A correctly rounded binary64 cube root

Robin Leroy (eggrobin)

REMOVE BEFORE FLIGHT 2021-04-36

This document describes the computations in `numerics/cbrt.cpp`.

On some abridged root-finding methods

We recall two families of root-finding methods from the late 17th century.

In [Fan91a], Thomas Fantet de Lagny first presents the iterations

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

hereafter the *(quadratic) 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 $p \geq 2$

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

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,

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

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

$$E_p \coloneqq S_1 - \tfrac{1}{2}R\left(x + \tfrac{1}{2}a\right) \text{ and } E_{p-1} \coloneqq S_2 - \tfrac{1}{2}R\left(x + \tfrac{1}{2}a\right). \tag{3}$$

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 in the quadratic irrational method, and the root of E_1 in 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.

Multiplicity of the irrational methods

Lagny does not require that the polynomial division be carried out all the way to E_2, merely until one gets *une valeur d'x* [...] *d'un degré commode*, by which he likely means one that is constructible. When f is a cubic, he uses the term *formule irrationnelle* for $x + \frac{1}{2}a$ where x is a root of E_2, but when it comes to computing the fifth root, the same term is used to refer to the case where x is a root of E_4. In order to avoid confusion, we use the term *quadratic irrational method* when x is a root of E_2, and we call the irrational formula for $\sqrt[5]{a^5 + b}$ from [Fan92, p. 43][2]

$$a \mapsto \frac{1}{2}a + \sqrt{\sqrt{\frac{1}{4}a^4 + \frac{b}{5a}} - \frac{1}{4}a^2}$$

Lagny's *quartic irrational method* for the fifth root; the quadratic irrational method for the same fifth root would be[3]

$$a \mapsto \frac{a\left(7b - \sqrt{100a^{10} + 100a^5b - 7b^2}\right)}{4b - 10a^5}.$$

More generally, we call $a \mapsto x + \frac{1}{2}a$ *Lagny's method of degree d* when x is a root of E_d. Note however that when $p = 3$, *i.e.*, when finding a root of a cubic, Lagny's only irrational method is the quadratic one; we can thus unambiguously refer to (1) as *Lagny's irrational method for the cube root.*

Generalization to arbitrary functions

Lagny's method of degree d and convergence order p may be generalized to functions f other than polynomials of degree p, by defining E_p and E_{p-1} in terms of Taylor polynomials for $f\left(x + \frac{1}{2}a\right)$ around $x = \frac{1}{2}a$:

$$E_p \coloneqq T_p - \frac{1}{2}T_{p-1} \text{ and } E_{p-1} \coloneqq \frac{1}{2}T_{p-1}, \tag{4}$$

where

$$T_n \coloneqq \sum_{k=0}^{n} \frac{f^{(n)}(a)}{k!}\left(x - \frac{1}{2}a\right)^k.$$

The rest of the method remains unchanged; the iteration is given by $a \mapsto x + \frac{1}{2}a$ for a root x of E_d. When f is a monic polynomial of degree p, the definitions (4) are equivalent to (3), so that we recover Lagny's method. When f is not a polynomial of degree p, we thus call the method defined by (4) the *generalized Lagny method of degree d and order p*; we use the terms “rational”, “quadratic irrational”, etc. for $d = 1$, $d = 2$, etc. respectively.

Note that while the E_n defined in this fashion may not have degree n if the higher derivatives of f vanish, *e.g.*, for a polynomial of degree less than p, the calculation may be carried out formally for an arbitrary f, and the offending function substituted in the result, taking limits as needed to remove singularities; the generalized methods of high order can thus be applied to polynomials of low degree.

These methods may equivalently be constructed using the Maclaurin series for $f(a + \varDelta)$ in the correction term $\varDelta$. Let

$$M_n \coloneqq \sum_{k=0}^{n} \frac{f^{(n)}(a)}{k!}\varDelta^k,$$

and consider the polynomials $\tilde{E}_p \coloneqq M_p$ and $\tilde{E}_{p-1} \coloneqq M_{p-1}$ of degree p and $p-1$ in $\varDelta$. Let $\tilde{E}_{n-1}$ be the remainder of the polynomial division of $\tilde{E}_{n+1}$ by $\tilde{E}_n$. The iteration is then $a \mapsto a + \varDelta$, where $\varDelta$ is a root of $\tilde{E}_d$.

[2] The formula has a misprint in [Fan92, p. 43], $-\frac{1}{2}a^2$ instead of $-\frac{1}{4}a^2$ under the radical. Halley remarks on it and gives the corrected formula in [Hal94, pp. 137, 140]. The misprint remains forty years later in [Fan33, p. 440 misnumbered 340]. Bateman writes in [**Bateman1938**] “we must not infer that [these expressions] are not correct simply because they differ from Halley's expression”, but with Lagny's construction, which was seemingly unknown to Bateman, the error is plain.

[3] Both are of order 5, but the reader who wishes to compute a fifth root should note that leading term of the error of the quartic method is $\frac{2}{7}$ of that of the quadratic.

