A nearly correctly-rounded cube root

Robin Leroy (eggrobin)

2021-03-07

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 $\operatorname{frac} a \coloneqq a - \lfloor a \rfloor \in [0, 1[$, regardless of the sign of a.

The quantities $p \in \mathbb{N}$ (precision in bits) and $bias \in \mathbb{N}$ 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 point[1],

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

and *vice versa*,

$$a \coloneqq 2^{\lfloor A \rfloor - bias}(1 + \operatorname{frac} A).$$

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

For both fixed- and floating-point numbers, given $\alpha \in \mathbb{R}$, we write $[\![\alpha]\!]$ for the nearest representable number (rounding ties to even). For fixed-point numbers, we write $[\![\alpha]\!]_0$ 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.

[1] The 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

$$Q \coloneqq C + \left[\!\left[\frac{Y}{3}\right]\!\right]_0,$$

where the fixed-point constant C is defined as[2]

$$C \coloneqq \left[\!\left[\frac{2\,bias - \gamma}{3}\right]\!\right]$$

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

Let $\varepsilon \coloneqq \frac{q}{\sqrt[3]{y}} - 1$, 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 $\eta \coloneqq 2^{\lfloor\gamma\rfloor}$, 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 $k \coloneqq \lfloor\log_2 y\rfloor - \lfloor\gamma\rfloor$; note that $k \in \{0,1,2\}$. Let $f \coloneqq \operatorname{frac}(2^{-\lfloor\log_2 y\rfloor} y) \in [0,1[$. Up to at most 1.5 units in the last place from rounding,

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

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

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

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

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

$$q \approx q' = \begin{cases} 1 + \frac{f - \operatorname{frac}\gamma}{6} & \text{if } k = 0 \text{ and } \operatorname{frac}\gamma > f, \\ 1 + \frac{k + f - \operatorname{frac}\gamma}{3} & \text{otherwise,} \end{cases}$$

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

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

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

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

The values $\gamma = 0.1009678$ and $\varepsilon < 3.2\%$ from [KB01] may be recovered by choosing γ minimizing the maximum of $|\varepsilon'|$ over $y \in [\eta, 8\eta[$, or equivalently.

$$\gamma_{\text{Kahan}} \coloneqq \operatorname*{argmin}_{\gamma\in\mathbb{R}} \max_{y \in [\eta, 8\eta[} |\varepsilon'| = \operatorname*{argmin}_{\gamma\in\mathbb{R}} \max_{(f,k)} |\varepsilon'|$$

[2] Note that there is a typo in the corresponding expression $C \coloneqq (B - 0.1009678)/3$ in [KB01]; a factor of 2 is missing on the bias term.

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

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

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

The values are more precisely[3]

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

with

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

leading to $C_{\text{Kahan}} = {}_{16}\text{2A9F 7625 3119 D328} \cdot 2^{-52}$ 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.

2 Getting to a third of the precision

We use a single iterate of Fantet de Lagny's method to compute ξ,

$$\xi \coloneqq \left[\!\!\left[q - \left[\!\!\left[\frac{[\![([\![[\![q^2]\!]q]\!] - y)q]\!]}{[\![2[\![[\![q^2]\!]q]\!] + y]\!]} \right]\!\!\right] \right]\!\!\right].$$

Note that the subtraction in the numerator is exact by Sterbenz's lemma. Let $\Delta \coloneqq \frac{\xi}{\sqrt[3]{y}} - 1$ and

$$\xi' = q' - \frac{({q'}^3 - y)q'}{2{q'}^3 + y}.$$

We have, up to rounding errors (TODO: bound those),

$$\Delta \approx \Delta' \coloneqq \frac{\xi'}{\sqrt[3]{y}} - 1.$$

With $q' = \sqrt[3]{y}(1+\varepsilon')$, we can express Δ' using the transformation of the relative error error by one step of Fantet de Lagny's method on the cube root,

$$\Delta' = \frac{2{\varepsilon'}^3 + {\varepsilon'}^4}{3 + 6\varepsilon' + 6{\varepsilon'}^2 + 2{\varepsilon'}^3}.$$

If q' is computed using $\gamma = \gamma_{\text{Kahan}}$, we get

$$\max_y |\Delta'| \approx 0.00002196,$$

$$\log_2 \max_y |\Delta'| \approx -15.47.$$

However, γ_{Kahan}, which minimizes $\max_y |\varepsilon|$, does not minimize $\max_y |\Delta'|$. This is because while Δ' is monotonic as a function of ε', it is not odd: positive errors are reduced more than negative errors are, so that the minimum is attained for a different value of γ. Specifically, we have

$$\begin{aligned}\gamma_{\text{L}} &\coloneqq \operatorname*{argmin}_{\gamma\in\mathbb{R}} \max_y |\Delta'| \\ &\approx 0.09918\,74615\,29855\,99525\,66149\,20761\,31234\,34720\,23067\,92759\end{aligned}$$

with

$$\max_y |\varepsilon'| \approx 0.03103\,20521\,29929\,93577\,08166\,75859\,02139\,33719\,41389\,93269,$$

but

$$\max_y |\Delta'| \approx 0.00002\,08686\,35536\,39593\,48770\,92008\,39844\,10254\,14831\,61229.$$

The corresponding fixed-point constant is $C_{\text{L}} \coloneqq {}_{16}\text{2A9F 7893 782D A1CE} \cdot 2^{-52}$ for binary64.

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

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://csclub.uwaterloo.ca/~pbarfuss/qbrt.pdf`. 21st Apr. 2001.