A correctly-rounded binary64 cube root

Robin Leroy (eggrobin)

REMOVE BEFORE FLIGHT 2021-04-36

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

On some abridged root-finding methods

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

In [Fan91a], Lagny first presents the iterations

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

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 $p \geq 2$

$$f(z) \coloneqq z^p + c_1 z^{p-1} + \cdots + c_{p-1}z + c_p =: z^p - R(z)$$

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,

$$S_1 \coloneqq \sum_{\substack{k=0 \\ 2|k}}^{p} \binom{p}{k} x^{p-k} \left(\tfrac{1}{2}a\right)^k \text{ and } S_2 \coloneqq \sum_{\substack{k=0 \\ 2\nmid k}}^{p} \binom{p}{k} x^{p-k} \left(\tfrac{1}{2}a\right)^k,$$

and consider the following polynomials, of degree p and $p - 1$ in x for almost all a:

$$E_p \coloneqq S_1 - \tfrac{1}{2}R\left(x + \tfrac{1}{2}a\right) \text{ and } E_{p-1} \coloneqq S_2 - \tfrac{1}{2}R\left(x + \tfrac{1}{2}a\right).$$

Let E_{n-1} be the remainder of the polynomial division[1] of E_{n+1} by E_n; its degree is $n - 1$ for almost all a. The iteration is $a \mapsto x + \frac{1}{2}a$, where x is a root of E_2 for the irrational method, and the root of E_1 for the rational method. Its order is p.

[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 prifes féparément, ou comparées enfemble felon la methode des problêmes plus que déterminez tirez en une valeur d'x rationelle, ou fimplement d'un degré commode.

It is assumed that the reader is familiar with this "comparison according to the method of more-than-determined problems". While the application 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 fur les Problemes plus que déterminez du quatriéme degré. La Regle générale eft d'égaler tout à zero, & de divifer la plus haute équation par la moins élevée, ou l'également élevée l'une par l'autre, continuellement jufques à ce que l'on trouve le refte ou le divifeur le plus fimple.

Modern calculus allows us to give a more straightfoward expression for the rational method than was available to Lagny; the proof of the following equivalence will be given at the end of this section.

Proposition. *The iteration of Lagny's rational method for a polynomial f of degree p is*

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

□

Names

The iteration (3) is a special case of the *Algorithmen* (A^{λ}_{ω}) defined by Schröder for an arbitrary polynomial f in [Sch70], equation (69) and p. 350; specifically, it is (A^{0}_{p-1}). As seen in the proof of the proposition, it is also a special case of Householder's equation (14) from [Hou70, p. 169], which generalizes it by substituting $\frac{f}{g}$ for f, and letting f be an arbitrary analytic function. The case $g \equiv 1$ is mentioned in theorem 4.4.2, and that expression is given explicitly in [SG01].

For $p = 2$ and f an arbitrary polynomial, (3) is Newton's method, presented by Wallis in [Wal85, p. 338].

For $p = 3$ and f an arbitrary polynomial, it is Halley's rational method, given in [Hal94, pp. 142–143] in an effort to generalize[2] Lagny's (2). It is usually 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 [ST95].

Considering, as remarked by [Sch70, p. 334], that a method can often be generalized from arbitrary polynomials or rational functions to arbitrary analytic functions, we call the iteration (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 3;
- the Lagny–Schröder method of order p otherwise.

We do not simply call this last case "Schröder's method", as it is only a special case of the methods defined in [Sch70], so that the expression would be ambiguous.

Note that we avoid the name "Householder's method" which appears in [SG01] and ulterior works (notably *MathWorld* and *Wikipedia*, both citing [SG01]), as it is variably used to refer to either (3) or to a method from a different family, namely φ_{p+1} from [Hou70, p. 168], equation (7), taking $\gamma_{p+1} \equiv 0$ in the resulting iteration; φ_3 is[3] the iteration given in section 3.0.3 of [SG01]. As mentioned by Householder, both of those were described by Schröder a century prior anyway: Householder's (7) is Schröder's (18) from [Sch70, p. 327].