Names of the irrational methods

Halley likewise generalized Lagny's (quadratic) irrational method for the cubic to arbitrary[4] polynomials, retaining cubic convergence; when f is not a polynomial of degree 3, we thus call the generalized Lagny quadratic irrational method of order 3 *Halley's irrational method.* This method was given in terms of derivatives by Bateman in [**Bateman1938**]:

$$a \mapsto a - \frac{f'(a)}{f''(a)} + \frac{\sqrt{{f'}^2(a) - 2f(a)f''(a)}}{f''(a)}.$$

Names of the rational method

Both special cases and generalizations of Lagny's rational method have been discovered multiple times and extensively studied; constructions that take advantage of modern calculus allow 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 the generalized Lagny rational method of order p for a root of the function f is*

$$a \mapsto a + (p-1)\frac{(1/f)^{(p-2)}(a)}{(1/f)^{(p-1)}(a)}. \tag{5}$$

□

The iteration (5) is a special case of the *Algorithmen* (A_ω^λ) defined by Schröder for an arbitrary polynomial f in [Sch70, pp. 349 sq.], equation (69); specifically, it is (A_{p-1}^0). 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 f/g for f. 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, (5) 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 Lagny's (2). It is usually simply known as Halley's method, as his aforementioned irrational method 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 (5)

- 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 rational method of order p otherwise.

[4] 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, præſertim Cubicâ, pauca tantum eaque perplexa nec ſatis demonſtrata de affectarum radicum extractione ſubjungit.

or, about the quartic irrational method for the fifth root, whereon Lagny does not elaborate as it is a direct application of the general method,

> 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.

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 (5) 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[5] 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].

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 into elementary operations on—and writing of—decimal digits [Fan92, pp. 5–9], with a remark on applicability to bases other than ten: a 17th century MIX.

Observing that the higher-order examples correspond to the well-known higher order method attributed to Householder in [SG01], we looked for its properties in [Hou70] so as to prove that observation, and found that Householder attributes them to Schröder. As mentioned in the translator’s note by Stewart in [SS93],

> 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 possible that the irrational methods could be expressed using Schröder’s results in one way or another, although most of his methods seem to be rational; in any case, such a formulation is unlikely to be something well-known, as irrational methods are far less popular nowadays—unjustifiedly so, as we shall see.

Our generalization of Lagny’s irrational methods to arbitrary f, which, in the polynomial case, decouples the degree of f from the convergence order, was inspired by Gander’s rephrasing in [**Gander1985**] of Halley’s construction from [Hal94], wherein the correction term of Halley’s irrational method is defined as a root of M_2.

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

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

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

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),

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

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

In particular, for $k = 1$,

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

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*

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

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 of the cube root of in `numerics/cbrt.cpp`.

Overview

Our general approach to computing a faithfully rounded cube root of $y > 0$ is 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 floating-point format used throughout is binary64; the quantities $p \in \mathbb{N}$ (precision in bits) and $bias \in \mathbb{N}$ are defined as in IEEE 754-2008, $p = 53$ and $bias = 1023$. Some of the individual methods discussed may be of general use; we thus give all inexact constants used in such methods rounded to thirty-five decimal places and twenty-nine hexadecimal places, which suffices for decimal128, binary128, and all smaller formats. A superscript sign after the last digit serves as the sticky bit[6]: the unrounded quantity is in excess of the rounded one if the sign is +, and in default if it is −.

We use capital Latin letters for fixed-point numbers involved in the computation, and $A > 0$ for the normal floating-point number $a > 0$ reinterpreted as a binary fixed-point[7] number with $p - 1$ bits after the binary point,

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

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:

[6] We learned of this practice from Steve Canon, who found it in a re-edition of [**Bruhns1870**]; there it is only present on the digit 5, to guard against double-rounding to the nearest decimal place. As mentioned in Hoüel's foreword to [**Schrön1873**], this practice, originally seen as a way to convey another bit of precision rather than a way to ensure correct rounding, dates back to at least 1827; see [**Babbage1827**], 8th rule. Like Babbage and Schrön, we give this bit regardless of the last digit; this allows for directed rounding.

[7] 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.

- $[\![\alpha]\!]$ for the nearest representable number, rounding ties to even: IEEE 754-2008 rounding-direction attribute roundTiesToEven;
- $[\![\alpha]\!]_+$ for the smallest positive representable number $\geq \alpha$: rounding-direction attribute roundTowardPositive;
- $[\![\alpha]\!]_-$ for the largest negative representable number $\leq \alpha$: rounding-direction attribute roundTowardNegative;
- $[\![\alpha]\!]_0$ for the representable number with the same sign as α and the largest magnitude $\leq |\alpha|$: rounding-direction attribute roundTowardZero.

We write the unit roundoff $u \coloneqq 2^{-p}$ (for rounding to nearest), and, after [Hig02, p. 63], $\gamma_n \coloneqq \frac{nu}{1-nu}$. We discuss other rounding modes in appendix D.

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. We work only with correctly rounded addition, subtraction, multiplication, division, and square root; FMA is treated separately in appendix A.

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[8]

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

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

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

[8]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.

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,

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

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 $(f, k) \in [0, \operatorname{frac} \Gamma[\times \{0\} \cup [0, 1[\times \{1, 2\}$,

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

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[9]

$$\begin{aligned} \Gamma_{\text{Kahan}} &\approx 0.10096\,78121\,55802\,88786\,36993\,42643\,55358\,1 \\ &\approx {}_{16}\text{E99.70D0 DEAD BEEF} \end{aligned}$$

with $\max_y |\varepsilon'_q| \approx 3.155\%$, yielding the constant

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

for IEEE 754-2008 binary64. However, as we will see in the next section, this value does not optimize the final error.

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.

[9]This value may be computed formally, but the expression is unwieldy.

