Arbitrary precision real arithmetic: design and algorithms

Valérie Ménissier-Morain

Arbitrary precision real arithmetic: design and algorithms

Valérie Ménissier-Morain*

*LIP6, Université Paris 6, 8 rue du Capitaine Scott, F-75015 Paris, France. *E-mail*: Valerie.Menissier-Morain@lip6.fr
Web page: http://calfor.lip6.fr/~vmm/

Abstract

We describe here a representation of computable real numbers and a set of algorithms for the elementary functions associated to this representation.

A real number is represented as a sequence of finite B-adic numbers and for each classical function (rational, algebraic or transcendental), we describe how to produce a sequence representing the result of the application of this function to its arguments, according to the sequences representing these arguments. For each algorithm we prove that the resulting sequence is a valid representation of the exact real result.

This arithmetic is the first abritrary precision real arithmetic with mathematically proved algorithms.

Résumé

Nous proposons une représentation des nombres réels calculables ainsi que des algorithmes pour les fonctions élémentaires usuelles pour cette représentation.

Un nombre réel est représenté par une suite de nombres B-adiques finis et pour chaque fonction classique (rationnelle, algébrique ou transcendante), nous montrons comment produire une suite représentant le résultat à partir de suites représentant les paramètres. Pour chacun de ces algorithmes nous démontrons que la suite qui en résulte représente bien le résultat réel exact.

Cette arithmétique est la première arithmétique réelle en précision arbitraire dotée d'un jeu complet d'algorithmes mathématiquement prouvés.

Contents

Chapter 1

Introduction

1.1 Motivation

We try to determine here what should be an arithmetic for a modern and reliable programming language.

The exactitude of the arithmetic is clearly an essential feature of a reliable programming language, but we show here that even a floating point representation with variable length, as it exists in symbolic computation softwares, is insufficient. Furthermore this arithmetic gives its users some wrong ideas, as disastrous as the round-off errors themselves. The floating point arithmetic [21] is obviously not an exact arithmetic: each partial result is systematically rounded off, and these successive round-off errors may lead to completely erroneous answers for ill-conditioned problems like the computation of $1/(y-x)$ where x is "much greater" than 1 (for example $x = 10^{20}$) and $y = x + 1$. For this particular computation, a single precision floating point computation induces an error due to a division by zero, but an exact computation yields a result, 1, that is furthermore exactly represented in a floating point arithmetic. However, in order to handle rational or real numbers, one represents them generally with the floating point arithmetic supplied by the computer with a fixed number of significant digits. The programmer is usually aware of these round-off problems, but nevertheless, he remains confident of the computed results because of "intuitive properties" of the floating point arithmetic. Among these pseudo-properties, one can cite:

1.. even if the result is not rigorously exact, it is certainly close to the exact result, that is to say that the round-off errors will be minor and in particular the order of magnitude is supposed to be preserved;

2.. a few floating point operations can only induce a slight inaccuracy in the computed result;

3.. if a result is computed with more digits, its accuracy will be better. More precisely, we hope that the distance between the computed value and the real value decreases according to the number of digits in the computation.

There are now numerous examples (see [38, 39, 14], etc.) where these common ideas are trampled on:

1. In [38], Jean-Michel Muller presents a rational sequence whose theoretical limit is 6, and after 10 (resp. 20) iterations for IEEE single (resp. double precision) the value of computed terms is always 100. The convergence to this wrong limit is fast and stable.

   The floating point arithmetic does not even preserve the order of magnitude of the limit. The floating point terms are rapidly irrelevant, even though the computation of these terms leads to very few operations. The speed of the convergence and the stability of the limit ensure neither the accuracy nor the order of magnitude of the computed limit.

   Any round-off error ejects the sequence out of the repulsive basin for 6 and project it in the attractive basin for 100.

2. In [39], Jean-Michel Muller presents a sequence for which the 25-th term is about 0.04, the result is about -10^{14} on a pocket computer and $+4 \times 10^9$ on a workstation.

3. In [14], Jean-Marie Chesneaux presents a second degree equation with a double root. According to the rounding mode, we obtain the real double root or two real roots or two complex roots.

1.1.1 Discussion

These examples might be considered as very particular problems, and thus we can consider that it is necessary for these special cases only to analyze the computation and the evolution of the precision without any modification of the arithmetic of the language.

But this analysis is so tedious that in practice the programmer ignores it.

In fact we have no easy way to recognize such problems and, furthermore, some traditional areas of floating point computations such as computer graphics and scientific computation show such problems in their usual practice. For example, in this second domain, algorithms produce positive definite matrices so a algorithm adapted to such matrices is used, but the floating point computation of the matrix may produce a not positive definite matrix so the result of the second algorithm is doubtful: at best it fails, it may even loop, at worst a completely wrong result is returned with almost no way to control its correctness. In other cases (for example the crash of the first try of Ariane 5), floating point computation fails because of a floating point overflow or underflow that the programmer didn't foresee. These areas are much concerned with money and safety (rockets, airplanes design structures).

Yet the programmer desires reliable arithmetic results, so the correction must be ensured by the programming language that should perform automatically a round-off error analysis to get back a real confidence in results obtained with floating point arithmetic.

1.1.2 Solutions

What solution can we envisage to use? The most usual solution to this question is interval analysis [41, 42, 1, 18]. The computation is performed using floating point arithmetic and propagates during all this computation an upper bound of the round-off error for rational operations, according to the IEEE-754 standard, so that one obtains at the end of the computation a floating point result 'and an upper bound for the round-off error on this result. Interval arithmetic can be only twice as expensive as ordinary floating point computations, at least in theory. This analysis indicates when the result is wrong. For the preceding sequences, a computation with such an arithmetic will indicate that the upper bound for the round-off error on this result is very big. However if we want to compute the exact result or a result as close as we want of the exact value and not only that the result of the floating point computation is completely wrong, this solution is not satisfactory.

We give here an answer to the programmer who needs reliable arithmetic results for which the correction is ensured by the programming language and not by an arithmetic that indicates an upper bound for the distance between the exact value and the obtained result. The programmer indicates an upper bound for the final round-off error and this bound is respected all along the computation, even if it requires a very high precision in a particular step of the computation. In some sense, this analysis is the reciprocal analysis of what would be a complete interval analysis (that is to say that concerns any classical elementary function).

The ML language was designed for safe programming and so should ensure safety in numerical programming. So we implement a small prototype of an arbitrary precision library in this language according to the work described here.

1.2 History of the problem

1.2.1 Two interesting attempts

In 1972, Bill Gosper [22] described in a small internal note how to perform rational operations on continued fractions and computes 17,000,000 terms of the continued fraction for π to discover a possible pattern.

Brent in [10] described in 1976 algorithms to compute quickly multiple-precision evaluations of elementary functions, but he did not consider real numbers as full members of a language. He compute the image of a floating-point number x (thus a rational number) by an algebraic or transcendental function f to a precision $\mathcal{O}(2^{-n})$. This work is however interesting from a practical point of view.

1.2.2 Sequences of redundant digits

Wiedmer proposed in 1977 a solution for real number computations in [55, 56, 57]. This solution can be considered as unbounded on-line arithmetic. However, Wiedmer proposed only an algorithm to add real numbers. These ideas were studied again and extended by Boehm in [8, 7]. Computable real numbers are represented by an infinite sequence of digits in a given base B. For such a representation the digits of the results are produced "from left to right", beginning with the most significant digits, in opposition to the usual algorithms for addition and multiplication for example,

but this technique is common for on-line arithmetic. Particularly, Avizienis in [2] and Wiedmer in [57] proved that an addition algorithm "from left to right" implies the redundancy of the representation: for example, digits are in the integer interval $[-B+1, B-1]$ rather than the classical interval $[0, B-1]$. The idea is that it is necessary to anticipate what will be the next digits of the arguments of the addition algorithm and for example to overestimate by one the absolute value of the sum, even if one needs to correct this trend on the next produced digit by a negative sign. We studied this representation, described and proved algorithms for rational operations, but we did not work out so far algorithms for transcendental functions. Perhaps the Cordic algorithms described by Lin and Sips in [33] may be used to compute these functions. The incrementality is a natural good point for this representation: if one need some more digits, one starts from the list of already computed digits rather than from the beginning of the computation. However, apart from the lack of well-integrated algorithms for transcendental functions, the algorithms for rational operations are intricate and rather inefficient.