---

[2] Lagny's method is general, in that an iteration 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 formulæ for 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, præfertim 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.

[3] We are grateful to Peter Barfuss for this observation.

Bibliographic note

Our foray into the history of these methods was prompted by finding the "historical background" section of [ST95] while looking for a reference for Halley's method: it is mentioned therein that this method, as applied to the cube root, is due to Lagny.

Searching for Lagny's work led us to the historical note [Can61], wherein a note by the editors Terquem and Gerono reads

> Naturellement, en mathématiques, séjour des propositions irréfragables, identiques en toute langue, en tout pays, ces rencontres ne peuvent manquer d'être assez fréquentes ; nulle part les plagiats *effectifs* sont si rares, et les plagiats *apparents* si communs que dans la science exacte par excellence ; mais les signaler est un devoir, un service rendu à l'histoire scientifique.

The editors then quote a letter by Prouhet, wherein he gives a reference to [Fan92].

Lagny's work proved far more extensive than we expected: besides the above root finding methods for arbitrary polynomials, it contains an error analysis, and even a discussion of the principles of performance analysis based on a decomposition in elementary operations on decimal digits.

We observed that the higher-order examples corresponded to the well-known higher order method attributed to Householder in [SG01]. Looking for properties of these methods in [Hou70] so as to prove that observation, we found that they are attributed to Schröder therein. As mentioned in the translator's note to the English translation [SS93] of [Sch70] by Stewart,

> A. S. Householder used to claim you could evaluate a paper on root finding by looking for a citation of Schröder's paper. If it was missing, the author had probably rediscovered something already known to Schröder.

It is quite possible that the irrational method could be expressed using Schröder's methods in one way or another, but while the result would probably be more convenient, it is unlikely to be something well-known, as irrational methods are far less popular nowadays—unjustifiedly so, as we shall see.

Prouhet's letter in [Can61] ends with these words:

> Tout cela est fort abrégé ; mais qui nous délivrera des méthodes abrégées, qui n'en finissent pas ?

Proof of the proposition

We now prove the above proposition, which, substituting the definition of Lagny's rational method, is that

$$x + \tfrac{1}{2}a = a + (p-1)\frac{(1/f)^{(p-2)}(a)}{(1/f)^{(p-1)}(a)} =: \psi(a)$$

if x is the root of E_1.

Proof. Let $E_p = d_0 x^p + \cdots + d_p$, $E_{p-1} = e_0 x^{p-1} + \cdots + e_{p-1}$. As shown in [Hou70, pp. 52–54], the polynomial remainders E_k are given up to a constant factor by [Hou70, p. 19] equation (23), *i.e.*, for some α,

$$\frac{E_k}{\alpha_k} = \det\begin{pmatrix}(E_p)_{p-1-k}\\(E_{p-1})_{p-k}\end{pmatrix},$$

where the expression on the right-hand side is the *bigradient* defined in [Hou68] (3.2) or [Hou70, p. 19] (20),

$$\frac{E_k}{\alpha_k} = \det \begin{pmatrix} d_0 & d_1 & d_2 & & \cdots & & d_{2(p-k)-3} & x^{p-k-2}E_p \\ & d_0 & d_1 & & \cdots & & d_{2(p-k)-4} & x^{p-k-3}E_p \\ & & \ddots & & & & \vdots & \\ & & & d_0 & d_1 & \cdots & d_{p-k-1} & x^0 E_p \\ \mathbf{0} & & & & e_0 & \cdots & e_{p-k-2} & x^0 E_{p-1} \\ & & & \iddots & & & \vdots & \\ & & e_0 & & \cdots & & e_{2(p-k)-4} & x^{p-k-2}E_p \\ & e_0 & e_1 & & \cdots & & e_{2(p-k)-4} & x^{p-k-2}E_p \\ e_0 & e_1 & e_2 & & \cdots & & e_{2(p-k)-3} & x^{p-k-1}E_p \end{pmatrix} =: \det \boldsymbol{E}_k,$$

where $d_n \coloneqq 0$ for $n > p$, and $e_n \coloneqq 0$ for $n > p-1$.