Lagny's rational method

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

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

with the error

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

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,

$$\varepsilon'_\xi = \frac{2{\varepsilon'_q}^3 + {\varepsilon'_q}^4}{3 + 6\varepsilon'_q + 6{\varepsilon'_q}^2 + 2{\varepsilon'_q}^3} = \tfrac{2}{3}{\varepsilon'_q}^3 + \mathcal{O}({\varepsilon'_q}^4).$$

If q' is computed using $\Gamma = \Gamma_{\mathrm{Kahan}}$, we get $\max_y|\varepsilon'_\xi| \approx 21.96 \cdot 10^{-6}$, $\log_2 \max_y|\varepsilon'_\xi| \approx -15.47$. 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

$$\begin{aligned}\Gamma_{\mathrm{L}^1} &\coloneqq \operatorname*{argmin}_{\Gamma\in\mathbb{R}} \max_y|\varepsilon'_\xi| \\ &\approx 0.09918\,74615\,29855\,99525\,66149\,20761\,31234\,35^- \\ &\approx {}_{16}\mathrm{E99.70D0\,DEAD\,BEEF}\end{aligned}$$

with $\max_y|\varepsilon'_q| \approx 3.2025\%$, but $\max_y|\varepsilon'_\xi| \approx 20.86 \cdot 10^{-6}$, $\log_2 \max_y|\varepsilon'_\xi| \approx -15.54$. The corresponding fixed-point constant is

$$C_{\mathrm{L}^1} \coloneqq {}_{16}\mathrm{2A9F\,7893\,782D\,A1CE} \cdot 2^{-52}$$

for binary64.

While it is not far from the seventeen 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 misrounding, see (7). Lagny's lesser-known irrational method provides us with a way to improve it.

Lagny's irrational method

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

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

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, rewriting it as

$$\xi' = \tfrac{1}{2}q' + \frac{1/\sqrt{12}}{q'}\sqrt{4yq' - {q'}^4}, \tag{6}$$

one can evaluate it with similar[10] performance to the rational method.

Its error is

$$\varepsilon'_\xi = \frac{-{\varepsilon'_q}^3}{3\left(\frac{1}{2} + \sqrt{\frac{1}{2} - 2{\varepsilon'_q}^2 - \frac{4}{3}{\varepsilon'_q}^3 - \frac{1}{3}{\varepsilon'_q}^4} - {\varepsilon'_q}^2\right)} = -\tfrac{1}{3}{\varepsilon'_q}^3 + \mathcal{O}({\varepsilon'_q}^4),$$

whose leading term is half that of the rational method; indeed we find that with $\Gamma = \Gamma_{\mathrm{Kahan}}$, we have $\max_y|\varepsilon'_\xi| \approx 10.48 \cdot 10^{-6}$, $\log_2 \max_y|\varepsilon'_\xi| \approx -16.54$, gaining one bit

[10]Other rewritings are preferable on some architectures; we discuss this in appendix C.

with respect to the rational method. Here $\Gamma = \Gamma_{\mathrm{Kahan}}$ is very close to optimal; with the optimal value

$$\begin{aligned}\Gamma_{\mathrm{L}^2} &\approx 0.10096\,82076\,65096\,37285\,40885\,52460\,33434\,6,\\ &\approx {}_{16}\mathrm{E99.70D0\ DEAD\ BEEF}\end{aligned}$$

the error remains the same within the precision to which we have given it. However, we have other ways of improving the error at no cost to performance.

Canon optimization of Lagny's irrational method

The idea for this optimization comes from [Can18a], reproduced here 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.

Canon later elaborated on this in [Can18b]:

> Quick version: we want to compute $1/\sqrt{y}$, we have an approximation x_0, we want to improve it to $x_1 = x_0 p(x_0, y)$. For efficiency, we want p to be a polynomial correction.
> *handwavy motivation for brevity* make p a polynomial in $x_0 x_0 y$, which is approximately 1.
>
> Specifically, if x_0 has relative error e, $x_0 x_0 y$ is bounded by something like $1 \pm 2e$. So, we want to find p that minimizes $|x/x_0 - p(x_0 x_0 y)|$ on $[1-2e, 1+2e]$. NR[11] uses the p = 1st order Taylor. We know that we can do better via usual approximation theory techniques.
>
> We can also use higher-order approximations to hit any specific accuracy target in a single step. This isn't always better than iterating, but sometimes it is.

We do not use a polynomial—nor even a rational function—, nor do we express our refinement as a function of a quantity bounded by the error. However, we take advantage of Canon's key idea of "apply[ing] optimization to the iteration, not just the starting guess"; the latter is what we have so far done with Γ.

The constants $\frac{1}{2}$, $\frac{1}{4}$, and 3 in Lagny's irrational method may be modified with no effect on performance; altering the first two of these introduces rounding errors, but these need not concern us here. We thus write

$$\xi' = \kappa q' + \sqrt{\lambda q'^2 + \frac{y - q'^3}{\mu q'}}$$

and choose Γ, κ, λ, and μ minimizing relative error in the Чебышёв norm,

$$(\Gamma_{\mathrm{L}^2\mathrm{C}}, \kappa_{\mathrm{L}^2\mathrm{C}}, \lambda_{\mathrm{L}^2\mathrm{C}}, \mu_{\mathrm{L}^2\mathrm{C}}) \coloneqq \operatorname*{argmin}_{\Gamma,\kappa,\lambda,\mu} \max_{y} |\varepsilon'_\xi|.$$

Unfortunately, computing $\max_y |\varepsilon'_\xi|$ is not as easy as for the standard methods; the introduction of κ, λ, and μ breaks the monotonicity of $\varepsilon'_\xi(\varepsilon'_q)$, so that the local

[11] Newton–Raphson, *i.e.*, Newton's method. REMOVE BEFORE FLIGHT: Cite Raphson, cite Lagrange's remarks

extrema of ε'_ξ are not found in the same place as those of ε'_q. Formally looking for zeros of the derivative of ε'_ξ with respect to f is impractical. Instead we find the local maxima by numerical maximization on the four pieces whereon q' is a smooth function of f.

