Correct rounding of algebraic functions

Jean-Michel Muller, Nicolas Brisebarre

RAIRO-Inf. Theor. Appl. **41** (2007) 71–83
DOI: 10.1051/ita:2007002

CORRECT ROUNDING OF ALGEBRAIC FUNCTIONS

Nicolas Brisebarre[1, 2] and Jean-Michel Muller[1]

Abstract. We explicit the link between the computer arithmetic problem of providing correctly rounded algebraic functions and some diophantine approximation issues. This allows to get bounds on the accuracy with which intermediate calculations must be performed to correctly round these functions.

Mathematics Subject Classification. 11J68, 65D20, 65G.

1. Introduction

On most current computer systems, the real numbers are approximated by floating-point numbers. For many years, floating-point arithmetic has been a mere set of cooking recipes. The consequences of this have sometimes been disastrous: numerical programs were not reliable nor portable. Without a clear specification of the underlying arithmetic, it was not possible to prove even simple properties of a sequence of operations, and the only way to feel comfortable with an important numerical program was to perform intensive tests.

The IEEE-754 [1,4] standard for binary floating-point arithmetic (and the radix independent IEEE-854 [3,9] standard that followed) put an end to this dangerous era. The IEEE-754 standard clearly specifies the formats of the floating-point representations of numbers, and the behaviour of the four arithmetic operations and the square root.

Keywords and phrases. Floating-point arithmetic, computer arithmetic, algebraic functions, correct rounding, diophantine approximation.

[1] Laboratoire LIP (CNRS/ENS Lyon/INRIA/Univ. Lyon 1), Projet Arénaire, 46 allée d'Italie, 69364 Lyon Cedex 07, France; Nicolas.Brisebarre@ens-lyon.fr; Jean-Michel.Muller@ens-lyon.fr

[2] Laboratoire LaMUSE, Université J. Monnet (Saint-Étienne), 23, rue du Dr P. Michelon, 42023 Saint-Étienne Cedex 02, France.

Define $\mathcal{D}_n^r$ as the set of exponent-unbounded, n-bit mantissa, radix-r floating-point numbers (with $n \geq 1$), that is:

$$\mathcal{D}_n^r = \left\{ M \times r^E, r^{n-1} \leq |M| \leq r^n - 1, M, E \in \mathbb{Z} \right\} \cup \{0\}.$$

$\mathcal{D}_n^r$ is not the set of available floating-point numbers on an existing system. It is an "ideal" system, with no overflows or underflows. We will show results in $\mathcal{D}_n^r$. These results will remain true in an actual system, provided that no overflows or underflows occur.

We will call *mantissa* of a nonzero element $x = M \times r^E$ of $\mathcal{D}_n^r$ the number $\mathcal{M}(x) = M/r^{n-1}$. The *exponent* of an element $M \times r^E$, with $r^{n-1} \leq |M| \leq r^n - 1$, will be the integer $E + n - 1$.

The most frequent choice for the radix is, by far, $r = 2$. And yet, some systems (for instance, most pocket calculators) still use $r = 10$. Other choices have sometimes been made during the early years of electronic computing. Small powers of 2 (4, 8 or 16) have frequently been chosen. There even existed a radix-3 computer, the SETUN calculator, built in Russia during the 60's. 50 copies of this computer were built. It used $r = 3$, and digits -1, 0 and $+1$ [2].

The result of an arithmetic operation whose input values belong to $\mathcal{D}_n^r$ may not belong to $\mathcal{D}_n^r$ (in general it does not). Hence that result must be *rounded*. The IEEE standard defines 4 different rounding modes:

- rounding towards $+\infty$, or upwards: $\circ_u(x)$ is the smallest element of $\mathcal{D}_n^r$ that is greater than or equal to x;
- rounding towards $-\infty$, or downwards: $\circ_d(x)$ is the largest element of $\mathcal{D}_n^r$ that is less than or equal to x;
- rounding towards 0: $\circ_z(x)$ is equal to $\circ_u(x)$ if $x < 0$, and to $\circ_d(x)$ otherwise;
- rounding to the nearest even: $\circ_n(x)$ is the element of $\mathcal{D}_n^r$ that is closest to x. If x is exactly halfway between two consecutive elements of $\mathcal{D}_n^r$, $\circ_n(x)$ is the one for which M is an even number.

The first three rounding modes are called *directed* rounding modes.

The standard requires that the user should be able to choose one rounding mode among these ones, called the *active rounding mode*. After that, when performing one of the 4 arithmetic operations, or when computing square roots, the obtained result should be equal to the rounding of the exact result.

The IEEE standard does not require correct rounding of other functions than the square root and the four arithmetic operations. This is mainly due to the *table maker's dilemma*, presented below. And yet, being able to provide correctly rounded functions is of uttermost interest: first, this is the best way to *standardize* the functions (that is, to provide specifications so that the output of a given function of a given input value will be the same on any system). Standardizing the functions would drastically improve the portability of numerical programs. Second, this would help much the implementation of efficient and cheap *interval arithmetic*.

