llbracket \alpha \rrbracket_0$ for directed rounding towards 0 to the fixed-point precision (as in division implemented with integer division).

Except in the section on rescaling, the input y and all intervening floating-point numbers are taken to be normal; the rescaling performed to avoid overflows also avoids subnormals.

¹The implementation uses integers (obtained by multiplying the fixed-point numbers by 2^{p-1}). For consistency with [KB01] we work with fixed-point numbers here. Since we do not multiply fixed point numbers together, the expressions are unchanged.

1 Quick approximation

The quick approximation q is computed using fixed-point arithmetic as

$$Q := C + \left\lfloor \frac{Y}{3} \right\rfloor_0,$$

where the fixed-point constant C is defined as²

$$C := \left\lfloor \frac{2 \text{bias} - \gamma}{3} \right\rfloor$$

for some $\gamma \in \mathbb{R}$.

Let $\varepsilon := \frac{q}{\sqrt[3]{y}} - 1$, so that $\sqrt[3]{y}(1 + \varepsilon) = q$; the relative error of q as an approximation of $\sqrt[3]{y}$ is $|\varepsilon|$. Considering Y , Q , q , and ε as functions of y , we have

$$\begin{aligned} Y(8y) &= Y(y) + 3, \\ Q(8y) &= Q(y) + 1, \\ q(8y) &= 2q(y), \\ \varepsilon(8y) &= \varepsilon(y), \end{aligned}$$

so that the properties of ε need only be studied on some interval of the form $[\eta, 8\eta]$.

Pick $\eta := 2^{\lfloor \gamma \rfloor}$, and $y \in [\eta, 8\eta] = [2^{\lfloor \gamma \rfloor}, 2^{\lfloor \gamma \rfloor+3}]$, so that $\log_2 y \in [\lfloor \gamma \rfloor, \lfloor \gamma \rfloor + 3]$. Let $k := \lfloor \log_2 y \rfloor - \lfloor \gamma \rfloor$; note that $k \in \{0, 1, 2\}$. Let $f := \text{frac}(2^{-\lfloor \log_2 y \rfloor} y) \in [0, 1]$. Up to at most 1.5 units in the last place from rounding,

$$\begin{aligned} Q \approx Q' &:= \text{bias} + \frac{\lfloor \log_2 y \rfloor}{3} + \frac{\text{frac}(2^{-\lfloor \log_2 y \rfloor} y) - \gamma}{3}, \\ &= \text{bias} + \frac{\lfloor \gamma \rfloor + k}{3} + \frac{f - \gamma}{3}, \\ &= \text{bias} + \frac{k + f - \text{frac} \gamma}{3}. \end{aligned}$$

Since $k \in [0, 2]$, the numerator $k + f - \text{frac} \gamma$ lies in $] -1, 3[$. Further, it is negative only if $k = 0$, so that

$$\begin{aligned} \lfloor Q' \rfloor &= \begin{cases} \text{bias} - 1 & \text{if } k = 0 \text{ and } \text{frac} \gamma > \text{frac}(2^{-\lfloor \gamma \rfloor} y), \\ \text{bias} & \text{otherwise,} \end{cases} \text{ and} \\ \text{frac} Q' &= \begin{cases} 1 + \frac{f - \text{frac} \gamma}{3} & \text{if } k = 0 \text{ and } \text{frac} \gamma > f, \\ \frac{k + f - \text{frac} \gamma}{3} & \text{otherwise.} \end{cases} \end{aligned}$$

Accordingly, for the quick approximation q , we have, again up to at most 1.5 units in the last place,

$$q \approx q' = \begin{cases} 1 + \frac{f - \text{frac} \gamma}{3} & \text{if } k = 0 \text{ and } \text{frac} \gamma > f, \\ 1 + \frac{k + f - \text{frac} \gamma}{3} & \text{otherwise,} \end{cases}$$

With $\sqrt[3]{y} = 2^{\frac{\lfloor \gamma \rfloor + k}{3}} \sqrt[3]{1 + f}$, we can define

$$\varepsilon' := \frac{q'}{\sqrt[3]{y}} - 1,$$

which we can express piecewise as a function of f and k . This gives us a bound on the relative error,

$$|\varepsilon| \leq |\varepsilon'| + 1.5 \cdot 2^{p-1} (1 + |\varepsilon'|).$$

The values $\gamma = 0.1009678$ and $\varepsilon < 3.2\%$ from [KBo1] may be recovered by choosing γ minimizing the maximum of $|\varepsilon'|$ over $y \in [\eta, 8\eta]$, or equivalently,

$$\gamma_{\text{Kahan}} := \underset{\gamma \in \mathbb{R}}{\text{argmin}} \max_{y \in [\eta, 8\eta]} |\varepsilon'| = \underset{\gamma \in \mathbb{R}}{\text{argmin}} \max_{(f, k)} |\varepsilon'|$$

²Note that there is a typo in the corresponding expression $C := (B - 0.1009678)/3$ in [KBo1]; a factor of 2 is missing on the bias term.

where the maximum is taken over $(f, k) \in [0, \text{frac } \gamma[\times \{0\} \cup [0, 1[\times \{1, 2\}$,

$$= \operatorname{argmin}_{\gamma \in \mathbb{R}} \max_{(f, k) \in \mathcal{E} \cup \mathcal{L}} |\varepsilon'|,$$

where $\mathcal{E} := \{(\text{frac } \gamma, 0)\} \cup \{(0, k) \mid k \in \{0, 1, 2\}\}$ is the set of the endpoints of the intervals whereon q' is piecewise affine, and $\mathcal{L} := \left\{\left(\frac{k - \text{frac } \gamma}{2}, k\right) \mid k \in \{1, 2\}\right\}$ are the local extrema.

The values are more precisely³

$$\gamma_{\text{Kahan}} \approx 0.10096\,78121\,55802\,88786\,36993\,42643\,55358\,06489\,88235\,75289$$

with

$$\max_y |\varepsilon'| \approx 0.03155\,46327\,73624\,80606\,11789\,73328\,17135\,58940\,02093\,40816,$$

leading to $C_{\text{Kahan}} = {}_{16}2\text{A9F } 7625\,3119\,\text{D328} \cdot 2^{-52}$ for IEEE 754-2008 binary64. However, as we will see in the next section, this value does not optimize the final error, so it is not the one that we use.

References

- [KB01] W. Kahan and D. Bindel. ‘Computing a Real Cube Root’. 2001 retypesetting by Bindel of a purported 1991 version by Kahan, at <https://cscclub.uwaterloo.ca/~pbarfuss/qbrt.pdf>. 21st Apr. 2001.

³These may be computed formally, but the expressions are unwieldy.