That maximum can be minimized by a straightforward hill-climbing[12] starting from $\Gamma = \frac{1}{10}$, $\kappa = \frac{1}{2}$, $\lambda = \frac{1}{4}$, and $\mu = 3$. We obtain the values

$$\begin{aligned}
\Gamma_{\mathrm{L^2C}} &\approx 0.10007\,61614\,69941\,46538\,73178\,74111\,71965\,6,\\
\kappa_{\mathrm{L^2C}} &\approx 0.49999\,99381\,08574\,04775\,14291\,72928\,30652\,9,\\
\lambda_{\mathrm{L^2C}} &\approx 0.25000\,00000\,00145\,58487\,81104\,01052\,77249\,3,\\
\mu_{\mathrm{L^2C}} &\approx 3.00074\,62871\,20756\,72280\,51404\,24030\,90920,
\end{aligned}$$

in hexadecimal

$$\begin{aligned}
\Gamma_{\mathrm{L^2C}} &\approx {}_{16}\mathrm{E99.70D0\ DEAD\ BEEF}\\
\kappa_{\mathrm{L^2C}} &\approx {}_{16}\mathrm{E99.70D0\ DEAD\ BEEF}\\
\lambda_{\mathrm{L^2C}} &\approx {}_{16}\mathrm{E99.70D0\ DEAD\ BEEF}\\
\mu_{\mathrm{L^2C}} &\approx {}_{16}\mathrm{E99.70D0\ DEAD\ BEEF}
\end{aligned}$$

for an error of $\max_y|\varepsilon'_\xi| \approx 2.6157 \cdot 10^{-6}$, $\log_2 \max_y|\varepsilon'_\xi| \approx -18.54$: this optimization gains two bits. The resulting ε'_ξ is remarkably equioscillating, as can be seen in figure ??.

With the rewriting (6), the constants $1/\sqrt{12}$ and 4 should be replaced by

$$\begin{aligned}
\sqrt{\frac{1-\lambda_{\mathrm{L^2C}}\mu_{\mathrm{L^2C}}}{\mu_{\mathrm{L^2C}}}} &\approx 0.28853\,15115\,62316\,71905\,38451\,44194\,38406\,3\\
&\approx {}_{16}\mathrm{E99.70D0\ DEAD\ BEEF}
\end{aligned}$$

and

$$\begin{aligned}
\frac{1}{1-\lambda_{\mathrm{L^2C}}\mu_{\mathrm{L^2C}}} &\approx 4.00298\,73779\,31697\,18250\,67433\,26901\,80421\\
&\approx {}_{16}\mathrm{E99.70D0\ DEAD\ BEEF}
\end{aligned}$$

respectively.

Note that a similar optimization could be applied to the rational method; however, it would not unconditionally be free: changing the 2 in the denominator turns an addition into a multiplication, and inserting additional constants adds more operations. Whether this hinders performance depends on the architecture. In any case, the optimization can scarcely gain more than two bits; such an optimized rational method would still have double the error of the optimized irrational method.

3 Rounded approximation

The number x is obtained from ξ by rounding to $\left\lfloor \frac{p}{3} \right\rfloor$ bits.

Directed rounding toward zero

An easy solution is to zero all but the most significant $\left\lfloor \frac{p}{3} \right\rfloor$ bits of ξ.

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, so that

$$\left|\frac{x}{\xi} - 1\right| \le 1 - \frac{1}{1 + 2^{-\left\lfloor \frac{p}{3} \right\rfloor + 1} - 2^{1-p}} < 2^{-\left\lfloor \frac{p}{3} \right\rfloor + 1} = 2^{-16}.$$

[12]It is plausible that some variation of Remez's algorithm could be used here, much like it can be adapted to rational functions; since the hill-climbing converged satisfactorily, and did so much faster than we were writing this document, we have not investigated this.

For the error of x as an approximation of the cube root,

$$\varepsilon_x \coloneqq \frac{x}{\sqrt[3]{y}} - 1,$$

we have the bound $|\varepsilon_x| < (1 + |\varepsilon_\xi|)(1 + 2^{-16}) - 1$.

Rounding to nearest

An alternative which has similar performance on some architectures, but half the error, is to use Veltkamp's method TODO CITE to round to the nearest number with the desired number of bits; see also [**Dekker1971**] in *mul12*:

$$\varpi \coloneqq [\![\xi(2^{p-\lfloor\frac{p}{3}\rfloor} + 1)]\!],\, x \coloneqq [\![[\![\xi - \varpi]\!] + \varpi]\!].$$

The rounding error is then bounded by $2^{-\lfloor\frac{p}{3}\rfloor}$; we have $|\varepsilon_x| < (1 + |\varepsilon_\xi|)(1 + 2^{-17}) - 1$.

4 High order iteration

We compute the faithfully rounded result r_0 as the correctly rounded sum of the rounded approximation and a correction term,

$$r \coloneqq x + \Delta,\, r_0 \coloneqq [\![r]\!],$$

where the correction term Δ is that of a high-order root finding method. As usual, we call the infinite-precision correction term Δ', and $r' \coloneqq x + \Delta'$. The truncation error is

$$\varepsilon_r' \coloneqq \frac{[\![x + \Delta']\!]}{\sqrt[3]{y}} - 1.$$

The rounding error on the correction term is $\delta \coloneqq \frac{\Delta}{\Delta'} - 1$. The error of r is thus

$$\begin{aligned}
\varepsilon_r &\coloneqq \frac{x + \Delta}{\sqrt[3]{y}} - 1 \\
&= \frac{x + \Delta'(1 + \delta)}{\sqrt[3]{y}} - 1 \\
&= \frac{x + \Delta' + (x + \Delta' - x)\delta}{\sqrt[3]{y}} - 1 \\
&= \varepsilon_r' + (\varepsilon_r' - \varepsilon_x)\delta,
\end{aligned} \tag{7}$$

and that of the faithfully-rounded result r_0 is $\varepsilon_r(1 + \upsilon)$ for some $|\upsilon| \leq u$.