When the function being implemented is simple enough (multiplication, division, etc.), there are specific algorithms that allow to evaluate them with correct rounding at low cost. For all other functions, the only thing we are able to do is to compute some *approximation* to the exact result, with a precision (which is a rational integer power of the radix r) somewhat higher than the "target" precision. The *table maker's dilemma* is the problem of determining what should the accuracy of the approximation be to make sure that rounding that approximation will always be equivalent to rounding the exact result.

It is worth noting that, once we know an approximation and an error bound on this approximation, we are able to determine whether it is possible to round it correctly. From the approximation and the error bound (and, possibly some knowledge such as the sign of the error), we deduce that the exact result lies in some interval $\mathcal{I}$. We are able to provide a correctly rounded result:

With the directed rounding modes: if $\mathcal{I}$ does not contain an element of $\mathcal{D}_n^r$;

With the rounding to nearest mode: if $\mathcal{I}$ does not contain the exact middle of two consecutive elements of $\mathcal{D}_n^r$.

We could use this property, and implement correctly rounded functions using the following strategy[1]:

- evaluate an initial approximation to $f(x)$;
- while the accuracy of the approximation does not suffice to provide a correctly rounded result, recompute a more accurate approximation.

Unfortunately, for many applications, this is not satisfactory, for two reasons:

- when the approximations are evaluated using a fast pipelined multiplier or multiplier-accumulator, the tests required by the previous strategy would require to wait many cycles before being able to restart the computation;
- for real-time applications, the delay of computation must be bounded.

In this paper, we are concerned with the correct rounding of what we will call *algebraic functions*. In this article, an *algebraic function* f is a function for which there exists a nonzero 2-variable polynomial P with rational integer coefficients such that $P(x, f(x)) = 0$. Examples are:

Division, reciprocation: $y = a/x$, with $P(x, y) = a - xy$;
Square roots: $y = \sqrt{x}$, with $P(x, y) = x - y^2$;
Roots: $y = x^{1/p}$, with $P(x, y) = x - y^p$;
Square root reciprocal: $y = 1/\sqrt{x}$, with $P(x, y) = 1 - xy^2$.

When $x \in \mathcal{D}_n^r$ and f is an algebraic function, the value $f(x)$ is an algebraic number *i.e.*, the root of a nonzero one-variable polynomial with rational integer coefficients. When α is an algebraic number, the minimal polynomial of α over $\mathbb{Z}$ is the polynomial $P \in \mathbb{Z}[X] \setminus \{0\}$, with relatively prime coefficients and positive leading coefficient, of least degree such that $P(\alpha) = 0$. Let d denote the degree of P, we say that α is an algebraic number of degree d.

[1]Assuming the exact result is not an element of $\mathcal{D}_n^r$ or the exact middle of two consecutive elements of $\mathcal{D}_n^r$.

In this paper, we use a classical number-theoretic approach to get bounds, as it is done in [10] and [12], on the accuracy with which intermediate calculations must be performed to correctly round algebraic functions. We state also some diophantine problems whose (even partial) solving should lead to an improvement of the bounds we give. We have to recall that there also exist nice algorithmic approaches to the TMD, *cf.* [13,14,20]. These approaches allow to solve the TMD for the IEEE double precision and should also allow to solve it for the IEEE double extended precision.

2. Formalization of the problem

Assume we wish to correctly round an algebraic function f. We consider that all input values are elements of $\mathcal{D}_n^r$ with the same exponent e_1. A different analysis must be done for each possible value of e_1. For "power functions" of the form $x^{p/q}$ it suffices to consider q consecutive values of the exponent, since

$$(r^q x)^{p/q} = r^p \left(x^{p/q}\right)$$

implies that all other cases are immediately deduced from these ones.

If the values of $f(x)$, for $x \in [r^{e_1}, r^{e_1+1})$ are not all included in an interval of the form $[r^{e_2}, r^{e_2+1})$, we split the input interval into subintervals such that for each subinterval, there is a rational integer e_2 such that the values $f(x)$, for x in the subinterval, are in $[r^{e_2}, r^{e_2+1})$.

We now consider the processing of one subinterval I included in $[r^{e_1}, r^{e_1+1})$.

2.1. Directed rounding modes

The problem to be solved is *Find the biggest precision (by this, we mean the biggest rational integer power of the radix r) such that, for all $x \in \mathcal{D}_n^r \bigcap I$ such that $f(x) \notin \mathcal{D}_n^r$ and for all $y \in \mathcal{D}_n^r$, the distance between $f(x)$ and y is greater than this precision.* By defining integers $X = r^{n-1-e_1}x$ and $Y = r^{n-1-e_2}y$, we get

Problem 1 (general problem, directed rounding modes)**.** *What is the minimum $\mu(n) \in \mathbb{Z}$ such that, for $r^{n-1} \leq X \leq r^n - 1$ (and, possibly, the restrictions implied by $x \in I$) such that $f\left(Xr^{e_1-n+1}\right) \notin \mathcal{D}_n^r$ and for $r^{n-1} \leq Y \leq r^n - 1$,*

$$\left|f\left(Xr^{e_1-n+1}\right) - Yr^{e_2-n+1}\right| > r^{-\mu(n)}.$$

Hence, the biggest precision sought will be $r^{-\mu(n)}$.