In particular, for $k = 1$,

$$\boldsymbol{E}_1 = \begin{pmatrix} d_0 & d_1 & d_2 & \cdots & d_{p-3} & d_{p-2} & d_{p-1} & d_p & & 0 & x^{p-3}E_p \\ & d_0 & d_1 & \cdots & d_{p-4} & d_{p-3} & d_{p-2} & d_{p-1} & d_p & & x^{p-4}E_p \\ & & \ddots & & & & & & & \ddots & \vdots \\ & & & d_0 & d_1 & d_2 & & \cdots & & d_{p-2} & x^0 E_p \\ \mathbf{0} & & & & e_0 & e_1 & & \cdots & & e_{p-3} & x^0 E_{p-1} \\ & & & \iddots & & & & & & \iddots & \vdots \\ & & e_0 & \cdots & e_{p-5} & e_{p-4} & e_{p-3} & e_{p-2} & e_{p-1} & & x^{p-5}E_{p-1} \\ & e_0 & e_1 & \cdots & e_{p-4} & e_{p-3} & e_{p-2} & e_{p-1} & & & x^{p-4}E_{p-1} \\ e_0 & e_1 & e_2 & \cdots & e_{p-3} & e_{p-2} & e_{p-1} & & & 0 & x^{p-3}E_{p-1} \end{pmatrix}.$$

Observe that, since the value of x used in the rational method is the root of E_1, for that value of x, $\det \boldsymbol{E}_1 = 0$, *i.e.*, $\boldsymbol{E}_1$ is singular.

Lemma. *The matrix $\boldsymbol{E}_1$ is singular if and only if $\boldsymbol{C}\left(x + \frac{1}{2}a\right)$ is singular, where*

$$\boldsymbol{C}(\Psi) \coloneqq \begin{pmatrix} \Psi - a & c_0 & & & 0 \\ -1 & c_1 & c_0 & & \\ 0 & c_2 & & \ddots & \\ \vdots & \vdots & \ddots & & c_0 \\ 0 & c_{p-1} & \cdots & c_2 & c_1 \end{pmatrix}$$

and $c_0 \coloneqq 1$.

TODO(egg): Prove the lemma.

Proof (Left as an exercise to the reviewer). Observe that:

— since $\det \boldsymbol{E}_1$ is a polynomial of degree 1, all terms divisible by x^2 must cancel out in the Laplace expansion of that determinant on the last column, the determinant is equal to

$$\pm\delta_1 x E_p \mp \delta_2 E_p \pm \delta_3 E_{p-1} \mp \delta_4 x E_{p-1},$$

where the δ_i are determinants of real matrices;

— by the same reasoning, only the linear and constant terms of the polynomials remain in the above expression, which simplifies to

$$\pm\delta_1 d_p x \mp \delta_2 (d_{p-1}x + d_p) \pm \delta_3 (e_{p-2}x + e_{p-1}) \mp \delta_4 e_{p-1} x;$$

— $E_p - E_{p-1} = S_1 - S_2$, which, by Lagny's *theoreme fondamental* [Fan92, p. 17], is $\left(x - \frac{1}{2}a\right)^p$.

The proof is a calculation. □

The proposition follows from the lemma and theorem 4.4.2 from [Hou70, p. 169]: $\psi(a)$ is Householder's (14) with $g \equiv 1$; for that value of g, theorem 4.4.2 states that (14) is the solution of (12) from the same page, which is $\det \boldsymbol{C}(\psi(a)) = 0$. By the lemma, for the value of x in Lagny's rational method, $x + \frac{1}{2}a$ solves that equation. □

A faithfully-rounded cube root

We now turn to the computation in `numerics/cbrt.cpp`.