It is easy to make ε_r' negligible by increasing the order of the method; the main contribution to misrounding is then $\delta\varepsilon_x$: this is why ε_x needed to be kept low in step 1. We now compute bounds for ε_r' and δ for two different methods.

Fifth order rational

We use one iteration of the Lagny–Schröder rational method of order 5:

$$\Delta = \left[\!\!\left[\frac{\left[\!\!\left[(y - x^3)\left[\!\!\left[\left[\!\!\left[[\![[\![10x^3]\!] + 16y]\!]x^3\right]\!\!\right] + [\![y^2]\!]\right]\!\!\right]\right]\!\!\right]}{\left[\!\!\left[x^2\left[\!\!\left[\left[\!\!\left[[\![[\![15x^3]\!] + [\![51y]\!]]\!]x^3\right]\!\!\right] + \left[\!\!\left[15[\![y^2]\!]\right]\!\!\right]\right]\!\!\right]\right]\!\!\right]}\right]\!\!\right]$$

where x^2 and x^3 are exact thanks to the trailing 0s of x, $[\![x^6]\!]$ is correctly rounded because it is computed as the square of x^3, and $y - x^3$ is exact by Sterbenz's lemma.

In infinite precision, this method is of such high order that if $\log_2|\varepsilon_x| < 14.5$, which is the case even if ξ is computed by the rational method, the relative error ε_r' of the result is less than 2^{-75}. We will not seek to bound the truncation error more closely, nor to tweak the constants in the method to optimize it: as we will see, it is dominated by rounding.

Thanks to the exact cube and exact difference, the rounding analysis of the correction term is straightforward. All remaining sums being of positive terms, their relative error is readily bounded by the largest of those of their terms. This leads to bounds of γ_5 on the numerator and γ_5 on the denominator, overall

$$\delta < \frac{1+\gamma_6}{1-\gamma_5} - 1 < \frac{\gamma_{11}}{1-\gamma_5} \approx 11u.$$

However, considering that our final bound on ε_r, and thus our misrounding estimate, is nearly proportional to this error, a more careful analysis is warranted. Observe that $x^3 = y(1+\varepsilon_x)^3$, so that a sum

$$\Sigma' = \alpha x^m y^{p-m} + \beta x^n y^{p-n}$$

evaluated with terms that carry the errors δ_i as

$$\Sigma = \alpha x^m y^{p-m}(1+\delta_1) + \beta x^n y^{p-n}(1+\delta_2)$$

has the error

$$\frac{\Sigma}{\Sigma'} - 1 = \frac{(1+\varepsilon_x)^{3m}\alpha\delta_1 + (1+\varepsilon_x)^{3n}\beta\delta_2}{(1+\varepsilon_x)^{3m}\alpha + (1+\varepsilon_x)^{3n}\beta},$$

which we may bound, assuming without loss of generality that $m \le n$, as

$$\left|\frac{\Sigma}{\Sigma'} - 1\right| \le \iota^{3n}\frac{\alpha\delta_1 + \beta\delta_2}{\alpha+\beta},$$

where $\iota \coloneqq \frac{1+|\varepsilon_x|}{1-|\varepsilon_x|}$.

With this property, we can, in particular, take advantage of the exact multiplication by 16 in the numerator; the rounding error analysis of the numerator is as follows.

Expression	Bound on the rounding error
$[\![10x^3]\!] + 16y$	$\iota^3\frac{10u}{26} < \gamma_{\frac{10}{26}\iota^3}$
$[\![[\![10x^3]\!] + 16y]\!]$	$\gamma_{\frac{10}{26}\iota^3+1}$
$[\![[\![[\![10x^3]\!] + 16y]\!]x^3]\!]$	$\gamma_{\frac{10}{26}\iota^3+2}$
$[\![[\![[\![10x^3]\!] + 16y]\!]x^3]\!] + [\![y^2]\!]$	$\iota^6\frac{1}{27}(26\gamma_{\frac{10}{26}\iota^3+2} + u) < \frac{1}{27}\gamma_{10\iota^9+(2\cdot 26+1)\iota^6}$
$[\![[\![[\![[\![10x^3]\!] + 16y]\!]x^3]\!] + [\![y^2]\!]]\!]$	$\frac{1}{27}\gamma_{10\iota^9+(2\cdot 26+1)\iota^6+27}$
$[\![(x^3-y)[\![[\![[\![[\![10x^3]\!] + 16y]\!]x^3]\!] + [\![y^2]\!]]\!]]\!]$	$\frac{1}{27}\gamma_{10\iota^9+(2\cdot 26+1)\iota^6+2\cdot 27}$

The resulting bound on the numerator is approximately $4.33u$, an improvement of $\frac{2}{3}u$ over the naïve bound of $5u$. We may build a similar 𒋛𒄣𒊏𒋫 for the denominator.

Transcription: *siqqurrata.* REMOVE BEFORE FLIGHT: do we want the akk-x-neoass spelling 𒍣𒄣𒊏? That one is invariable though. The akk-x-midass spelling is also used by Šulmānu-ašarēdu III.