2.2. Rounding to nearest mode

Here, the problem to be solved is *Find the biggest precision such that, for all $x \in \mathcal{D}_n^r \bigcap I$ such that $f(x)$ is not the middle of two consecutive elements of $\mathcal{D}_n^r$ and for all $y \in \mathcal{D}_n^r$, the distance between $f(x)$ and $y + \frac{1}{2r^{n-e_2-1}}$ is greater than this precision.* By defining integers $X = r^{n-1-e_1}x$ and $Y = r^{n-1-e_2}y$, we get

Problem 2 (general problem, rounding to nearest mode)**.** *What is the minimum* $\mu(n) \in \mathbb{Z}$ *such that, for* $r^{n-1} \leq X \leq r^n - 1$ *(and, possibly, the restrictions implied by* $x \in I$*) such that* $f\left(Xr^{e_1-n+1}\right)$ *is not the middle of two consecutive elements of* $\mathcal{D}_n^r$ *and for* $r^{n-1} \leq Y \leq r^n - 1$*,*

$$\left| f\left(Xr^{e_1-n+1}\right) - \left(Y + \frac{1}{2}\right) r^{e_2-n+1} \right| > r^{-\mu(n)}.$$

Here again, the biggest precision sought will be $r^{-\mu(n)}$.

These two problems are a special kind of best rational approximation problem, for which the denominators of all considered rationals are powers of r or double of powers of r. Hence we give them a diophantine approximation approach in the following.

3. Some heuristics for the value of $\mu(n)$

It is generally expected that $\mu(n)$ is close to $n + \log_r N$ where N is the number of floating-point numbers of the domain being considered.

Before giving a number-theoretical heuristic, we briefly give an idea of some probabilistic arguments in favor of such an estimate. A probabilistic study of that issue has been done by Dunham [6] and by Gal and Bachelis [7].

Assume that after the n^{th} digit, the digits of the mantissas of the values $f(x)$, where x is a floating-point number, are $0, \ldots, r-1$ with equal probability $1/r$. The probability that after digit n, we have

- in **rounding to nearest mode**, the digit sequence

$$\underbrace{\frac{r}{2}00\cdots0}_{k \text{ digits}} \quad \text{or} \quad \underbrace{\left(\frac{r}{2}-1\right)(r-1)(r-1)\cdots(r-1)}_{k \text{ digits}}$$

  for r even, or the digit sequence

$$\underbrace{\frac{r-1}{2}\frac{r-1}{2}\cdots\frac{r-1}{2}}_{k \text{ digits}} \quad \text{or} \quad \underbrace{\frac{r-1}{2}\frac{r-1}{2}\cdots\frac{r-1}{2}\frac{r-3}{2}}_{k \text{ digits}}$$

  for r odd,
- or, in **directed rounding mode**, the digit sequence

$$\underbrace{00\cdots0}_{k \text{ digits}} \quad \text{or} \quad \underbrace{(r-1)(r-1)\cdots(r-1)}_{k \text{ digits}}$$

is $2r^{-k}$. Hence, if we have N floating-point numbers in the domain being considered, the number of values x for which we will have a digit sequence of the form indicated above is around $2Nr^{-k}$. This number vanishes as soon as k is large

enough compared with $\log_r(2N)$. This means that in practice the largest value of k is slightly above $\log_r(2N)$ which means that the value of $\mu(n)$ is slightly above $n + \log_r(N) + \log_r(2)$.

Now, we would like to recall also diophantine approximation results that confirm that such an estimate seems reasonable. In the following five results, the reader should replace, cf. Problems 1 and 2, the real number α with $f\left(Xr^{e_1-n+1}\right)$, the rational integer u with the numerator of Yr^{e_2-n+1} (resp. $(Y+1/2)r^{e_2-n+1}$) and the rational integer v with the denominator of Yr^{e_2-n+1} (resp. $(Y+1/2)r^{e_2-n+1}$).

First, we recall a result by Hurwitz [8] that suggests that, for a function f which can take some irrational values when evaluated on the set of floating-point numbers, we cannot expect $\mu(n)$ to be less that $n + \log_r N$.

Theorem 1 (Hurwitz)**.** *For all $\alpha \in \mathbb{R} \setminus \mathbb{Q}$, the inequality*

$$\left|\alpha - \frac{u}{v}\right| < \frac{1}{\sqrt{5}v^2}$$

has infinitely many integer solutions u and $v \neq 0$. The factor $\sqrt{5}$ is optimal.