1.2.3 Computable Cauchy sequences

In [8, 7], Boehm studied a more natural representation. This representation is designed for almost automatic evaluation of round-off errors in programs written in Fortran. In his implementation, the classical operations on floating point numbers are transparently replaced by exact operations on real numbers, then some numerical tests of small size are performed with each arithmetic, so that if a floating point result does not correspond to the expected value, one can attribute this computation error either to a round-off error if the real result is correct or to an error in the implementation of the algorithm by the program if the real result is wrong. Boehm describes algorithms for addition and multiplication on this representation.

Boehm developed an implementation for each of these two representations. The comparison of the running times indicates clearly that the second one is much faster than the first one.

We studied this second representation and now we propose a complete and entirely proved set of algorithms for all elementary functions. This work leads to an implementation in the Caml implementation of the ML language.

1.2.4 Continued fractions

Finally, in [51, 52, 53], Vuillemin interprets Bill Gosper's work on the continued fractions arithmetic (essentially rational operations) [22] and represents real numbers by continued fractions, with the underlying idea that continued fractions are the "closest" rational numbers to the real numbers. However, apart from the fact that these algorithms are principally not proved, this representation is rather inadequate to the architecture of current computers so it is inefficient. We implement a complete prototype for this representation that exhibits poor running times despite the natural incrementality of the method.

1.2.5 And now?

We present these three representations with all details in [37]. We describe completely in this report the second representation mentioned above. Since this work take place in 1994, we present before only the state of the art in 1994 and the later state of the art is postponed after the description of this work.

1.2.6 Plan of the report

We first recall the main properties of computable real numbers. We deduce from one definition, among the three definitions of this notion, a representation of these numbers as sequence of finite B-adic numbers and then we describe algorithms for rational operations and transcendental functions for this representation. Finally we describe briefly the prototype written in Caml.

Chapter 2

Computable real numbers

Unlike Bishop [5], Martin-Löf [35] and Stolzenberg [47], we use here the classical real analysis as the framework to state properties of computable real numbers*, as in Rice fundamental paper [45].

In this section we will use the Gödel-Kleene representation of recursive functions as μ-recursive functions [54, 17, 29, 28, 15].

2.1 Definitions

Let us note $f_{\mathbb{Q}\to\mathbb{N}}$ a recursive bijection from $\mathbb{Q}$ to $\mathbb{N}$. We define first the notion of *recursive Cauchy sequence* for rational numbers and for intervals with rational bounds.

Definition 1 (Recursive Cauchy sequence)

1.. *A sequence of rational numbers* $(q_n)_{n\in\mathbb{N}}$ *is called* recursively enumerable *if the function* $n \mapsto f_{\mathbb{Q}\to\mathbb{N}}(q_n)$ *is recursive.*

2.. *A sequence of intervals with rational bounds* $(I_n = [i_n, s_n])_{n\in\mathbb{N}}$ *is called* recursively enumerable *if the sequences* $(i_n)_{n\in\mathbb{N}}$ *and* $(s_n)_{n\in\mathbb{N}}$ *are recursively enumerable sequences of rational numbers.*

3.. *A sequence of rational numbers* $(q_n)_{n\in\mathbb{N}}$ *is called* a recursive Cauchy sequence *if it is recursively enumerable and there exists a recursive function* g*, the* convergence function *of the sequence, such that for any strictly positive integer* N *and all pair of integers* n *and* m *with* $n \geq m \geq g(N)$*, we have:*

$$|q_n - q_m| < \frac{1}{N}.$$

We can now give several definitions of the notion of *computable real numbers.*

Definition 2 (Computable real numbers, first definition)
A real number r *is a* computable real number *if and only if there exists a recursively enumerable sequence* $(I_n = [i_n, s_n])_{n\in\mathbb{N}}$ *of nested intervals with rational bounds, enclosing* r *and the sequence of the lengths of these intervals* $(|s_n - i_n|)_{n\in\mathbb{N}}$ *converges to 0.*

Definition 3 (Computable real numbers, second definition)
A real number is a computable real number *if and only if it is the limit of a recursive Cauchy sequence of rational numbers.*

These two definitions lead naturally to the following property:

Property 1 *These two definitions of the notion of computable real numbers are equivalent.*

*We use here "computable real numbers as Turing did [48] rather than the expression "recursive real numbers" employed by Rice, but these two terms are of course equivalent

Proof.

Let r be a real number enclosed in any interval of a recursively enumerable sequence $([i_n, s_n])_{n\in\mathbb{N}}$ of nested intervals with rational bounds and length decreasing to 0.

This number is the limit of a recursively enumerable sequence $(i_n)_{n\in\mathbb{N}}$ since for any $n \in \mathbb{N}$ we have: $0 \leq r - i_n \leq s_n - i_n$ and $\lim_{n\to\infty}(s_n - i_n) = 0$.

Furthermore let g be the function such that for any $N \in \mathbb{N}$ we have the definition equality

$$g(N) = \mu\, n \left[(s_n - i_n) < \frac{1}{N}\right],$$

this is a recursive function (unbound total μ scheme total since the sequence $(s_n - i_n)_{n\in\mathbb{N}}$ is decreasing to 0, according to the classical denomination about computable functions).

Moreover g is a convergence function for the sequence $(i_n)_{n\in\mathbb{N}}$ since if $n \geq m \geq g(N)$, we have

$$|i_n - i_m| \leq s_m - i_m < \frac{1}{N}.$$

Consequently the sequence $(i_n)_{n\in\mathbb{N}}$ is a recursive Cauchy sequence and r is the limit of a recursively enumerable sequence of rational numbers.

Reciprocally, let r be the limit of a recursive Cauchy sequence of rational numbers $(r_n)_{n\in\mathbb{N}}$ with convergence function g. Then we define for any $n \in \mathbb{N}$, i_n and s_n by the following equations

$$\begin{aligned} i_n &= r_{g(n+1)} - \frac{1}{n+1} \\ s_n &= r_{g(n+1)} + \frac{1}{n+1}. \end{aligned}$$

The function g is recursive, hence the sequences $(i_n)_{n\in\mathbb{N}}$ and $(s_n)_{n\in\mathbb{N}}$ are recursively enumerable.

Furthermore for any $n \in \mathbb{N}$, when taking the limit on the first index up to infinity in the definition inequation of a recursive Cauchy sequence, we have

$$\left|r - r_{g(n+1)}\right| \leq \frac{1}{n+1},$$

consequently r is inclosed in $[i_n, s_n]$ and $s_n - i_n$ is equal to $2/(n+1)$ and finally the length of the interval $[i_n, s_n]$ converges to 0. □

We will now define the related notion of *finite B-adic numbers* for a given base B and deduce the notion of *B-approximable real number.*

Definition 4 (Finite B-adic number)

If B is an integer greater than or equal to 2, a rational number r is called a finite B-adic number *if there exists two integers p and q such that $r = p/B^q$ and q is a positive integer.*

This definition generalizes the notion of *dyadic* number. We define now the notion of *B-approximable real number.*

Definition 5 (B-approximable real number)

A real number x is called B-approximable *if there exists a recursive function g such that, for any integer N, $g(N)$ is a finite B-adic number and*

$$|x - g(N)| < \frac{1}{B^N}.$$

This third definition leads naturally to the following property:

Property 2 *The notions of computable real numbers and B-approximable real number are equivalent.*

Proof.

We will prove that the second definition of computable real numbers is equivalent to the definition of B-approximable real numbers.