Expression	Bound on the rounding error
$[\![15x^3]\!] + [\![51y]\!]$	γ_1
$[\![[\![15x^3]\!] + [\![51y]\!]]\!]$	γ_2
$[\![[\![[\![15x^3]\!] + [\![51y]\!]]\!]x^3]\!]$	γ_3
$[\![[\![[\![15x^3]\!] + [\![51y]\!]]\!]x^3]\!] + [\![15[\![y^2]\!]]\!]$	$\iota^6\frac{66\gamma_3+15\gamma_2}{81} < \frac{1}{81}\gamma_{228\iota^6}$
$[\![[\![[\![[\![15x^3]\!] + [\![51y]\!]]\!]x^3]\!] + [\![15[\![y^2]\!]]\!]]\!]$	$\frac{1}{81}\gamma_{228\iota^6+81}$
$[\![x^2[\![[\![[\![[\![15x^3]\!] + [\![51y]\!]]\!]x^3]\!] + [\![15[\![y^2]\!]]\!]]\!]]\!]$	$\frac{1}{81}\gamma_{228\iota^6+2\cdot 81}$

This is a bound of about $4.81u$ on the denominator, a more modest improvement of $\frac{5}{27}u$ over our earlier $5u$. Overall, we get the bound

$$\delta < \frac{1+\frac{1}{27}\gamma_{10\iota^9+(2\cdot 26+1)\iota^6+2\cdot 27} + \gamma_1}{1-\frac{1}{81}\gamma_{228\iota^6+2\cdot 81}} - 1 < \frac{1+\frac{1}{27}\gamma_{10\iota^9+(2\cdot 26+1)\iota^6+3\cdot 27}}{1-\frac{1}{81}\gamma_{228\iota^6+2\cdot 81}} - 1,$$

approximately $10.14u$.

REMOVE BEFORE FLIGHT: give the resulting bound for ε_r using Lagny's irrational method with Canon optimization.

Sixth order rational

An alternative is the Lagny–Schröder rational method of order 6:

$$\varDelta = \left\llbracket \frac{\left\llbracket \llbracket x(y-x^3)\rrbracket \left\llbracket \left\llbracket \llbracket\llbracket 5x^3\rrbracket + \llbracket 17y\rrbracket\rrbracket x^3 \right\rrbracket + \llbracket 5\llbracket y^2\rrbracket\rrbracket \right\rrbracket \right\rrbracket}{\left\llbracket \left\llbracket \llbracket\llbracket 7x^3\rrbracket + \llbracket 42y\rrbracket\rrbracket\llbracket x^6\rrbracket \right\rrbracket + \left\llbracket \llbracket\llbracket 30x^3\rrbracket + 2y\rrbracket\llbracket y^2\rrbracket \right\rrbracket \right\rrbracket} \right\rrbracket,$$

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

Here we have $|\varepsilon'_r| < 2^{-100}$ if $|\varepsilon_x| < 2^{-14}$, so the truncation error is even more negligible. For rounding error, the maximum bound on the error of the sums gives us a naïve bound of γ_6 on the numerator, γ_5 on the denominator, overall

$$\delta < \frac{1+\gamma_7}{1-\gamma_5} - 1 < \frac{\gamma_{12}}{1-\gamma_5} \approx 12u.$$

Transcription: *siqqurrātu.*

The following 𒋛𒄣𒌨𒊏𒀀𒌅 tighten the bounds on the sums in the numerator and the denominator in the same way as for the fifth order method.

Expression	Bound on the rounding error
$\llbracket 5x^3\rrbracket + \llbracket 17y\rrbracket$	γ_1
$\llbracket\llbracket 5x^3\rrbracket + \llbracket 17y\rrbracket\rrbracket$	γ_2
$\left\llbracket\llbracket\llbracket 5x^3\rrbracket + \llbracket 17y\rrbracket\rrbracket x^3\right\rrbracket$	γ_3
$\left\llbracket\llbracket\llbracket 5x^3\rrbracket + \llbracket 17y\rrbracket\rrbracket x^3\right\rrbracket + \llbracket 5\llbracket y^2\rrbracket\rrbracket$	$\iota^6\frac{22\gamma_3+5\gamma_2}{27} < \frac{1}{27}\gamma_{76\iota^6}$
$\left\llbracket\left\llbracket\llbracket\llbracket 5x^3\rrbracket + \llbracket 17y\rrbracket\rrbracket x^3\right\rrbracket + \llbracket 5\llbracket y^2\rrbracket\rrbracket\right\rrbracket$	$\frac{1}{27}\gamma_{76\iota^6+27}$
$\llbracket x(y-x^3)\rrbracket\left\llbracket\left\llbracket\llbracket\llbracket 5x^3\rrbracket + \llbracket 17y\rrbracket\rrbracket x^3\right\rrbracket + \llbracket 5\llbracket y^2\rrbracket\rrbracket\right\rrbracket$	$\frac{1}{27}\gamma_{76\iota^6+2\cdot 27}$
$\left\llbracket\llbracket x(y-x^3)\rrbracket\left\llbracket\left\llbracket\llbracket\llbracket 5x^3\rrbracket + \llbracket 17y\rrbracket\rrbracket x^3\right\rrbracket + \llbracket 5\llbracket y^2\rrbracket\rrbracket\right\rrbracket\right\rrbracket$	$\frac{1}{27}\gamma_{76\iota^6+3\cdot 27}$

This bound is approximately $5.81u$ on the numerator.