Overview

The general approach to compute a faithfully-rounded 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 iteration of a root finding method is used to get the faithfully-rounded result r_0.

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 defined as in IEEE 754-2008.

We use capital Latin 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 $p - 1$ bits after the binary point[4],

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

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). We write the unit roundoff $u \coloneqq 2^{-p}$, and, after [Hig02, p. 63], $\gamma_n \coloneqq \frac{nu}{1-nu}$.

To quote [Tre97], "If rounding errors vanished, 95% of numerical analysis would remain". While we keep track of rounding errors throughout, they are of very little importance until the last step; when it is convenient to solely study the truncation error, we work with ideal quantities affected with a prime, which correspond to their primeless counterparts by removal of all intervening roundings.

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

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

[4] 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.

where the fixed-point constant C is defined as[5]

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

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

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

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

so that the properties of ε_q 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 3 half-units in the last place from rounding (2 from the directed rounding of the division by three and 1 from the definition of C), we have

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

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 3 half-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

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

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

$$|\varepsilon|_q \le |\varepsilon'_q|(1 + 3u).$$

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

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

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

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

[5] 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 $\mathcal{E} \coloneqq \{(\operatorname{frac} \Gamma, 0)\} \cup \{(0, k) \mid k \in \{0, 1, 2\}\}$ 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. We get more precisely[6]

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

with $\max_y |\varepsilon'_q| \approx 0.03155$, yielding the constant

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

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.

2 Getting to a third of the precision

We now consider multiple methods for the refinement of q to ξ. The rounding error in this step being both negligible and tedious to bound, its analysis is relegated to appendix B. Here we will study only the truncation error, and thus work only with the primed quantities.

Lagny's rational method

One way to compute ξ is Lagny's rational method,

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

so that, up to rounding errors[7],

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

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

$$\varepsilon'_\xi = \frac{2{\varepsilon'_q}^3 + {\varepsilon'_q}^4}{3 + 6\varepsilon'_q + 6{\varepsilon'_q}^2 + 2{\varepsilon'_q}^3}.$$

If q' is computed using $\Gamma = \Gamma_{\text{Kahan}}$, we get $\max_y |\varepsilon'_\xi| \approx 0.00002196$, $\log_2 \max_y |\varepsilon'_\xi| \approx -15.47$. However, Γ_{Kahan}, which minimizes $\max_y |\varepsilon_q|$, does not minimize $\max_y |\varepsilon'_\xi|$. This is because while ε'_ξ is monotonic as a function of ε'_q, 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

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

with $\max_y |\varepsilon'_q| \approx 0.0032025$, but $\max_y |\varepsilon'_\xi| \approx 0.00002086$, $\log_2 \max_y |\varepsilon'_\xi| \approx -15.54$. The corresponding fixed-point constant is

$$C_{\text{L}} \coloneqq {}_{16}\text{2A9F}\,7893\,782\text{D}\,\text{A1CE} \cdot 2^{-52}$$

for binary64.

While it is close to the 16 bits to which we will round in the next step, this error is still larger, and in any case is not comparatively negligible. As a result, it significantly contributes to the final error above $1u$, *i.e.*, to misrounding. Lagny's lesser-known irrational method provides us with a way to improve it.

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

[7] The .

Lagny's irrational method

As written in (1), Lagny's irrational method

$$\xi' \coloneqq \tfrac{1}{2}q' + \sqrt{\tfrac{1}{4}q'^2 + \frac{y - q'^3}{3q'}}$$

seems prohibitively computationally expensive in comparison to the rational one: it adds a square root on the critical path, dependent on the result of a division. However, evaluating it as

$$\xi = \tfrac{1}{2}q' + \frac{1/\sqrt{12}}{q'}\sqrt{4yq' - q'^4}$$

yields performance comparable to the rational method.