Let B be an integer greater than or equal to 2 and r a computable real number. Let $(r_n)_{n\in\mathbb{N}}$ be a recursive Cauchy sequence of rational numbers with limit equal to and h be the convergence function of this sequence. Let N be an integer, we have for any $n, m \geq h(2B^N)$,

$$|r_n - r_m| < \frac{1}{2B^N}.$$

Thus as n tends to infinity, we have for any $n \geq h(2B^N)$,

$$|r_n - r| \leq \frac{1}{2B^N}.$$

Let p_n (resp. q_n) be the numerator (resp. the denominator) of $r_n = p_n/q_n$, we define $k = \max(0, n - \lfloor \log_B q_n \rfloor)$ and we write the Euclidean division of $p_n B^k B^n$ by $B^k q_n$: $p_n B^k B^n = B^k q_n p d_n + p r_n$ with $0 \leq p r_n < |B^k q_n|$ and then we obtain

$$\left|\frac{p_n}{q_n} - \frac{p d_n}{B^n}\right| = \left|\frac{p_n B^k}{q_n B^k} - \frac{p d_n}{B^n}\right| = \left|\frac{(p_n B^k)B^n - p d_n (q_n B^k)}{B^n (B^k q_n)}\right| = \frac{1}{B^n}\frac{p r_n}{|B^k q_n|} < \frac{1}{B^n}$$

and if $n \geq \max(h(2B^N), N+1)$, then

$$\left|\frac{p d_n}{B^n} - r\right| < \frac{1}{2B^N} + \frac{1}{B^{N+1}} \leq \frac{1}{B^N}.$$

We construct the rational number $x_n = p d_n / B^n$ as a finite B-adic number. Furthermore the function $n \mapsto p d_k / B^k$ with $k = \max(h(2B^n), n+1)$ is recursive thus r is a B-approximable real number.

Reciprocally, if r is a B-approximable real number, let us prove that r is a computable real number. let g be the recursive function associated to r. We define, for any integer N, $h(N) = \lceil \log_B N + 1 \rceil$. The function defined in this way is a recursive one (unlimited total μ scheme) and we have for any strictly positive integer N and any $n, m \geq h(N)$,

$$|g(n) - g(m)| \leq |g(n) - r| + |r - g(m)| \leq \frac{1}{B^n} + \frac{1}{B^m} < \frac{2}{B(N+1)} \leq \frac{1}{N}.$$

The sequence $(g(n))_{n\in\mathbb{N}}$ converges to r by definition thus r is a computable real number. □

And now a last definition that will be useful afterwards.

Definition 6 (Recursively enumerable sequence of real numbers)

A sequence $(x_n)_{n\in\mathbb{N}}$ of computable real numbers is called recursively enumerable *if there exists two recursive functions g and h such that $(g(n,k))_{k\in\mathbb{N}}$ is a recursive Cauchy sequence of rational numbers that converges to x_n as k tends to infinity, with $k \mapsto h(n,k)$ as convergence function.*

2.2 Properties of $\mathcal{R}$

We denote up to the end of this section by $\mathcal{R}$ (resp. $\mathcal{C}$) the set of computable real (resp. complex) numbers and as usual by $\mathbb{R}$ (resp. $\mathbb{C}$) the set of real (resp. complex) numbers.

The set $\mathcal{R}$ has the following properties:

2.2.1 Algebraic properties

Theorem 3

1.. *The set $\mathcal{R}$ with the addition and the multiplication of $\mathbb{R}$ is an Archimedean commutative field.*

2.. *The set $\mathcal{C}$ with the addition and the multiplication of $\mathbb{C}$ is an algebraically closed commutative field.*

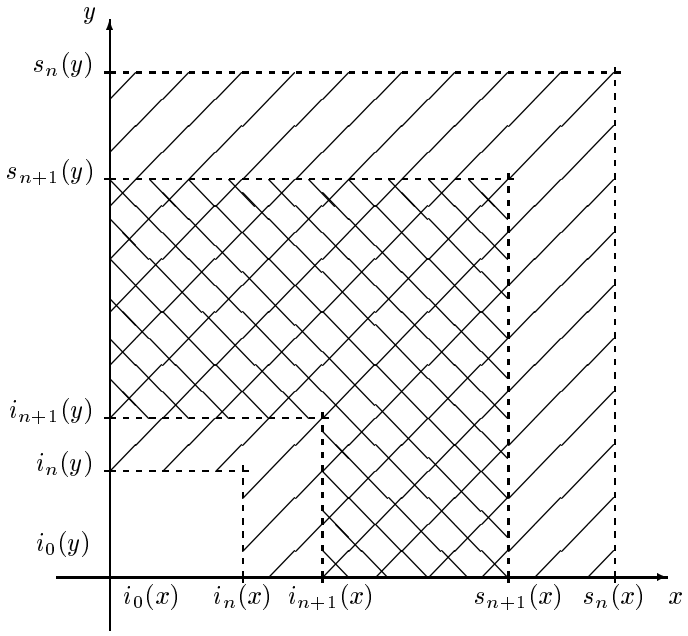


Figure 2.1: The case of multiplication

Proof.

First of all, $\mathcal{R}$ and $\mathcal{C}$ are respectively a subset of $\mathbb{R}$ and $\mathbb{C}$ including $\mathbb{Q}$ by definition.

Let x and y be two elements of $\mathcal{R}$ defined, according to the first definition of computable real number, by the sequence $([i_n(x), s_n(x)])_{n\in\mathbb{N}}$ and $([i_n(y), s_n(y)])_{n\in\mathbb{N}}$.

We have, for any $n \in \mathbb{N}$, $x + y \in [i_n(x) + i_n(y), s_n(x) + s_n(y)]$. Moreover, the sequence $([i_n(x) + i_n(y), s_n(x) + s_n(y)])_{n\in\mathbb{N}}$ is a recursively enumerable sequence of nested intervals with rational bounds and the length of the n-th interval $([i_n(x) + i_n(y), s_n(x) + s_n(y)])$ is

$$|(i_n(x) + i_n(y)) - (s_n(x) + s_n(y))| \leq |s_n(x) - i_n(x)| + |s_n(y) - i_n(y)|$$

so tends to 0 and $x + y$ is a computable real number.

We have, for any $n \in \mathbb{N}$, $-x \in [-s_n(x), -i_n(x)]$. Furthermore, the sequence $([-s_n(x), -i_n(x)])_{n\in\mathbb{N}}$ is a recursively enumerable sequence of nested intervals with rational bounds and the length of the n-th interval $([-s_n(x), -i_n(x)])$ is $s_n(x) - i_n(x)$ so tends to 0 and $-x$ is a computable real number.

Consequently, $\mathcal{R}$ is a subgroup of $\mathbb{R}$. From the definition of addition on $\mathbb{C}$ using addition on $\mathbb{R}$, it follows immediately that $\mathcal{C}$ is a subgroup of $\mathbb{C}$.

We will now consider the case of multiplication. We might present a complete proof for this case, but it will be technical and boring. We chose here to present only the idea of this proof.

If we move the origin of the coordinates from $(0, 0)$ to $(i_0(x), i_0(y))$, we can only consider the case of sequences of positive rational numbers (it is precisely this shift that complicates the proof even if it is natural with a geometrical point of view). Then $([i_n(x) \times i_n(y), s_n(x) \times s_n(y)])_{n\in\mathbb{N}}$ is a recursively enumerable sequence of nested intervals with rational bounds and we interpret graphically these intervals as the surface of the juxtaposition of tree rectangles joined in a "L" form (hatched on figure 2.1). These juxtapositions are nested and their surface is decreasing with a null limit. Thus these intervals are nested, with length converging to zero and xy is a computable real number.

We deduce from this that $\mathcal{R}$ is a subring of $\mathbb{R}$. Furthermore, multiplication on $\mathbb{C}$ is simply defined using addition and multiplication on $\mathbb{R}$, thus $\mathcal{C}$ is a subring of $\mathbb{C}$.

We will now consider the case of inversion.

Let x be a computable real number different of zero, then there exists $k \in \mathbb{N}$ such that for any $n \geq k$ the n-th

interval $[i_n(x), s_n(x)]$ does not contain 0. Thus we have

$$\frac{1}{s_n(x)} \leq \frac{1}{x} \leq \frac{1}{i_n(x)}$$

thus $1/x \in [1/s_n(x), 1/i_n(x)]$. Furthermore, the sequence $([1/s_n(x), 1/i_n(x)])_{n\geq k}$ is a recursively enumerable sequence of nested intervals with rational bounds and if ε is a positive real number lesser than x, then theres exists $k' \in \mathbb{N}$ such that for any $n \geq k'$, we have $\min(|i_n(x)|, |s_n(x)|) \geq |s_{k'}(x)| - \varepsilon > 0$ et $|s_n(x) - i_n(x)| \leq \varepsilon$, thus the length of the interval $[1/s_n(x), 1/i_n(x)]$ is

$$\frac{1}{i_n(x)} - \frac{1}{s_n(x)} \leq \frac{\varepsilon}{(|s_{k'}(x)| - \varepsilon)^2},$$

thus tends to zero and $1/x$ is a computable real number.