Expression	Bound on the rounding error
$\llbracket 30x^3\rrbracket + 2y$	$\iota^3\frac{30u}{32} < \gamma_{\frac{10}{32}\iota^3}$
$\llbracket\llbracket 30x^3\rrbracket + 2y\rrbracket$	$\gamma_{\frac{10}{32}\iota^3+1}$
$\llbracket\llbracket 30x^3\rrbracket + 2y\rrbracket\llbracket y^2\rrbracket$	$\gamma_{\frac{10}{32}\iota^3+2}$
$\left\llbracket\llbracket\llbracket 30x^3\rrbracket + 2y\rrbracket\llbracket y^2\rrbracket\right\rrbracket$	$\gamma_{\frac{10}{32}\iota^3+3}$
$\left\llbracket\llbracket\llbracket 7x^3\rrbracket + \llbracket 42y\rrbracket\rrbracket\llbracket x^6\rrbracket\right\rrbracket + \left\llbracket\llbracket\llbracket 30x^3\rrbracket + 2y\rrbracket\llbracket y^2\rrbracket\right\rrbracket$	$\iota^6\frac{49\gamma_4+32\gamma_{\frac{10}{32}\iota^3+3}}{81} < \frac{1}{81}\gamma_{10\iota^9+282\iota^6}$
$\left\llbracket\left\llbracket\llbracket\llbracket 7x^3\rrbracket + \llbracket 42y\rrbracket\rrbracket\llbracket x^6\rrbracket\right\rrbracket + \left\llbracket\llbracket\llbracket 30x^3\rrbracket + 2y\rrbracket\llbracket y^2\rrbracket\right\rrbracket\right\rrbracket$	$\frac{1}{81}\gamma_{10\iota^9+282\iota^6+81}$

This bound is approximately $4.61u$ on the denominator. We thus have, for the method of order 6,

$$\delta < \frac{1+\frac{1}{27}\gamma_{76\iota^6+3\cdot 27}+\gamma_1}{1-\frac{1}{81}\gamma_{10\iota^9+282\iota^6+81}} < \frac{1+\frac{1}{27}\gamma_{76\iota^6+4\cdot 27}}{1-\frac{1}{81}\gamma_{10\iota^9+282\iota^6+81}}$$

approximately $11.42u$.

REMOVE BEFORE FLIGHT: give the resulting bound for ε_r using Lagny's irrational method with Canon optimization.

The unreasonable effectiveness of the sixth order rational method

The bound on the rounding error in the sixth order rational method is a little more than $1u$ worse than the one on its fifth order rational method; this is to be expected, as it involves more calculations and does not benefit an exact multiplication by 16. Accordingly, since truncation error is negligible before rounding error for both, one might expect that the fifth order method would be superior. Numerical experiments suggest otherwise: using the method of order 6 leads to 4.33 misroundings per million, whereas the method of order 5 leads to 4.43 per million.

The bounds on δ may need to be tightened further; for instance, we have not taken any account of rounding errors systematically compensating each other, the error induced by $[\![y^2]\!]$ is the same in the numerator and the denominator, but we bound it as if it had opposite signs. The explanation may however lie elsewhere; the larger number of operations involved for the overall maximal rounding error may lead to lower average error.

Whatever the reason may be, the method of order 6 appears to be preferable should one wish to implement a faithful cube root—they are about equally fast on modern architectures. However, the poorer bound on the maximal error means that we must instead use the method of order 5 in our correctly-rounded cube root: its better bound means that we go through the “potential misrounding” path less often.

Correct rounding

We have $r = \sqrt[3]{y}(1 + \varepsilon_r)$, thus

$$\sqrt[3]{y} \in \left[\frac{r}{1 + \max|\varepsilon_r|}, \frac{r}{1 - \max|\varepsilon_r|}\right] =: \mathcal{I}.$$

Consider the *ties*, *i.e.*, the number halfway betwen r_0 and its binary64 successor and the number halfway between r_0 and its predecessor. If $\mathcal{I}$ contains neither of the ties, $r_0 = [\![\sqrt[3]{y}]\!]$: the faithful method returned a correct result.

This criterion, slightly weakened, may be determined as follows. The difference between $r_0 = [\![r]\!]$ and the unrounded r can be computed as described in [**Dekker1971**], $r_1 \coloneqq x - r_0 + \Delta$, evaluated as written (both operations are exact). The potential other candidate for $[\![\sqrt[3]{y}]\!]$ may be computed as $\tilde{r} \coloneqq [\![r_0 + 2r_1]\!]$, which is r_0 only if r_1 is below a quarter-unit in the last place—in which case the small size of $\mathcal{I}$ ensures that it does not contain a tie. If $\tilde{r} \neq r_0$, we must ascertain whether the tie lies in $\mathcal{I}$,

$$\frac{\tilde{r} + r_0}{2} \in \left[\frac{r}{1 + \max|\varepsilon_r|}, \frac{r}{1 - \max|\varepsilon_r|}\right],$$

subtracting r_0 to get rid of the unrepresentable r in the right-hand side,

$$\frac{\tilde{r} - r_0}{2} \in \left[\frac{r_1 - \max|\varepsilon_r| r_0}{1 + \max|\varepsilon_r|}, \frac{r_1 + \max|\varepsilon_r| r_0}{1 - \max|\varepsilon_r|}\right],$$

subtracting r_1 to remove the cancellation,

$$\frac{\tilde{r} - r_0}{2} - r_1 \in \left[-\max|\varepsilon_r| \frac{r_0 + r_1}{1 + \max|\varepsilon_r|}, \max|\varepsilon_r| \frac{r_0 + r_1}{1 - \max|\varepsilon_r|}\right].$$

The left-hand-side may be computed as written; however the bounds of the interval above are not representable. We must relax them a little,

$$\frac{\tilde{r} - r_0}{2} - r_1 \in \left[-\frac{\max|\varepsilon_r|}{1 + \max|\varepsilon_r|}(1 + u) r_0, \frac{\max|\varepsilon_r|}{1 - \max|\varepsilon_r|}(1 + u) r_0\right].$$

On an architecture where multiplications with directed rounding are not more expensive when the surrounding computation uses the roundTiesToEven rounding-direction attribute, these bounds may be used directly, provided that the constants therein have been rounded toward their respective signs, and the multiplication by r_0 be similarly rounded.