Canon optimization of Lagny's irrational method

TODO(egg): ε'_ξ accounts for more than half of the excess above correct rounding in the final bound (for $\varepsilon_\xi = 0$, the bound becomes $1.0001832u$). Consider using either a higher order iteration or Lagny's irrational formula. Also consider optimizing the constants in the method, as described in [Can18], reproduced below with the author's permission.

> A trick I've used for years and should write up: you can apply optimization to the iteration, not just the starting guess: $x' = xp(x)$, select $p(x)$ to be minimax error on bounded initial error in x. This yields a nice family of tunable approximations.
>
> Everyone else seems to worry about starting estimate, but use standard iterations, which is appropriate for arbitrary precision, but silly with a fixed precision target.
>
> Note that as p gets to be high-order, it converges quickly to the Taylor series for the correction, but there's a nice space with cheap initial approximations and order 2–5 or so, because we can evaluate these polynomials with lower latency [than] serially-dependent iterations.

3 Rounded approximation

The number x is obtained from ξ by zeroing all but the most significant $\left\lfloor \frac{p}{3} \right\rfloor$ bits of its significand. The resulting relative error $\left|\frac{x}{\xi} - 1\right|$ is greatest when the zeroed bits are all 1 and the remaining bits (except for the leading 1) are all 0; this is the case when the significand of ξ is $1 + 2^{-\left\lfloor \frac{p}{3} \right\rfloor + 1} - 2^{1-p}$, in which case that of x is 1.

The relative error arising from $\xi \mapsto x$ is thus at most

$$\left|\frac{x}{\xi} - 1\right| \le 1 - \frac{1}{1 + 2^{-\left\lfloor \frac{p}{3} \right\rfloor + 1} - 2^{1-p}} < 2^{-\left\lfloor \frac{p}{3} \right\rfloor + 1},$$

2^{-16} in binary64.

4 High order iteration

We use one iteration of the Lagny–Schröder method of order 6:

$$r \coloneqq \left[\!\!\left[x - \left[\!\!\left[\frac{\left[\!\!\left[[\![x(x^3 - y)]\!] \left[\!\!\left[\left[\!\!\left[[\![5x^3]\!] + [\![17y]\!] \right]\!\!\right] [\![x^3]\!] + [\![5[\![y^2]\!]]\!] \right]\!\!\right] \right]\!\!\right]}{\left[\!\!\left[\left[\!\!\left[[\![7x^3]\!] + [\![42y]\!] \right]\!\!\right] [\![x^6]\!] + \left[\!\!\left[[\![30x^3]\!] + 2y \right]\!\!\right] [\![y^2]\!] \right]\!\!\right]} \right]\!\!\right] \right]\!\!\right],$$

where x^3 is exact thanks to the trailing 0s of x, $[\![x^6]\!]$ is correctly-rounded because it is computed as the square of x^3, and $x^3 - y$ is exact by Sterbenz's lemma.

In infinite precision, this method is of such high order that if x has a relative error below 2^{-14}, which is the case here, the relative error of the result is less than 2^{-100}; we will not seek to bound the truncation error more closely, nor to tweak Γ nor the method itself to optimize it: as we will see, it is dominated by rounding.

Thanks to the exact cube and exact difference, the rounding analysis is straightforward; all remaining sums, being of positive terms, have a relative error bounded by the largest of those of their terms, so that we get no more than γ_4 on the second factor of the numerator, γ_6 on the whole numerator, and similarly γ_5 on the denominator, overall $\frac{1+\gamma_7}{1-\gamma_5} - 1 < \frac{\gamma_{12}}{1-\gamma_5}$ on the correction term.

Note however that in the rational method for the cube root there is little room for optimization without altering performance; the only constant is 2, and changing that turns an addition into a multiplication.

TODO(egg): consider order 5; the truncation error is dominated by the rounding error here. It is not immediately clear that order 5 would incur fewer roundings though, for the same reason that it has the same performance.