We deduce from this that $\mathcal{R}$ is a subfield of $\mathbb{R}$. Furthermore, inversion in $\mathbb{C}$ is simply defined using addition, multiplication and inversion in $\mathbb{R}$, thus $\mathcal{C}$ is a subfield of $\mathbb{C}$.

Moreover, just as commutativity is stable by subset, $\mathbb{R}$ is a totally ordered field thus $\mathcal{R}$ is a totally ordered field. Finally if a and b are two computable real numbers with a greater than zero, respectively represented by the sequences of intervals $([i_n(a), s_n(a)])_{n\in\mathbb{N}}$ and $([i_n(b), s_n(b)])_{n\in\mathbb{N}}$, then there exists an integer k such that $i_k(a) > 0$. Let

$$n = \left\lceil \frac{|s_0(b)|}{i_k(a)} \right\rceil + 1,$$

we have

$$n\ a \geq n\ i_k(a) > |s_0(b)| \geq s_0(b) \geq b,$$

and $\mathcal{R}$ is Archimedean.

Finally, in [45], Rice proved that $\mathcal{C}$ is algebraically closed by adaptation of a classical proof on $\mathbb{C}$ to its subset $\mathcal{C}$. □

2.2.2 Topological properties

We have for computable real (complex) numbers similar results of completeness by making constructive the notion of Cauchy sequence, that is to say, using recursive Cauchy sequences of real numbers instead of classical Cauchy sequences. Precisely we define the notion of recursively enumerable sequence of computable real (resp. complex) numbers as follows:

Definition 7 (Recursively enumerable sequences of computable real or complex numbers)

1.. *A sequence $(x_n)_{n\in\mathbb{N}}$ of computable real numbers is said recursively enumerable if there exists two recursive functions g and h such that $(g(n,k))_{k\in\mathbb{N}}$ is a recursive Cauchy sequence of rational numbers that tends to x_n when k tends to infinity, with $k \mapsto h(n,k)$ as convergence function.*

2.. *A sequence $(x_n)_{n\in\mathbb{N}}$ of computable complex numbers is said recursively enumerable if the sequences $(\Re(x_n))_{n\in\mathbb{N}}$ and $(\Im(x_n))_{n\in\mathbb{N}}$ are recursively enumerable sequences of computable real numbers.*

We define the notion of recursive Cauchy sequences of computable real numbers as for rational numbers. For computable complex numbers, we turn absolute values into modules.

Theorem 4 *The limit of any recursive Cauchy sequence of computable real numbers is a computable real number.*

Proof.

Let ℓ and x design respectively the convergence function and the limit of a sequence $(x_n)_{n\in\mathbb{N}}$ of computable real numbers, then let us prove that the sequence $(y_n)_{n\in\mathbb{N}}$ with

$$y_n = g(n, h(n,n))$$

is a recursive Cauchy sequence of rational numbers with convergence function

$$N \mapsto \max(\ell(3N), 3N)$$

and limit x: let $n, m \geq \max(\ell(3N), 3N)$, then we have

$$|g(n, h(n,n)) - g(m, h(m,m))| \leq |g(n, h(n,n)) - x_n| + |x_n - x_m| + |x_m - g(m, h(m,m))|.$$

But $n, m \geq \ell(3N)$, thus

$$|x_n - x_m| < \frac{1}{3N}$$

and by definition of h and because $n, m \geq 3N$, we have:

$$\begin{array}{l} |g(n, h(n,n)) - x_n| < \frac{1}{n} < \frac{1}{3N} \\ |g(m, h(m,m)) - x_m| < \frac{1}{m} < \frac{1}{3N} \end{array}$$

thus

$$|g(n, h(n,n)) - g(m, h(m,m))| < \frac{1}{N},$$

and x is a computable real number. □

Corollary 1 *The limit of any recursive Cauchy sequence of computable complex number is a computable complex number.*

2.2.3 Functions class with computable real (or complex) values

Since $\mathcal{R}$ is a field, we already know that $\mathcal{R}$ is closed for addition, opposite, multiplication and inverse of a number different from zero. We will now prove that in fact $\mathcal{R}$ is closed for any elementary function. To prove this, we first establish the following lemma:

Lemma 1 *Let $(a_n)_{n\in\mathbb{N}}$ be a recursively enumerable sequence of positive rational numbers decreasing to zero, the sum of the alternated series of general term $(-1)^n a_n$ is a computable real number.*

Proof.
Let us name

$$\begin{aligned} s &= \sum_{k\in\mathbb{N}} (-1)^k a_k \\ s_n &= \sum_{k=0}^{k=n} (-1)^k a_k, \end{aligned}$$

then the sequence $([s_{2n+1}, s_{2n}])$ is a recursively enumerable sequence of nested intervals including s with rational bounds. The length of these nested intervals is decreasing to 0 thus s is a computable real number.

We apply it now to elementary functions. We start with complex exponential.

If x is a rational negative number, then e^x is the sum of $n = \lfloor |x| \rfloor$ rational numbers (the n first terms of the Taylor series of exp: $x^i/i!$) and the limit of an alternated series (the sign of $x^i/i!$ alternates and this term decreases to zero from rank k to infinity, thus the series beginning to the $n+1$-th term of this series is an alternated series) with an infinite convergence domain thus is a computable real number. If x is a rational positive number, we have $\mathrm{e}^x = \frac{1}{\mathrm{e}^{-x}}$ and e^{-x} is a computable real number different from zero so its inverse e^x is also a computable real number. Finally, if x is a computable real number, for example the limit of the recursive Cauchy sequence of rational numbers $(x_n)_{n\in\mathbb{N}}$, then $\exp(x)$ is the limit of the recursive Cauchy sequence of computable real numbers $(\exp(x_n))_{n\in\mathbb{N}}$ since exp is a continue function and $\exp(x)$ is a computable real number according to theorem 4.

Let x be a computable real number, $\sin x$ and $\cos x$ are defined by an alternated series (or the opposite of an alternated series for the sine function if x is a negative number) thus are computable real numbers.

If $a + ib$ is a computable complex number, then

$$\begin{aligned} \mathrm{e}^{a+ib} &= \mathrm{e}^a(\cos(b) + i\sin(b)) \\ \sin(a+ib) &= \frac{\mathrm{e}^{-b}(\cos(a) + i\sin(a)) - \mathrm{e}^b(\cos(a) - i\sin(a))}{2i} \\ \cos(a+ib) &= \frac{\mathrm{e}^{-b}\cos(a) + i\sin(a)) + \mathrm{e}^b(\cos(a) - i\sin(a))}{2} \end{aligned}$$

are also computable complex numbers. We have proved that $\mathcal{R}$ and $\mathcal{C}$ are closed for exp.

Let us now prove that $\mathcal{R}$ is closed for log and arctan and we will then deduce that $\mathcal{C}$ is closed for log too.

If x is a computable real number between 1 and 2, $\log x$ is the sum of an alternated Taylor series (in $x-1$)thus is a computable real number. If x is a real number between 2^n and 2^{n+1}, with $n \in \mathbb{Z}$, then

$$\log x = \log\left(\frac{x}{2^n}\right) + n\log(2).$$

But $\log(\frac{x}{2^n})$ and $\log(2)$ are computable real numbers according to the preceding case and then $\log x$ is a computable real number.

If x is a computable real number between -1 and 1, then $\arctan(x)$ is defined by an alternated series with absolute value decreasing to zero thus is a computable real number (consequently $\pi = 4 \times \arctan(1)$ is a computable real number). If x is a computable real number greater than 1 (resp. lesser than -1) then $\arctan(x) = \pi/2 - \arctan(1/x)$ (resp. $\arctan(x) = -\pi/2 - \arctan(1/x)$) and $1/x$ is a computable real number between -1 and 1, thus $\arctan(x)$ is a computable real number.