If we restrict ourselves to roundTiesToEven at runtime, we must relax the bounds some more,

$$\left[\left[\!\!\left[\left[\!\!\left[-\frac{\max|\varepsilon_r|}{1 + \max|\varepsilon_r|}\left(1 + \frac{2u}{1 - u}\right)\right]\!\!\right]_{-} r_0\right]\!\!\right], \left[\!\!\left[\left[\!\!\left[\frac{\max|\varepsilon_r|}{1 - \max|\varepsilon_r|}\left(1 + \frac{2u}{1 - u}\right)\right]\!\!\right]_{+} r_0\right]\!\!\right]\right].$$

We may widen this interval slightly into one that is symmetric about 0, thus requiring only one comparison

$$|\tilde{r} - r_1| \leq [\![\tau r_0]\!],$$

with

$$\begin{aligned}\tau &:= \left[\!\!\left[\frac{\max|\varepsilon_r|}{1-\max|\varepsilon_r|}\left(1+\frac{2u}{1-u}\right)\right]\!\!\right]_+ \\ &= {}_{16}\mathrm{DEAD\ BEEF},\end{aligned}$$

where we have used the constants for Lagny's irrational method with Canon optimization, followed by the method of fifth order.

If this inequality holds, there may be a misrounding. The correctly-rounded result may then readily be computed using the ordinary cube root algorithm, described, *e.g.*, in Lagny's [Fan97, pp. 286 sqq.], used with binary digits.

Extracting a digit

This method extracts a single digit of $[\![\sqrt[3]{y}]\!]_0$ at each step. Since $\mathcal{I}_\pm$ is small enough that it cannot contain both a floating-point number and a tie; thus if $\mathcal{I}_\pm$ contains a tie, we know that $[\![\sqrt[3]{y}]\!]_0 = [\![r]\!]_0 =: a$; we thus have the first 53 bits already, as

$$a = \min(r_0, \tilde{r}).$$

Correct rounding may be achieved by extracting a single additional bit: the number is in excess of the tie—and thus must be rounded up—if and only if that bit is 1, because, as remarked in [**LangMuller2000**], there are no halfway cases for the cube root.

In the ordinary cube root method, our remainder is $\rho_{53} := y - a^3$, and the next bit is 1 if and only if the next remainder would be positive with that bit,

$$\rho_{54|1} := \rho_{53} - 3a^2b - 3ab^2 - b^3 \geq 0,$$

where b is the power of two corresponding to this bit (the difference between a and the tie).

Using Veltkamp's algorithm from [**Dekker1971**] to express a^2 as $a_0^2 + a_1^2$ with $a_0^2 = [\![a^2]\!]$, the remainder is

$$\rho_{53} = y - (a_0^2 + a_1^2)a;$$

with two more applications of Veltkamp's algorithm,

$$\rho_{53} = y - a_{00}^3 - a_{01}^3 - a_{10}^3 - a_{11}^3,$$

where the first subtraction is representable by Sterbenz's lemma. The remainder cannot have more digits than two plus twice the number of digits computed for the cube root, *i.e.*, $2 \cdot 53 + 2$ significant bits, for otherwise the "digit" $2 = {}_2 10$ would fit in the place b. By repeated exact subtractions we may thus express it as

$$\rho_{53} = \rho_{53;0} + \rho_{53;1} + \rho_{53;2},$$

where $\rho_{53;i} = [\![\rho_{53;i} + \rho_{53;i+1}]\!]$, with ample room to spare. The terms $3a^2b$, $3ab^2$, and b^3 may then be subtracted exactly while retaining triple precision: they add at most three significant digits. The first two of these terms should be split into representable parts,

$$\begin{aligned}3a^2b &= 2a_0^2b + a_0^2b + 2a_1^2b + a_1^2b, \\ 3ab^2 &= 2ab^2 + ab^2.\end{aligned}$$

The sign of $\rho_{54|1}$ may then be checked.

A FMA

The overall strategy is different with FMA, since we need only round x to half the precision (26 bits), rather than a third; this is because the expression $y - x^3$ may be computed as $[\![y - (x^2)x]\!]$, requiring only an exact square.

This means that ε'_ξ should ideally be somewhat less than 2^{-26} or 2^{-27} depending on the manner in which x is rounded. Even with Canon optimization, Lagny's irrational method cannot achieve that from q. The error of the Lagny–Schröder rational method of order 4 reaches a below 2^{-21}, and that of the generalized Lagny quadratic irrational method of order 4 below 2^{-23}. It seems that Canon optimization on the latter can readily bring down its rerror below 2^{-24}, but not much further. With the methods of order 5, the errors are below 2^{-28}; little stands to be gained from optimization.

Conversely the computation of r_0 from x may use a lower-order method, since it starts from 26 rather than 17, but still cannot add more than 53.

B Rounding error analysis for the second step

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

Note that the subtraction in the numerator is exact by Sterbenz's lemma [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

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

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 $b \coloneqq [\![[\![q^2]\!]q]\!] - y$ 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

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

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 $p \geq 12$.

C Performance considerations

D Other rounding modes

E Comparison with other faithful implementations

References

[Can18a] 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.

[Can18b] S. Canon. "A trick I've used for years and should write up: Quick version". Tweets at https://twitter.com/stephentyrone/status/1057788315699687424 *sqq.* 1st Nov. 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`.