For example, let $f = \sqrt[3]{}$, it is very likely (and some examples show that this is indeed the case) that there exist some floating-point numbers x in $[1, r^3)$ and $y = Y/r^{n-1} \in [1, r)$ such that

$$\left|\sqrt[3]{x} - \frac{Y}{r^{n-1}}\right| < \frac{1}{\sqrt{5}r^{2(n-1)}} = \frac{1}{\sqrt{5}r^{n+\log_r N-1}}.$$

Thus, the problem is to determine how close the values of f over $\mathcal{D}_n^r$ can be approximated by floating-point numbers or middle of two consecutive floating-point numbers. First, we can notice that the infinitely many solutions of Hurwitz theorem share the same specific property: they all are convergents of α (see [11] for an account on continued fractions). This is a classical result of Legendre [15].

Theorem 2 (Legendre)**.** *Let $\alpha \in \mathbb{R}$, let u, v be nonzero integers, with* $\gcd(u, v) = 1$*. If*

$$\left|\alpha - \frac{u}{v}\right| < \frac{1}{2v^2}$$

then u/v is a convergent of α.

In other words, our problem is equivalent to finding the values of f that admit a floating-point number or the middle of two floating-point numbers as a convergent and then to check how good the rational approximation can be. A result of Khintchine says that "in general" the approximation is the worst possible.

Theorem 3 (Khintchine)**.** *Let $\varphi : \mathbb{N} \setminus \{0\} \to \mathbb{N} \setminus \{0\}$, such that $k \mapsto k\varphi(k)$ is decreasing. Then, for almost all $\alpha \in \mathbb{R}$, the inequality*

$$\left|\alpha - \frac{u}{v}\right| < \frac{\varphi(v)}{v}$$

has finitely many integer solutions u and v, $v > 0$ if and only if the series $\sum_{k=1}^{\infty} \varphi(k)$ is convergent.

In particular, that implies the following result: for almost all $\alpha \in \mathbb{R}$, for all $\varepsilon > 0$, the inequality

$$\left|\alpha - \frac{u}{v}\right| < \frac{1}{v^{2+\varepsilon}}$$

has finitely many integer solutions u and v, $v > 0$. If we compare this result to Theorem 1, we see that it indicates that the exponent 2 is expected to be the best possible. This is the case for algebraic numbers, as it was proved by Roth in 1955 [19].

Theorem 4 (Roth)**.** *Let α be an algebraic number of degree $d \geq 2$. For all $\varepsilon > 0$, there exists $C_{\varepsilon,\alpha} > 0$ such that, for all u, $v \in \mathbb{Z}$, $v \geq 1$,*

$$\left|x - \frac{u}{v}\right| > \frac{C_{\varepsilon,\alpha}}{v^{2+\varepsilon}}. \tag{1}$$

The constant $C_{\varepsilon,\alpha}$, that depends only on ε and α, is, unfortunately, not effectively computable. Hence, this theorem is useless for computing an effective upper bound for $\mu(n)$. To do that, we use a result by Liouville that proves a worse exponent but gives effective results. Before stating Liouville's theorem, we want to mention that in the case of radix 2 and the reciprocal square root[2], Croot, Li and Zhu show in [5] that the *abc* conjecture of Masser and Oesterlé implies that (1) is true with a constant independant of α, which implies that $\mu(n)$ is equal to $n + \log_r(N)$ plus a few more bits. But this result is useless in practice since the constant is not effectively computable, which prevents us from estimating the number of the "few more bits".

Theorem 5 (Liouville)**.** *Let α be an algebraic number of degree $d \geq 2$. There exists a constant C_α such that, for all u, $v \in \mathbb{Z}$, $v \geq 1$,*

$$\left|\alpha - \frac{u}{v}\right| > \frac{C_\alpha}{v^d}.$$

The effective constant C_α is given by $C_\alpha = \frac{1}{\max_{|t-\alpha|\leq 1/2} |P'(t)|}$ where $P \in \mathbb{Z}[X]$ is the minimal polynomial of α over $\mathbb{Z}$.

The exponent d is worse (except in the quadratic – $d = 2$ – case) than Roth's theorem exponent but, this time, the constant is effective and allows practical applications. In Section 5, we follow the classical Liouville's approach [16–18] in order to obtain effective bounds for the intermediate precision necessary to get correct rounding. But, before, we deal with the peculiar case of the division.

[2]The adaptation to the other power functions and radices is not difficult.

4. Division

Let $a = A/r^{n-1} \in \mathcal{D}_n^r$ of exponent 0. We consider input values x of exponent 0. We assume $x \neq 1$ and a/x is is not an element of $\mathcal{D}_n^r$ (resp. not the middle of two consecutive elements of $\mathcal{D}_n^r$) in directed rounding modes (resp. in the rounding to nearest mode).