We deduce from this that the and the principal argument of a computable complex number different of zero $a+ib$,

$$|a+ib| = \sqrt{a \times a + b \times b} = \exp\left(\frac{\log(a \times a + b \times b)}{2}\right)$$

and

$$Arg(a+ib) = 2\arctan\left(\frac{b}{a+|a+ib|}\right)$$

are computable real numbers.

Consequently, for any computable complex number with a non null imaginary part or with a strictly positive real part $z = \rho e^{i\theta}$, $\log(z) = \log(\rho) + i\theta$ is a computable complex number and $\mathcal{C}$ is closed for the principal determination of log on $\mathbb{C}$.

Furthermore $2i\pi \in \mathcal{C}$, thus other determinations of the argument and of the logarithm of a non null computable complex number are computable complex numbers. □

And we conclude by the following general theorem:

Theorem 5 *$\mathcal{C}$ is closed for elementary functions.*

and its corollary on $\mathcal{R}$:

Corollary 2 *$\mathcal{R}$ is closed for* exp, log, $(x,y) \mapsto x^y$, sin, cos, tan, sinh, cosh, tanh, arcsin, arccos, arctan, arcsinh, arccosh, arctanh.

Proof.

An elementary function is defined on an open subset of the complex plane from the rational operations, the exponential and the logarithm functions. But $\mathcal{C}$ is closed for these operations and the composition of functions preserves this property and we deduce the general result.

About its real corollary, the concerned functions map real numbers to real images thus the result is obtained by projection on $\mathcal{R}$. □

2.2.4 Cardinal of $\mathcal{R}$ and $\mathcal{C}$

We will now prove that, even if $\mathcal{R}$ contains every "interesting" number, $\mathcal{R}$ is a strict subset of $\mathbb{R}$ by examination of its cardinality.

Theorem 6 *$\mathcal{R}$ is a denumerable subset of $\mathbb{R}$ and is dense in $\mathbb{R}$.*

Proof.

First of all, $\mathbb{Q} \subset \mathcal{R}$, thus $\mathcal{R}$ is at least denumerable and dense in $\mathbb{R}$. Now let r be a computable real number represented by the recursive Cauchy sequence $(r_n)_{n\in\mathbb{N}}$, we define the recursive function $g : \mathbb{N} \mapsto \mathbb{N}$ by the formula $g(n) = f_{\mathbb{Q}\to\mathbb{N}}(r_n)$ and finally the Gödel index of this function. We have build in this way an injective function from $\mathcal{R}$ to $\mathbb{N}$ thus $\mathcal{R}$ is at the most a denumerable set and finally $\mathcal{R}$ is a denumerable and dense subset of $\mathbb{R}$. □

Corollary 3 *$\mathcal{C}$ is a denumerable subset of $\mathbb{C}$.*

Proof.

Let g be a bijection from $\mathcal{R}$ to $\mathbb{N}$ and h be a bijection from $\mathbb{N}^2$ to $\mathbb{N}$, then $z \mapsto h(g(\Re(z)), g(\Im(z)))$ is a bijection from $\mathcal{C}$ to $\mathbb{N}$ and $\mathcal{C}$ is denumerable. □

2.3 Indecidability theorems about $\mathcal{R}$

Rice proved the following result:

Theorem 7 (Rice)
There exists no general algorithm to determine whether a computable real number is zero or not.

Proof (*Principle of the proof*).

As a first point, if such an algorithm exists, then we can decide for all recursive function from $\mathbb{N}$ into $\{0,1\}$ if this function is single-valued or not. As a second point, the question of the stopping of a Turing machine can be describe by the question of the single-valuation of a recursive function from $\mathbb{N}$ into $\{0,1\}$. Then since we know that the question of the stopping of a Turing machine is undecidable, the result follows. □

We deduce from this the following results:

Corollary 4 *1.. There exists no general algorithm to determine the image of a computable real number by a function with a discontinuity at this point.*

2.. There exists no general algorithm to determine if a computable real number is greater than another one.

3.. There exists no general algorithm to determine the integer part of a computable real number.

4.. There exists no general algorithm to determine if a computable real number is rational.

Proof.

If a method exists to determine the image of a computable real number by a function with a discontinuity at this point, it will be possible to evaluate in every point the characteristic function of zero that is refuted by Rice's theorem expressed before.

The three next properties are immediate by application of this property to boolean functions on real numbers, as follows:

2.. g_b maps any computable real number a to the boolean value $a \geq b$,

3.. g maps any computable real number r to its integer part $\lfloor r \rfloor$,

4.. g maps any computable real number r to the boolean value $r \in \mathbb{Q}$.

□

A consequence of the third proposition of this corollary is that one cannot determine exactly the classical continued fraction expansion or the development in a given base of any computable real number.

However, one should not attach an excessive importance to these impossibilities because according to the last definition of computable real numbers, any computable real numbers may be known to a precision within B^{-n} in a given base B for any integer n. As far as the comparison is concerned, the following theorem establishes that if two computable real numbers differ, then there is an algorithm that indicate which one is the greater one and at which rank their definition sequences differ.

Property 8 *Let $A = (a_n)_{n\in\mathbb{N}}$ and $B = (b_n)_{n\in\mathbb{N}}$ be two recursive Cauchy sequences of rational numbers with distinct respective limits a and b, then there exists an algorithm to compare a and b that terminates.*

Proof.

Since a and b are distinct, there exists $k(A,B)$ such that $|a-b| > 4/k(A,B)$. Let g and h be the convergence functions respectively of the sequences $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$. We have

$$\frac{4}{k(A,B)} < |a-b| < \left|a - a_{g(k(A,B))}\right| + \left|a_{g(k(A,B))} - b_{h(k(A,B))}\right| + \left|b_{h(k(A,B))} - b\right|.$$

But if $n \geq g(k(A,B))$ and $m \geq h(k(A,B))$, we have $\left|a_n - a_{g(k(A,B))}\right| < \frac{1}{k(A,B)}$ and $\left|b_{h(k(A,B))} - b_n\right| < \frac{1}{k(A,B)}$. When n tends to infinity, we obtain $\left|a - a_{g(k(A,B))}\right| \leq \frac{1}{k(A,B)}$ and $\left|b_{h(k(A,B))} - b\right| \leq \frac{1}{k(A,B)}$, thus we deduce:

$$\frac{4}{k(A,B)} < |a-b| \leq \frac{2}{k(A,B)} + \left|a_{g(k(A,B))} - b_{h(k(A,B))}\right|,$$

and finally $\left|a_{g(k(A,B))} - b_{h(k(A,B))}\right| > \frac{2}{k(A,B)}$.

Consequently there exists n indices such that $\left|a_{g(n)} - b_{h(n)}\right| > \frac{2}{n}$, $|a - a_{g(n)}| \leq \frac{1}{n}$ and $|b - b_{h(n)}| \leq \frac{1}{n}$. If n is such an integer, we have $a - b = a - a_{g(n)} + a_{g(n)} - b_{h(n)} + b_{h(n)} - b$ and

$$\left|a - a_{g(n)}\right| + \left|b_{h(n)} - b\right| \leq \frac{2}{n} < \left|a_{g(n)} - b_{h(n)}\right|$$

thus $a - b$ has same sign that $a_{g(n)} - b_{h(n)}$. It is sufficient to choose

$$n(A,B) = \mu m \left[\left|a_{g(m)} - b_{h(m)}\right| > \frac{2}{m}\right]$$

that defines a recursive function on the sequences A and B (total unbound μ scheme). The fact that comparison between rational numbers is recursive terminates the demonstration. □

Chapter 3

Description of a representation of computable real numbers with particular sequences of B-adic numbers

According to the third definition, computable real numbers are considered here as B-approximable real numbers. Precisely reals numbers will be represented by B-adic numbers and as in Boehm's work, we represent the B-adic numbers by longer and longer integer corresponding to the numerator of B-adic approximations more and more precise.

It is well-known that the limit of a sequence of B-adic numbers $(a_n/B^{k_n})_{n\in\mathbb{N}}$ is also the limit of this other sequence of B-adic numbers $(b_n/B^n)_{n\in\mathbb{N}}$ with $b_n = \lfloor a_n B^{n-k_n} \rfloor$. This second sequence is interesting because the denominator of each B-adic is exactly B raised to its rank in the sequence so we need only the sequence of integers $(b_n)_{n\in\mathbb{N}}$ to represent the limit of this sequence.

