# A nearly correctly-rounded cube root

Robin Leroy (eggrobin)

TODO(egg): 2017-03-36

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

# On root finding methods studied by Lagny, Halley, Householder, *et alii*

We start with a historical overview of a family of root-finding methods.

In [Fan91a], Lagny first presents the iterates

$$a \mapsto \frac{1}{2}a + \sqrt{\frac{1}{4}a^2 + \frac{b}{3a}}, \tag{1}$$

hereafter the *irrational method*, and

$$a \mapsto a + \frac{ab}{3a^3 + b}, \tag{2}$$

the *rational method*, for the computation of the cube root $\sqrt[3]{a^3+b}$, mentioning the existence of similar methods for arbitrarily higher powers. In [Fan91b] the above methods are again given, with an outline of the general method for higher powers, and a mention of their applicability to finding roots of polynomials other than $z^p - r$.

That general method is given in detail in [Fan92, p. 19]. Modernizing the notation, the general rule is as follows for finding a root of the monic polynomial of degree $p \geq 2$

$$f(z) \coloneqq z^p - \sum_{k=1}^{p-1} c_k z^{p-k} \eqqcolon z^p - R(z)$$

with an initial approximation $a$.

Separate the binomial expansion of $\left(x + \frac{1}{2}a\right)^p$ into alternating sums of degree $p$ and $p-1$ in $z$,

$$S_1 \coloneqq \sum_{\substack{k=0\\2|k}}^{p} \binom{p}{k} x^{p-k} \left(\frac{1}{2}a\right)^k$$

$$S_2 \coloneqq \sum_{\substack{k=0\\2\nmid k}}^{p} \binom{p}{k} x^{p-k} \left(\frac{1}{2}a\right)^k$$

and consider the polynomials in $x$

$$P_p \coloneqq S_1 - \frac{1}{2}R\left(x + \frac{1}{2}a\right)$$

$$P_{p-1} \coloneqq S_2 - \frac{1}{2}R\left(x + \frac{1}{2}a\right).$$

Let $P_{n-1}$ be the remainder of the polynomial division[1] of $P_{n+1}$ by $P_n$.

The iterate is $x + \frac{1}{2}a$, where $x$ is a root of $P_2$ for the irrational method, and the root of $P_1$ for the rational method.

**Theorem.** *The iterate of Lagny's rational method for a polynomial $f$ of degree $p$ is*

$$z \mapsto z + (p-1)\frac{(1/f)^{(p-2)}(z)}{(1/f)^{(p-1)}(z)} \tag{3}$$

**Proof.** TODO(egg): The proof is left as an eggsercise to the reader. □

This iterate was described by Householder for an arbitrary analytic functions $f$ in [**Householder1970**] (see equation (14), wherein one should take $g \equiv 1$, and theorem 4.4.2 which mentions this special case; see also [**SebahGourdon2001**] which explicitly gives that form).

For $p = 2$ and $f$ an arbitrary polynomial, (3) is Newton's method, presented by Wallis in [**Wallis1685**].

For $p = 3$ and $f$ an arbitrary polynomial, it is Halley's rational method, given by him in [Hal94, pp. 142–143] in an effort to generalize[2] Lagny's (2). It is generally simply known as Halley's method, as the irrational method (which likewise generalizes Lagny's irrational method for $p = 3$, while retaining constant order as the degree changes) has comparatively fallen into obscurity; see [**ScavoThoo1995**].

Considering that generalizing from arbitrary polynomials to other functions does not change the method, we call the method given by the iterate (3)

- Newton's method when $p = 2$, for arbitrary $f$;
- Lagny's rational method when $p > 2$ and $f$ is a polynomial of degree $p$;
- Halley's (rational) method when $p = 3$ and $f$ is not a polynomial of degree $p$;
- the Lagny–Householder method otherwise.

Note that we avoid the name "Householder's method" in this last case, as it is often used to refer to a different third order iterate instead, see TODO CITATION.

[1] While the rest of the method is a straightforward translation, this step bears some explanations; its description in [Fan92] is

> De ces deux égalitez, ou priſes ſéparément, ou comparées enſemble ſelon la methode des problêmes plus que déterminez tirez en une valeur d'$x$ rationelle, ou ſimplement d'un degré commode.

It is assumed that the reader is familiar with this "method of more-than-determined problems". While the derivation of the root-finding method is described in painstaking detail in [Fan33], which outlines the treatment of overdetermined problems, it is perhaps this remark from [Fan97, p. 494] which lays it out most clearly:

> Il n'y a rien de nouveau à remarquer ſur les Problemes plus que déterminez du quatriéme degré. La Regle générale eſt d'égaler tout à zero, & de diviſer la plus haute équation par la moins élevée, ou l'également élevée l'une par l'autre, continuellement juſques à ce que l'on trouve le reſte ou le diviſeur le plus ſimple.