Since the relative error of x is at most $(1 + \varepsilon_\xi)(1 + 2^{-16}) - 1$, the rounding-free correction term is no larger than $((1 + \varepsilon_\xi)(1 + 2^{-16}) - 1)\sqrt[3]{y}$, so that the absolute error of the computed correction term is bounded by

$$\frac{\gamma_{12}}{1 - \gamma_5}((1 + \varepsilon_\xi)(1 + 2^{-16}) - 1)\sqrt[3]{y}.$$

This gives us our bound for the relative error of the result,

$$\left|\frac{r}{\sqrt[3]{y}} - 1\right| < \left(1 + \frac{\gamma_{12}}{1 - \gamma_5}((1 + \varepsilon_\xi)(1 + 2^{-16}) - 1)\right)(1 + u) - 1.$$

For binary64, this bound is $\left|\frac{r}{\sqrt[3]{y}} - 1\right| < 1.0004336u$.

Correct rounding

A FMA

B Rounding error analysis for the second step

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

Note that the subtraction in the numerator is exact by Sterbenz's lemma [Ste74, p. 138, theorem 4.3.1].

Let $\varepsilon_\xi \coloneqq \frac{\xi}{\sqrt[3]{y}} - 1$ and

It is fairly clear that ε'_ξ dominates the rounding error in the approximation $\varepsilon_\xi \approx \varepsilon'_\xi$; for the sake of completeness we quantify this.

Since we have a cancellation in the expression for ξ, a little bit of care is needed to bound those errors. As mentioned above, q approximates q' to a relative error of at most $3u < \gamma_3$. The sum in the denominator

$$d \coloneqq \left[\!\!\left[2\left[\!\!\left[\left[\!\!\left[q^2 \right]\!\!\right] q \right]\!\!\right] + y \right]\!\!\right]$$

has positive terms, so its relative error with respect to $d' \coloneqq 2{q'}^3 + y$ is readily bounded,

$$\frac{d}{d'} - 1 < \gamma_{3\cdot 3+3} = \gamma_{12}.$$

The cancelling difference $b \coloneqq \left[\!\!\left[\left[\!\!\left[q^2 \right]\!\!\right] q \right]\!\!\right] - y$ differs from $b' \coloneqq {q'}^3 - y$ by at most

$$\delta \coloneqq {q'}^3 \gamma_{3\cdot 3+2} = {q'}^3 \gamma_{11},$$

and the numerator $[\![bq]\!]$ from $b'q'$ by at most

$$((b' + \delta)q'(1 + \gamma_3))(1 + \gamma) - b'q' = (1 + \gamma_4)\delta q' + \gamma_4 b'q'.$$

We can bound the absolute error of the correction term as

$$\begin{aligned}\left| \left[\!\!\left[\frac{[\![bq]\!]}{d} \right]\!\!\right] - \frac{b'q'}{d'} \right| < \frac{b'q' + (1 + \gamma_4)\delta q' + \gamma_4 b'q'}{d'(1 - \gamma_{12})} - \frac{b'q'}{d'} &= \frac{(1 + \gamma_4)\delta q' + (\gamma_4 + \gamma_{12})b'q'}{d'(1 - \gamma_{12})} \\ &< q' \frac{(1 + \gamma_4)\delta + \gamma_{16} b'}{d'(1 - \gamma_{12})},\end{aligned}$$

and that of ξ as

$$|\xi - \xi'| < q' \left(\gamma_3 + \frac{(1 + \gamma_4)\delta + \gamma_{16} b'}{d'(1 - \gamma_{12})} \right).$$

Substituting the truncation errors and the definition of δ, we can bound the relative error arising from rounding:

$$\left| \frac{\xi}{\xi'} - 1 \right| < \frac{1 + \varepsilon'_q}{1 + \varepsilon'_\xi} \left(\gamma_3 + \frac{(1 + \gamma_4)\gamma_{11}(1 + \varepsilon'_q)^3 + \gamma_{16}((1 + \varepsilon'_q)^3 - 1)}{((1 + \varepsilon'_q)^3 + 1) + (1 - \gamma_{12})} \right).$$

Linearizing, this bound is $\left(6 + \frac{2}{3} + \mathcal{O}(\varepsilon'_q + \varepsilon'_\xi)\right)u + \mathcal{O}(u^2)$. More palatably, for either choice of γ, we have

$$\left| \frac{\xi}{\xi'} - 1 \right| < 8u$$

provided that $p \geq 12$.