Thus we approximate a computable real number r with a sequence of integers $(c_n)_{n\in\mathbb{N}}$ such that $|r - c_n B^{-n}| < B^{-n}$ for any integer n.

We present now precisely the definitions and general properties of this representation to prepare the algorithms for elementary functions for this representation.

Let B be a given base, i.e. an integer greater than or equal to 2. A computable real number is represented by a sequence of integers that satisfy the following property:

Definition 8 (Bounds property)
Let x be a computable real number, for any integer p, the bounds property *of x by p for order n is characterized by the following inequality $|x - pB^{-n}| < B^{-n}$ i.e. $(p-1)B^{-n} < x < (p+1)B^{-n}$.*

We authorize negative indices for practical reasons because sometimes we need only to know the order of magnitude of a real number rather than its integer part. The bounds property apply easily to negative indices and it saves some time during the computation. We will now express some properties of the integers that satisfy the bounds property for a given real number and a given order.

Property 9 *Let x be a computable real number, n an integer and p be an integer. Suppose that the bounds property of x by p for order n is satisfied. Then $p = \lfloor B^n x \rfloor$ or $p = -\lfloor B^n(-x) \rfloor$. Furthermore if $B^n x$ is an integer then $p = \lfloor B^n x \rfloor$.*

Proof.
First of all, we will prove that $\lfloor B^n x \rfloor$ and $-\lfloor B^n(-x) \rfloor$ satisfy the bounds property of x for order n.

We have $\lfloor B^n x \rfloor \leq B^n x < \lfloor B^n x \rfloor + 1$, thus $\lfloor B^n x \rfloor B^{-n} \leq x < (\lfloor B^n x \rfloor + 1)B^{-n}$ and $\lfloor B^n x \rfloor$ satisfies *a fortiori* the bounds property of x for order n.

Just as for $\lfloor B^n x \rfloor$, we have $\lfloor B^n(-x) \rfloor \leq B^n(-x) < \lfloor B^n(-x) \rfloor + 1$, so we deduce that $\lfloor B^n(-x) \rfloor B^{-n} \leq (-x) < (\lfloor B^n(-x) \rfloor + 1)B^{-n}$ and $(-\lfloor B^n(-x) \rfloor - 1)B^{-n} < x \leq (-\lfloor B^n(-x) \rfloor)B^{-n}$ and, consequently $-\lfloor B^n(-x) \rfloor$ satisfies also the bounds property of x for order n.

Now, let p be an integer that satisfies the bounds property of x for order n, we will prove that $p = \lfloor B^n x \rfloor$ or $p = \lfloor B^n x \rfloor + 1$.

From $\lfloor B^n x \rfloor B^{-n} \leq x$ and $x < (p+1)B^{-n}$, we deduce that $\lfloor B^n x \rfloor < p+1$ and by combining the inequalities $x < (\lfloor B^n x \rfloor + 1)B^{-n}$ and $(p-1)B^{-n} < x$, we obtain $p - 1 < \lfloor B^n x \rfloor + 1$. Consequently p satisfies the inequality $\lfloor B^n x \rfloor \leq p \leq \lfloor B^n x \rfloor + 1$ that is to say that $p = \lfloor B^n x \rfloor$ or $p = \lfloor B^n x \rfloor + 1$. Furthermore if $B^n x$ is not an integer, then $\lfloor B^n x \rfloor + 1 = -\lfloor B^n(-x) \rfloor$ and if $B^n x$ is an integer $\lfloor B^n x \rfloor = -\lfloor B^n(-x) \rfloor$. But in this case $\lfloor B^n x \rfloor + 1$ doesn't verify the bounds property of x for order n, since the left inequality is not a strict inequality. □

The real numbers that we will consider in this section are computable real numbers x represented by sequences of integers $(x_n)_{n\in\mathbb{N}}$ such that the bounds property for x by x_n for order n is satisfied.

We will now define the *sign* function before we express the following property that describes the relations between the integers that satisfy the bounds property for x and $|x|$ for the same order.

Definition 9 (*sign*)

The function sign *is defined from* $\mathbb{R}$ *to* $\{-1, 0, 1\}$ *with the usual following equality:*

$$\text{sign}(x) = \begin{cases} -1 & \text{if } x < 0 \\ 0 & \text{if } x = 0 \\ 1 & \text{otherwise} \end{cases}$$

Practically, for each non-zero real number x and for each integer n, the sign of x is the sign of each non-zero value x_n since if $x_n > 0$, then $x > (x_n - 1)B^{-n} \geq (1-1)B^{-n} = 0$ and if $x_n < 0$, then $x < (x_n + 1)B^{-n} \leq (-1+1)B^{-n} = 0$. Furthermore its computation terminates for any not null number.

Property 10 *Let x be a real number represented by the sequence $(x_n)_{n\in\mathbb{Z}}$ and n be an integer, $|x_n|$ satisfies the bounds property of $|x|$ for order n and if p is a positive integer that satisfies the bounds property of $|x|$ for order n, then the integer q defined by $q = \text{sign}(x) \times p$ satisfies the bounds property of x for order n.*

Proof.

Let us prove that $(|x_n| - 1)B^{-n} < |x| < (|x_n| + 1)B^{-n}$.

If $x_n \geq 1$, then $|x_n| = x_n$ and $x > 0$, thus $|x| = x$ and this inequality is exactly the definition formula of the bounds property of x for order n satisfied by x_n.

If $x_n \leq -1$, then $|x_n| = -x_n$ and $x < 0$, thus $|x| = -x$ and $(|x_n| - 1)B^{-n} < |x| < (|x_n| + 1)B^{-n}$ and finally $(-x_n - 1)B^{-n} < -x < (-x_n + 1)B^{-n}$, that is to say the definition formula of the bounds property of x for order n satisfied by x_n multiplied by (-1).

Finally if $x_n = 0$, then $(|x_n| - 1)B^{-n} < |x| < (|x_n| + 1)B^{-n}$ that may be rewritten in $-B^{-n} < |x| < B^{-n}$ that is to say the definition formula of the bounds property of x for order n satisfied by x_n multiplied by the sign of x.

Let us now prove the second part of this property. According to the hypothesis, we have $(p-1)B^{-n} < |x| < (p+1)B^{-n}$.

If p is null, then we have $0 \leq |x| < B^{-n}$ that is to say $(0-1)B^{-n} < x < (0+1)B^{-n}$ and consequently $q = 0$ satisfies the bounds property of x for order n.

If p is not null, then $0 \leq (p-1)B^{-n} < |x| < (p+1)B^{-n}$ and as $x = sign(x) \times |x| \neq 0$, we have $(p-1)B^{-n} < x < (p+1)B^{-n}$ if $sign(x) = 1$ and $-(p+1)B^{-n} < x < -(p-1)B^{-n}$ otherwise, that is to say $(sign(x) \times p - 1)B^{-n} < x < (sign(x) \times p + 1)B^{-n}$ and q satisfies the bounds property of x for order n. □

We have also the following technical properties:

Property 11 *Let x, α be two real numbers and n be an integer. If $\alpha B^{-n} < x < (\alpha+1)B^{-n}$, then $\lfloor \alpha \rfloor + 1$ and $\lceil \alpha \rceil$ satisfy the bounds property of x for order n.*

Proof.

If $\alpha B^{-n} < x \leq \alpha + 1 B^{-n}$, then we have

$$(\lfloor \alpha \rfloor + 1) - 1 = \lfloor \alpha \rfloor \leq \alpha < B^n x \leq \alpha + 1 \leq \alpha + 1 < \lfloor \alpha \rfloor + 2 = (\lfloor \alpha \rfloor + 1) + 1$$

and $\lfloor \alpha \rfloor + 1$ satisfies the bounds property of x for order n. Just as before, if $\alpha B^{-n} \leq x < (\alpha+1)B^{-n}$, then we have

$$\lceil \alpha \rceil - 1 < \alpha \leq B^n x < \alpha + 1 \leq \lceil \alpha \rceil + 1$$

and $\lceil \alpha \rceil$ satisfies the bounds property of x for order n. □

Property 12 *Let x be a real number represented by the sequence $(x_n)_{n\in\mathbb{Z}}$, n and m be integers such that $n \leq m$, then the integer $\left\lfloor \frac{x_m}{B^{m-n}} \right\rfloor$ satisfies the bounds property of x for order n.*

Proof.