4.1. Directed rounding

First, we assume $a < x$. For all $y \in \mathcal{D}_n^r \cap (1/r, 1)$, we have

$$\left|\frac{a}{x} - y\right| = \left|\frac{A}{X} - \frac{Y}{r^n}\right| = \frac{|Ar^n - XY|}{Xr^n}.$$

As $|a/x - y| \neq 0$, the integer $|Ar^n - XY|$ cannot be 0. Hence,

$$\left|\frac{a}{x} - y\right| \geq \frac{1}{Xr^n} \geq \frac{1}{r^n(r^n - 1)} > r^{-2n}.$$

Then, if we assume $a > x$, for $y \in \mathcal{D}_n^r \cap (1, r)$, we get

$$\left|\frac{a}{x} - y\right| \geq \frac{1}{Xr^{n-1}} \geq \frac{1}{r^{n-1}(r^n - 1)} > r^{-2n+1}.$$

Hence, we obtain $\mu(n) \leq 2n$.

4.2. Rounding to nearest

First, we assume $a < x$. If $y \in \mathcal{D}_n^r \cap [1/r, 1)$, we have

$$\left|\frac{a}{x} - y - \frac{1}{2r^n}\right| = \left|\frac{A}{X} - \frac{2Y+1}{2r^n}\right| = \frac{|2r^nA - X(2Y+1)|}{2Xr^n} \geq \frac{1}{2Xr^n} > \frac{1}{2r^{2n}}.$$

The numerator $|2r^nA - X(2Y+1)|$ is a nonzero integer since $|a/x - y - 1/(2r^n)| \neq 0$.

Then, if we assume $a > x$, for $y \in \mathcal{D}_n^r \cap (1, r)$, we get

$$\left|\frac{a}{x} - y - \frac{1}{2r^{n-1}}\right| \geq \frac{1}{2Xr^{n-1}} \geq \frac{1}{2r^{n-1}(r^n - 1)} > \frac{1}{2r^{2n-1}}.$$

Thus, we have $\mu(n) \leq 2n + 1$.

5. Power functions

We consider the functions $t \mapsto t^{p/q}$ with $p \in \mathbb{Z}\backslash\{0\}$ and $q \in \mathbb{N}\backslash\{0\}$, $gcd(p, q) = 1$.

5.1. Case when p positive

For $k \in \{0, \ldots, p-1\}$, let $x \in [r^{kq/p}, r^{(k+1)q/p})$ such that x in $\mathcal{D}_n^r$, let $y \in [r^k, r^{k+1})$. Let $j_k \in \mathbb{N}, \lfloor kq/p \rfloor \leq j_k \leq [(k+1)q/p] - 1$ such that $x \in [r^{j_k}, r^{j_k+1})$. We assume that $f(x) \notin \mathcal{D}_n^r$ in the directed rounding cases and $f(x)$ is not the middle of two consecutive elements of $\mathcal{D}_n^r$ in the rounding to nearest case.

The number $x^{p/q}$ is a root of the polynomial $P(t) = t^q - x^p$.

5.1.1. *Directed rounding*

There exists c strictly between $x^{p/q}$ and y such that

$$P(y) - P\left(x^{p/q}\right) = P'(c)\left(y - x^{p/q}\right)$$

i.e.,

$$y^q - x^p = qc^{q-1}\left(y - x^{p/q}\right).$$

Hence, we have

$$\begin{aligned}\left|y - x^{p/q}\right| &= \frac{1}{qc^{q-1}}\left|\frac{Y^q}{r^{q(n-1-k)}} - \frac{X^p}{r^{p(n-1-j_k)}}\right| \\ &= \frac{1}{qc^{q-1}}\left|\frac{r^{kq+\max(p-q,0)(n-1)}Y^q - r^{pj_k+\max(q-p,0)(n-1)}X^p}{r^{\max(p,q)(n-1)}}\right| \\ &= \frac{r^{\min(a,b)}}{qc^{q-1}}\frac{\left|r^{a-\min(a,b)}Y^q - r^{b-\min(a,b)}X^p\right|}{r^{\max(p,q)(n-1)}} \qquad (2)\end{aligned}$$

where $a = kq + \max(p-q, 0)(n-1)$ and $b = pj_k + \max(q-p, 0)(n-1)$.

As $c < r^{k+1}$, Equality (2) implies

$$\left|y - x^{p/q}\right| > \frac{r^{\min(a,b)}}{qr^{(q-1)(k+1)}}\frac{\left|r^{a-\min(a,b)}Y^q - r^{b-\min(a,b)}X^p\right|}{r^{\max(p,q)(n-1)}}.$$

As we assumed $\left|y - x^{p/q}\right| \neq 0$, $r^{a-\min(a,b)}Y^q - r^{b-\min(a,b)}X^p$ is a nonzero integer. Therefore, we get the following bound

$$\begin{aligned}\mu(n) \leq -\log_r\left(\left|y - x^{p/q}\right|\right) \leq \max(p,q)(n-1) + (q-1)(k+1) + \log_r(q) \\ - \min(kq + \max(p-q,0)(n-1), pj_k + \max(q-p,0)(n-1)).\end{aligned}$$

Our analysis leads to the following diophantine problem whose solving should decrease the upper bound for $\mu(n)$.