[2] Lagny's method is general, in that an iterate is given for any polynomial, albeit one whose order changes with the degree. However, while he refers to its results—and even corrects a misprint therein—, Halley did not have access to a copy of [Fan92],

> Has Regulas, cum nondum librum videram, ab amico communicatas habui

and it appears that said friend communicated only the formulæfor the cube and fifth root, as opposed to the general method and its proof, as Halley writes

> [...] *D. de Lagney* [...] qui cum totus fere sit in eliciendis Potestatum purarum radicibus, praefertim Cubicâ, pauca tantum eaque perplexa nec satis demonstrata de affectarum radicum extractione subjungit.

or, about Lagny's irrational method for the fifth root,

> Author autem nullibi inveniendi methodum ejusve demonstrationem concedit, etiamsi maxime desiderari videatur [...].

Being unaware of this generality, Halley sets out to generalize (1) and (2) to arbitrary polynomials, and does so by keeping the order constant.

# Computing a real cube root

We now turn to the computation 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 $\xi$ with a third of the precision;
3. $\xi$ 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 $\mathit{bias} \in \mathbb{N}$ 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[3],

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

and *vice versa*,

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

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 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[4]

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

[3] 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.

[4] 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.

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 $\varepsilon$ 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 $\varepsilon$ 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 $f := \operatorname{frac}(2^{-\lfloor \log_2 y \rfloor} y) \in [0, 1[$. Up to at most 1.5 units in the last place from rounding,

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

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

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

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 [KB01] may be recovered by choosing $\gamma$ minimizing the maximum of $|\varepsilon'|$ over $y \in [\eta, 8\eta[$, or equivalently.

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

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

The values are more precisely[5]

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

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

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 Lagny's method to compute $\xi$,

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

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 $\Delta'$ 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

$$\begin{aligned} \max_y|\Delta'| &\approx 0.00002196, \\ \log_2 \max_y|\Delta'| &\approx -15.47. \end{aligned}$$

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

## References

[Fan33] T. Fantet de Lagny. "Analyse Générale, ou Méthodes nouvelles pour résoudre les Problêmes de tous les Genres & de tous les Degrez à l'infini". In: *Recueil des Mémoires de l'Académie Royale des Sciences depuis 1666 jusqu'à 1699*. Ed. by C. Richer. Vol. XI. Par la compagnie des libraires, 1733. eprint: https://books.google.fr/books?id=KwP-kH6gmlEC.

[Fan91a] T. Fantet de Lagny. "Nouvelle methode de Mr. T. F. de Lagny pour l'approximation des Racines cubiques". In: *Le Journal des sçavans* 1691.17 (14th May 1691), pp. 200–203. eprint: https://gallica.bnf.fr/ark:/12148/bpt6k56538h/f202.double.

[Fan91b] T. Fantet de Lagny. *Méthode nouvelle, infiniment générale et infiniment abrégée, Pour l'Extraction des Racines quarrées, cubiques, &c. & pour l'Approximation des mêmes Racines à l'infini dans toutes sortes d'égalitez. Proposée à examiner aux Mathématiciens de l'Europe*. De l'Imprimerie d'Antoine Lambin, ruë S. Jacques, au Miroir, 1691. eprint: https://gallica.bnf.fr/ark:/12148/bpt6k1039787.

[Fan92] T. Fantet de Lagny. *Methodes nouvelles et abbregées pour l'extraction et l'approximation des racines. Et pour resoudre par le cercle et la ligne droite, plusieurs problêmes solides & sursolides ; comme la duplication du cube, l'invention de deux & de quatre moyennes proportionnelles, &c. dans toute la précision possible, & d'une maniere praticable. Avec une dissertation sur les methodes d'arithmetique & d'analyse ; où l'on établit des principes generaux pour en juger*. De l'Imprimerie de Jean Cusson, ruë saint Jacques, à l'Image de saint Jean Baptiste, 1692. eprint: https://nubis.univ-paris1.fr/ark:/15733/3415.

[Fan97] T. Fantet de Lagny. *Nouveaux elemens d'arithmetique et d'algebre, ou introduction aux mathematiques*. Chez Jean Jombert, prés des Augustins, à l'Image Nôtre-Dame, 1697. eprint: https://books.google.fr/books?id=IbTtzq_fixAC.

[Hal94] E. Halley. "Methodus Nova Accurata & Facilis Inveniendi Radices Æquationum quarumcumque generaliter, sine prævia Reductione". In: *Philosophical Transactions of the Royal Society* 18.210 (May 1694), pp. 136–148. DOI: 10.1098/rstl.1694.0029.

[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.