We will prove this property by demonstration of each inequality of the definition formula of the bounds property. According to the definition of $x \mapsto \lfloor x \rfloor$, we have

$$\frac{x_m}{B^{m-n}} - 1 < \left\lfloor \frac{x_m}{B^{m-n}} \right\rfloor \leq \frac{x_m}{B^{m-n}}$$

thus we have

$$\left(\left\lfloor \frac{x_m}{B^{m-n}} \right\rfloor - 1 \right) B^{-n} \leq \left(\frac{x_m}{B^{m-n}} - 1 \right) B^{-n}.$$

We supposed that $n \leq m$ consequently $-B^{-n} \leq -B^{-m}$ and we obtain

$$\left(\left\lfloor \frac{x_m}{B^{m-n}} \right\rfloor - 1 \right) B^{-n} \leq (x_m - 1)\, B^{-m} < x.$$

We will now prove the other inequality. Let ℓ be define by $x = (x_m + \ell)B^{-m}$ with $-1 < \ell < 1$ and k is the remainder of the Euclidean division of x_m by B^{m-n}: $x_m = \left\lfloor \frac{x_m}{B^{m-n}} \right\rfloor B^{m-n} + k$ and $0 \leq k \leq B^{m-n} - 1$. Consequently we have

$$x = \left(\left\lfloor \frac{x_m}{B^{m-n}} \right\rfloor B^{m-n} + k + \ell \right) B^{-m},$$

that is to say

$$x = \left\lfloor \frac{x_m}{B^{m-n}} \right\rfloor B^{-n} + (k + \ell)B^{-m}.$$

But $k + \ell < B^{m-n}$, thus

$$x < \left(\left\lfloor \frac{x_m}{B^{m-n}} \right\rfloor + 1 \right) B^{-n}.$$

□

For efficiency reasons, the implemented representation includes for each real number x represented by a sequence $(x_n)_{n\in\mathbb{Z}}$ the most precise approximation ever computed for x, $x_{\mathrm{mpa}(x)}$ for order $\mathrm{mpa}(x)$. In this way, any approximation less precise for x may be computed by a simple shift operation on $x_{\mathrm{mpa}(x)}$ rather than by a possibly complex computation that we have in some sense already performed before:

$$\text{if } n \leq \mathrm{mpa}(x), \text{ then we take } x_n = \left\lfloor \frac{x_{\mathrm{mpa}(x)}}{B^{\mathrm{mpa}(x)-n}} \right\rfloor.$$

The value of x_n may slightly vary according to the $\mathrm{mpa}(x)$ value.

We will now define the msd function that indicates the order of magnitude of a real number.

Definition 10 (msd)

The function msd *("most significant digit") is defined from $\mathbb{R}$ to $\mathbb{Z}$ for any real number x represented by the sequence $(x_n)_{n\in\mathbb{Z}}$, by the equality* $\mathrm{msd}(x) = \min_{n\in\mathbb{Z}}(|x_n| > 1)$.

Practically, the function msd is recursively computed and does not terminate for zero. This function satisfies the following properties:

Properties 13

1.. *For any non-zero real number x,* $\mathrm{msd}(x)$ *exists and is unique (at the exact time of its computation, see the remark below), with $2 \leq \left|x_{\mathrm{msd}(x)}\right| \leq 2B$ and $\mathrm{msd}(x) = -\lfloor \log_B |x| \rfloor$ or $\mathrm{msd}(x) = -\lfloor \log_B |x| \rfloor + 1$.*

2.. *For any non-zero real number x and any integer $n < \mathrm{msd}(x)$, we have $|x_n| \leq 1$.*

3.. *For any non-zero real number x and any integer $n \geq \mathrm{msd}(x)$, then*

$$B^{n-\mathrm{msd}(x)} \leq |x_n| \leq B^{n-\mathrm{msd}(x)}(2B + 1)$$

and

$$1 \leq \left| \left\lfloor \frac{x_n}{B^{n-\mathrm{msd}(x)}} \right\rfloor \right| \leq 2B + 1.$$

Proof.

Let x be a not null real number, we will prove there exists an integer n such that $|x_n| > 1$.

By definition of $x \mapsto \lfloor x \rfloor$, we have

$$B^{\lfloor \log_B |x| \rfloor} \leq |x| < B^{\lfloor \log_B |x| \rfloor + 1} \tag{3.1}$$

but according to property (10), $|x_{-\lfloor \log_B |x| \rfloor}|$ satisfies the bounds property for order $-\lfloor \log_B |x| \rfloor$ of $|x|$:

$$\frac{|x_{-\lfloor \log_B |x| \rfloor}| - 1}{B^{-\lfloor \log_B |x| \rfloor}} < |x| < \frac{|x_{-\lfloor \log_B |x| \rfloor}| + 1}{B^{-\lfloor \log_B |x| \rfloor}}.$$

We combine these two inequalities and we obtain

$$\frac{|x_{-\lfloor \log_B |x| \rfloor}| - 1}{B^{-\lfloor \log_B |x| \rfloor}} < B^{\lfloor \log_B |x| \rfloor + 1}$$

and

$$B^{\lfloor \log_B |x| \rfloor} < \frac{|x_{-\lfloor \log_B |x| \rfloor}| + 1}{B^{-\lfloor \log_B |x| \rfloor}}.$$

We reduce to lowest terms and obtain the following inequalities between the concerned numerators: $|x_{-\lfloor \log_B |x| \rfloor}| - 1 < B$ et $1 < |x_{-\lfloor \log_B |x| \rfloor}| + 1$.

These inequalities concern integers thus we deduce that $|x_{-\lfloor \log_B |x| \rfloor}| \leq B$ and $1 \leq |x_{-\lfloor \log_B |x| \rfloor}|$ that is to say $1 \leq |x_{-\lfloor \log_B |x| \rfloor}| \leq B$.

Let us suppose that $\left|x_{-\lfloor \log_B |x| \rfloor}\right| = 1$ and prove that $1 < \left|x_{-\lfloor \log_B |x| \rfloor + 1}\right|$.

According to the bounds property of $|x|$ for order $-\lfloor \log_B |x| \rfloor + 1$ satisfied by $|x_{-\lfloor \log_B |x| \rfloor + 1}|$, we have

$$\frac{|x_{-\lfloor \log_B |x| \rfloor + 1}| - 1}{B^{-\lfloor \log_B |x| \rfloor + 1}} < |x| < \frac{|x_{-\lfloor \log_B |x| \rfloor + 1}| + 1}{B^{-\lfloor \log_B |x| \rfloor + 1}}.$$

We combine this inequality with (3.1), we obtain

$$\frac{|x_{-\lfloor \log_B |x| \rfloor + 1}| - 1}{B^{-\lfloor \log_B |x| \rfloor + 1}} < B^{\lfloor \log_B |x| \rfloor + 1}$$

and

$$B^{\lfloor \log_B |x| \rfloor} < \frac{|x_{-\lfloor \log_B |x| \rfloor + 1}| + 1}{B^{-\lfloor \log_B |x| \rfloor + 1}}.$$

We reduce to lowest term and obtain the following inequalities between the concerned numerators: $|x_{-\lfloor \log_B |x| \rfloor + 1}| - 1 < B^2$ et $B < |x_{-\lfloor \log_B |x| \rfloor + 1}| + 1$, that is to say $B \leq |x_{-\lfloor \log_B |x| \rfloor + 1}| \leq B^2$ and *a fortiori* $1 < |x_{-\lfloor \log_B |x| \rfloor + 1}|$ since $B \geq 2$.

Since we consider the smallest n such that $|x_n| > 1$, the result is completely determined when the computation is performed even if it may vary according to the computed approximation of x.