Problem 3 ($x^{p/q}$ for $x \in [1, r^q)$, directed rounding modes)**.** *For $k \in \{0, \ldots, p-1\}$, for $j_k \in \mathbb{N}, \lfloor kq/p \rfloor \leq j_k \leq [(k+1)q/p] - 1$, find two integers $X \in [r^{j_k+n-1}, r^{j_k+n})$ and $Y \in [r^{k+n-1}, r^{k+n})$ that minimize*

$$\left|r^{a-\min(a,b)}Y^q - r^{b-\min(a,b)}X^p\right|$$

where $a = kq + \max(p-q,0)(n-1)$ *and* $b = pj_k + \max(q-p,0)(n-1)$.

5.1.2. *Rounding to nearest*

There exists c strictly between $x^{p/q}$ and $y + 1/(2r^{n-1-k})$ such that

$$P\left(y + \frac{1}{2r^{n-1-k}}\right) - P\left(x^{p/q}\right) = P'(c)\left(y + \frac{1}{2r^{n-1-k}} - x^{p/q}\right)$$

i.e.,

$$\left(y + \frac{1}{2r^{n-1-k}}\right)^q - x^p = qc^{q-1}\left(y + \frac{1}{2r^{n-1-k}} - x^{p/q}\right).$$

Hence, we have

$$\begin{aligned}\left|y + \frac{1}{2r^{n-1-k}} - x^{p/q}\right| &= \frac{1}{qc^{q-1}}\left|\frac{(2Y+1)^q}{2^q r^{q(n-1-k)}} - \frac{X^p}{r^{p(n-1-j_k)}}\right| \\ &= \frac{1}{qc^{q-1}}\frac{\left|r^{kq+\max(p-q,0)(n-1)}(2Y+1)^q - 2^q r^{pj_k+\max(q-p,0)(n-1)}X^p\right|}{2^q r^{\max(p,q)(n-1)}} \\ &= \frac{r^{\min(a,b)}D}{qc^{q-1}}\frac{\left|(2Y+1)^q r^{a-\min(a,b)}/D - X^p 2^q r^{b-\min(a,b)}/D\right|}{2^q r^{\max(p,q)(n-1)}} \qquad (3)\end{aligned}$$

with $a = kq + \max(p-q,0)(n-1)$, $b = pj_k + \max(q-p,0)(n-1)$ and $D = \gcd(2^q, r^{\max(a-b,0)})$.

As $c < r^{k+1}$, Equality (3) implies

$$\left|y - x^{p/q}\right| > \frac{r^{\min(a,b)}D}{qr^{(q-1)(k+1)}}\frac{\left|(2Y+1)^q r^{a-\min(a,b)}/D - X^p 2^q r^{b-\min(a,b)}/D\right|}{2^q r^{\max(p,q)(n-1)}}$$

from which we deduce the following upper bound, since $(2Y+1)^q r^{a-\min(a,b)}/D - X^p 2^q r^{b-\min(a,b)}/D$ is a nonzero integer,

$$\begin{aligned}\mu(n) \le -\log_r\left(\left|y - x^{p/q}\right|\right) &\le \max(p,q)(n-1) + (q-1)(k+1) \\ &\quad + \log_r(q) + q\log_r(2) - \min(a,b) - \log_r(\gcd(2^q, r^{\max(a-b,0)})).\end{aligned}$$

Here again, we notice that the solving of the following diophantine problem should allow to decrease the upper bound for $\mu(n)$.

Problem 4 ($x^{p/q}$ for $x \in [1, r^q)$, rounding to nearest mode)**.** *For* $k \in \{0, \ldots, p-1\}$, *for* $j_k \in \mathbb{N}$, $\lfloor kq/p \rfloor \le j_k \le \lceil (k+1)q/p \rceil - 1$, *find two integers* $X \in [r^{j_k+n-1}, r^{j_k+n})$ *and* $Y \in [r^{k+n-1}, r^{k+n})$ *that minimize*

$$\left|(2Y+1)^q r^{a-\min(a,b)}/D - X^p 2^q r^{b-\min(a,b)}/D\right|$$

with $a = kq + \max(p-q,0)(n-1)$, $b = pj_k + \max(q-p,0)(n-1)$ *and* $D = \gcd(2^q, r^{\max(a-b,0)})$.