References

[Can18] S. Canon. “A trick I’ve used for years and should write up”. Tweets at https://twitter.com/stephentyrone/status/1016283784067665920 *sqq.*, with a note at https://twitter.com/stephentyrone/status/1016328842296864770. 9th July 2018.

[Can61] M. Cantor. “Note historique sur l’extraction abrégée de la racine carrée”. In: *Nouvelles annales de mathématiques. Journal des candidats aux écoles polytechnique et normale* 1.20 (1861). Ed. by O. Terquem and C.-C. Gerono, pp. 46–47.
eprint: http://www.numdam.org/item/?id=NAM_1861_1_20__46_1.

[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.

[Hal09] E. Halley. “A new, exact, and easy Method of finding the Roots of any Equations generally, and that without any previous Reduction”. In: *The Philosophical Transactions of the Royal Society of London, from their commencement, in 1665, to the year 1800; Abridged, with notes and biographic illustrations.* Ed. by C. Hutton, G. Shaw and R. Pearson. Vol. III from 1683 to 1694. Translated from the Latin [Hal94]. 1809, pp. 640–649.

[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. English translation: [Hal09].

[Hig02] N. J. Higham. *Accuracy and Stability of Numerical Algorithms*. Society for Industrial and Applied Mathematics, 2002.

[Hou68] A. S. Householder. “Bigradients and the Problem of Routh and Hurwitz”. In: *SIAM Review* 10.1 (Jan. 1968), pp. 56–66.
DOI: 10.1137/1010003.

[Hou70] A. S. Householder. *The Numerical Treatment of a Single Nonlinear Equation.* International Series in Pure and Applied Mathematics. McGraw-Hill, 1970.

[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.

[Sch70] E. Schröder. "Ueber unendlich viele Algorithmen zur Auflösung der Gleichungen". In: *Mathematische Annalen* 2 (June 1870), pp. 317–365. DOI: `10.1007/BF01444024`. English translation: [SS93].

[SG01] P. Sebah and X. Gourdon. "Newton's method and high order iterations". In: *Numbers, constants and computation.* 3rd Oct. 2001. eprint: `http://numbers.computation.free.fr/Constants/Algorithms/newton.html`.

[SS93] E. Schröder and G. W. Stewart. *On Infinitely Many Algorithms for Solving Equations.* Tech. rep. UMIACS-TR-92-121. Translated from the German [Sch70]. Institute for Advanced Computer Studies, University of Maryland, College Park, Jan. 1993. eprint: `http://hdl.handle.net/1903/577`.

[ST95] T. R. Scavo and J. B. Thoo. "On the Geometry of Halley's Method". In: *The American Mathematical Monthly* 102.5 (May 1995), pp. 417–426. DOI: `10.2307/2975033`.

[Ste74] P. H. Sterbenz. *Floating-point computation.* Prentice-Hall, 1974.

[Tre97] L. N. Trefethen. *Maxims about numerical mathematics, science, computers, and life on Earth.* Maxims. Cornell University, 1997. eprint: `https://people.maths.ox.ac.uk/trefethen/maxims.html`.

[Wal85] J. Wallis. *A Treatise of Algebra, both Historical and Practical. Shewing, the Original, Progress, and Advancement thereof, from time to time; and by what Steps it hath attained to the Heighth at which now it is.* Richard Davis, 1685. eprint: `https://books.google.fr/books?id=TXpmAAAAcAAJ`.