5.2. Case when p negative

For $k \in \{0, \ldots, p-1\}$, let $x \in [r^{kq/|p|}, r^{(k+1)q/|p|})$ such that x in $\mathcal{D}_n^r$, let $y \in (r^{-k-1}, r^{-k}]$. Let $j_k \in \mathbb{N}, \lfloor kq/|p| \rfloor \leq j_k \leq [(k+1)q/|p|] - 1$ such that $x \in [r^{j_k}, r^{j_k+1})$. We assume that $f(x) \notin \mathcal{D}_n^r$ in the directed rounding cases and $f(x)$ is not the middle of two consecutive elements of $\mathcal{D}_n^r$ in the rounding to nearest case.

Here again, we notice that $x^{p/q}$ is a root of the polynomial $P(t) = t^q - x^p$.

5.2.1. *Directed rounding*

There exists c strictly between $x^{p/q}$ and y such that

$$P(y) - P\left(x^{p/q}\right) = P'(c)\left(y - x^{p/q}\right)$$

i.e.,

$$y^q - x^p = qc^{q-1}\left(y - x^{p/q}\right).$$

Hence, we have

$$\begin{aligned}\left|y - x^{p/q}\right| &= \frac{1}{qc^{q-1}}\left|\frac{Y^q}{r^{q(n+k)}} - \frac{r^{|p|(n-1-j_k)}}{X^{|p|}}\right| \\ &= \frac{1}{qc^{q-1}}\left|\frac{X^{|p|}Y^q - r^{q(n+k)+|p|(n-1-j_k)}}{r^{q(n+k)}X^{|p|}}\right|. \end{aligned} \tag{4}$$

As $c < r^{-k}$ and $X \leq r^{n-1-j_k}$, Equality (4) implies

$$\left|y - x^{p/q}\right| > \frac{r^{k(q-1)}}{q}\left|\frac{X^{|p|}Y^q - r^{q(n+k)+|p|(n-1-j_k)}}{r^{q(n+k)}r^{|p|(n-1-j_k)}}\right|.$$

As we assumed $\left|y - x^{p/q}\right| \neq 0$, the integer $X^{|p|}Y^q - r^{q(n+k)+|p|(n-1-j_k)}$ is not equal to zero. Thus, we obtain the following upper bound for $\mu(n)$:

$$\mu(n) \leq -\log_r\left(\left|y - x^{p/q}\right|\right) \leq qn + k + \log_r(q) + |p|(n-1-j_k).$$

Hence, we state the following diophantine problem, still in order to decrease the upper bound for $\mu(n)$.

Problem 5 ($x^{p/q}$ for $x \in [1, r^q)$, directed rounding modes)**.** *For $k \in \{0, \ldots, p-1\}$, for $j_k \in \mathbb{N}, \lfloor kq/|p| \rfloor \leq j_k \leq [(k+1)q/|p|] - 1$, find two integers $X \in [r^{j_k+n-1}, r^{j_k+n})$ and $Y \in [r^{k+n}, r^{k+n+1})$ that minimize*

$$\left|X^{|p|}Y^q - r^{q(n+k)+|p|(n-1-j_k)}\right|.$$

5.2.2. *Rounding to nearest*

There exists c strictly between $x^{p/q}$ and $y + 1/(2r^{n-1-k})$ such that

$$P\left(y + \frac{1}{2r^{n-1-k}}\right) - P\left(x^{p/q}\right) = P'(c)\left(y + \frac{1}{2r^{n-1-k}} - x^{p/q}\right)$$

i.e.,

$$\left(y + \frac{1}{2r^{n-1-k}}\right)^q - x^p = qc^{q-1}\left(y + \frac{1}{2r^{n-1-k}} - x^{p/q}\right).$$

Hence, we have

$$\begin{aligned}\left|y + \frac{1}{2r^{n+k}} - x^{p/q}\right| &= \frac{1}{qc^{q-1}}\left|\frac{(2Y+1)^q}{2^q r^{q(n+k)}} - \frac{r^{|p|(n-1-j_k)}}{X^{|p|}}\right| \\ &= \frac{1}{qc^{q-1}}\left|\frac{X^{|p|}(2Y+1)^q - 2^q r^{q(n+k)+|p|(n-1-j_k)}}{2^q r^{q(n+k)} X^{|p|}}\right|. \quad (5)\end{aligned}$$

As $c < r^{-k}$ and $X \leq r^{n-1-j_k}$, Equality (5) implies

$$\left|y - x^{p/q}\right| > \frac{r^{k(q-1)}}{q}\left|\frac{X^{|p|}(2Y+1)^q - 2^q r^{q(n+k)+|p|(n-1-j_k)}}{2^q r^{q(n+k)} r^{|p|(n-1-j_k)}}\right|.$$

For $X^{|p|}(2Y+1)^q - 2^q r^{q(n+k)+|p|(n-1-j_k)}$ is a nonzero integer, we have

$$\mu(n) \leq -\log_r\left(\left|y - x^{p/q}\right|\right) \leq q(n + \log_r(2)) + k + \log_r(q) + |p|(n-1-j_k).$$

To finish, we state the following diophantine problem that arises naturally in our analysis in order to get an upper bound for $\mu(n)$ as small as possible.

Problem 6 ($x^{p/q}$ for $x \in [1, r^q)$, rounding to nearest mode)**.** *For $k \in \{0, \ldots, p-1\}$, for $j_k \in \mathbb{N}$, $\lfloor kq/|p| \rfloor \leq j_k \leq [(k+1)q/|p|]-1$, find two integers $X \in [r^{j_k+n-1}, r^{j_k+n})$ and $Y \in [r^{k+n}, r^{k+n+1})$ that minimize*

$$\left|X^{|p|}(2Y+1)^q - 2^q r^{q(n+k)+|p|(n-1-j_k)}\right|.$$

References

[1] American National Standards Institute and Institute of Electrical and Electronic Engineers. *IEEE standard for binary floating-point arithmetic.* ANSI/IEEE Standard, Std 754-1985, New York (1985).

[2] N.P. Brousentsov, S.P. Maslov, J. Ramil Alvarez and E.A. Zhogolev, *Development of ternary computers at Moscow State University.* Technical report, Dept. VMK MGU (2000). Available at `http://www.computer-museum.ru/english/setun.htm`.

[3] W.J. Cody, A proposed radix and word length independent standard for floating-point arithmetic. *ACM SIGNUM Newsletter* **20** (1985) 37–51.

[4] W.J. Cody, J.T. Coonen, D.M. Gay, K. Hanson, D. Hough, W. Kahan, R. Karpinski, J. Palmer, F.N. Ris and D. Stevenson, A proposed radix-and-word-length-independent standard for floating-point arithmetic. *IEEE MICRO* **4** (1984) 86–100.
[5] E. Croot, R.-C. Li, and H.J. Zhu, The *abc* conjecture and correctly rounded reciprocal square roots. *Theor. Comput. Sci.* **315** (2004) 405–417.
[6] C.B. Dunham, Feasibility of "perfect" function evaluation. *SIGNUM Newsletter* **25** (1990) 25–26.
[7] S. Gal and B. Bachelis, An accurate elementary mathematical library for the IEEE floating point standard. *ACM Trans. Math. Software* **17** (1991) 26–45.
[8] A. Hurwitz, Über die angenäherte Darstellung der Irrationalzahlen durch rationale Brüche. *Math. Ann.* **46** (1891) 279–284.
[9] American National Standards Institute, Institute of Electrical, and Electronic Engineers. *IEEE standard for radix independent floating-point arithmetic*. ANSI/IEEE Standard, Std 854-1987, New York (1987).
[10] C. Iordache and D.W. Matula, On infinitely precise rounding for division, square root, reciprocal and square root reciprocal, in *Proceedings of the 14th IEEE Symposium on Computer Arithmetic (Adelaide, Australia)*, edited by Koren and Kornerup, IEEE Computer Society Press, Los Alamitos, CA (1999) 233–240.
[11] A.Ya. Khintchine, *Continued fractions*. Translated by Peter Wynn. P. Noordhoff Ltd., Groningen (1963).
[12] T. Lang and J.-M. Muller, Bound on runs of zeros and ones for algebraic functions, in *Proc. of the 15th IEEE Symposium on Computer Arithmetic (Arith-15)*, edited by Burgess and Ciminiera, IEEE Computer Society Press (2001).
[13] V. Lefèvre, *Moyens arithmétiques pour un calcul fiable*. Thèse, École normale supérieure de Lyon, Lyon, France (2000).
[14] V. Lefèvre and J.-M. Muller, Worst cases for correct rounding of the elementary functions in double precision, in *Proc. of the 15th IEEE Symposium on Computer Arithmetic (Arith-15)*, edited by Burgess and Ciminiera, IEEE Computer Society Press (2001).
[15] A.-M. Legendre, *Essai sur la théorie des nombres*. Duprat, Paris, An VI (1798).
[16] J. Liouville, Nouvelle démonstration d'un théorème sur les irrationnelles algébriques inséré dans le compte rendu de la dernière séance. *C.R. Acad. Sci. Paris, Sér. A* **18** (1844) 910–911.
[17] J. Liouville, Sur des classes très étendues de quantités dont la valeur n'est ni algébrique, ni même réductible à des irrationnelles algébriques. *C.R. Acad. Sci. Paris, Sér. A* **18** (1844) 883–885.
[18] J. Liouville, Sur des classes très étendues de quantités dont la valeur n'est ni algébrique, ni même réductible à des irrationnelles algébriques. *J. Math. Pures Appl.* **16** (1851) 133–142.
[19] K.F. Roth, Rational approximations to algebraic numbers. *Mathematika* **2** (1955) 1–20; corrigendum **168** (1955).
[20] D. Stehlé, V. Lefèvre and P. Zimmermann, Searching worst cases of a one-variable function using lattice reduction. *IEEE Trans. Comput.* **54** (2005) 340–346.