We will now prove that for any $n < -\lfloor \log_B |x| \rfloor$, we have $|x_n| \leq 1$. According to the bounds property of $|x|$ satisfied by $|x_n|$ for order n, we have

$$\frac{|x_n| - 1}{B^n} < |x|$$

and according to (3.1), we have $|x| < B^{\lfloor \log_B |x| \rfloor + 1}$. We deduce from these two inequalities that $|x_n| - 1 < B^{n + \lfloor \log_B |x| \rfloor + 1}$. But, according to an hypothesis, we have $n + \lfloor \log_B |x| \rfloor + 1 \leq 0$, thus $|x_n| - 1 < 1$ and since this inequality concerns integers, we have $|x_n| \leq 1$.

We will now prove the inequalities for $\left|x_{\mathrm{msd}(x)}\right|$.

If $\mathrm{msd}(x) = -\lfloor \log_B |x| \rfloor$, then we have $1 < \left|x_{\mathrm{msd}(x)}\right| \leq B$.

If $\mathrm{msd}(x) = -\lfloor \log_B |x| \rfloor + 1$, we have $\left|x_{-\lfloor \log_B |x| \rfloor}\right| = 1$, thus $\left|x_{-\lfloor \log_B |x| \rfloor}\right| + 1 = 2$ and according to the bounds property of $|x|$ for order $-\lfloor \log_B |x| \rfloor$ satisfied by $\left|x_{-\lfloor \log_B |x| \rfloor}\right|$, $|x| < 2B^{\lfloor \log_B |x| \rfloor}$ and $\left|x_{\mathrm{msd}(x)}\right| - 1 < 2B$, thus $2 \leq \left|x_{\mathrm{msd}(x)}\right| \leq 2B$.

The lower bound could be reach in an obvious way. The upper bound $|x_{smsd(x)}| \leq 2B$, may also be reach as illustrated on the following example: let us choose $x = 2 - 1/(2B)$, $x_0 = 1$, $x_1 = 2B$ with $\mathrm{msd}(x) = 1$.

Let n be an integer such that $n \geq \mathrm{msd}(x)$.

According to the definition of $x_{\mathrm{msd}(x)}$, we have

$$\left(\left|x_{\mathrm{msd}(x)}\right|-1\right)B^{n-\mathrm{msd}(x)}B^{-n} < |x| < \left(\left|x_{\mathrm{msd}(x)}\right|+1\right)B^{n-\mathrm{msd}(x)}B^{-n},$$

thus

$$\left(\left|x_{\mathrm{msd}(x)}\right|-1\right)B^{n-\mathrm{msd}(x)} \leq B^n x \leq \left(\left|x_{\mathrm{msd}(x)}\right|+1\right)B^{n-\mathrm{msd}(x)}$$

and

$$\left(\left|x_{\mathrm{msd}(x)}\right|-1\right)B^{n-\mathrm{msd}(x)} \leq |x_n| \leq \left(\left|x_{\mathrm{msd}(x)}\right|+1\right)B^{n-\mathrm{msd}(x)}.$$

But $2 \leq \left|x_{\mathrm{msd}(x)}\right| \leq 2B$, thus $B^{n-\mathrm{msd}(x)} \leq |x_n| \leq B^{n-\mathrm{msd}(x)}(2B+1)$.

We suppose that $|x_n| \geq 2B\,B^{n-\mathrm{msd}(x)}+1$. We already know that this is impossible if $n = \mathrm{msd}(x)$ so we can consider directly that $n - \mathrm{msd}(x) \geq 1$. Thus we have $B^{\mathrm{msd}(x)-1}\,|x| > (|x_n|-1)B^{-(n-\mathrm{msd}(x)+1)} \geq 2$ thus $|x_{\mathrm{msd}(x)-1}| \geq 2$ that is refuted by the minimality of $\mathrm{msd}(x)$. Consequently

$$\frac{|x_n|}{B^{n-\mathrm{msd}(x)}} \leq 2B$$

and

$$\left\lfloor \frac{|x_n|}{B^{n-\mathrm{msd}(x)}} \right\rfloor \leq 2B.$$

But we may have $B^{n-\mathrm{msd}(x)} \leq |x_n| < B^{n-\mathrm{msd}(x)}+1$ for $n \geq \mathrm{msd}(x)+1$ as illustrated in the following example: let us choose $x = 1+1/2B$, $x_0 = 2$, $x_1 = B$, $\mathrm{msd}(x) = 0$ and $n = 1$. □

Let us notice that the value of x_n and of $msd(x)$ may vary of one unit according to the value to $\mathrm{mpa}(x)$. However the rather strict definition that we give of this function ensures, independently of the value of $\mathrm{mpa}(x)$, the fundamental property we wanted to ensure, i.e. $1 < \left|x_{\mathrm{msd}(x)}\right| \leq B$ whatever the value of $\mathrm{mpa}(x)$ is when $\mathrm{msd}(x)$ is computed.

We present now a complete set of algorithms to compute elementary functions for this representation, using the corresponding algorithms for rational numbers.

Chapter 4

Algorithms for the usual elementary functions

4.1 Introduction to algorithms for the computation of elementary functions on $\mathcal{R}$

We will now describe algorithms for computing elementary functions on $\mathcal{R}$.

We associate to each elementary function $f : \mathbb{R}^p \to \mathbb{R}$ its representation $\overline{f} : \mathcal{R}^p \to \mathcal{R}$ and for any computable real x, we note $\overline{x}$ its representation.

For each elementary function f with p arguments $(x_1, \ldots, x_p)$, for any integer n and for any $1 \leq i \leq p$, we have to describe to what precision k_i each argument x_i is supposed to be computed in a $\overline{x_i}_{k_i}$ approximation and give a formula to apply to these approximations to produce $\overline{f}(\overline{x_1}, \ldots, \overline{x_p})_n$.

After this description of the algorithm we have to establish a theorem of correction as follows:

> *Theorem (Correction of the algorithm for computing a function f on computable real arguments $(x_1, \ldots, x_p)$): The sequence $\overline{f}(\overline{x_1}, \ldots, \overline{x_p})$ is a representation of $f(x_1, \ldots, x_p)$.*

In other words, for any order n, $\overline{f}(\overline{x_1}, \ldots, \overline{x_p})_n$ satisfies the bounds property of $f(x_1, \ldots, x_p)$. This can be also interpreted as the fact that it is right to define $\overline{f(x_1, \ldots, x_p)}$ as $\overline{f}(\overline{x_1}, \ldots, \overline{x_p})$.

These algorithms are designed for any integer $B \geq 2$. In almost all algorithms we distinguish the case where $B \geq 4$ and $B = 2$ or 3. In fact, for $B \geq 4$ we can generally give more precise constants and to distinguish this case rather than to adopt the constants determined by the formulas for $B \geq 2$. This should improve the efficiency of the algorithms. We could of course distinguish other cases for which the constant may be still smaller but we have to stop this sequence of improvements and in fact the distinction of $B \geq 4$ is relevant according to the implementation (see section 5 below).

4.2 Algorithms for rational operations

We will now describe the representation of the image of any computable real number by an elementary function and we begin with the heart of these algorithms: the representation of rational numbers.

4.2.1 Representation of rational numbers

Each rational number q is represented by the sequence of integers $(q_n)_{n \in \mathbb{N}}$ such that q_n is defined, for any integer n by the equality:

$$q_n = \lfloor B^n q \rfloor.$$

Proof.

Immediate. □

4.2.2 Addition of real numbers

Let x and y be two real numbers represented by the sequences $(x_n)_{n\in\mathbb{Z}}$ and $(y_n)_{n\in\mathbb{Z}}$ respectively, we represent the sum of these two numbers $x+y$ by the sequence $(\overline{x+y}_n)_{n\in\mathbb{Z}}$ such that:

$$\overline{x+y}_n = \left\lfloor \frac{x_{n+w}+y_{n+w}}{B^w} \right\rceil \text{ with } w = \begin{cases} 1 & \text{if } B \geq 4 \\ 2 & \text{if } B = 2 \text{ or } 3. \end{cases}$$

Theorem 14 (Correction of the addition algorithm)
For any integer n, $\overline{x+y}_n$ satisfies the bounds property of x for order n.

Proof.
We will first consider the case $B \geq 4$. Let n be an integer. According to the definition of $x \mapsto \lfloor x \rceil$, we have:

$$\frac{x_{n+w}+y_{n+w}}{B^w} - \frac{1}{2} < \overline{x+y}_n \leq \frac{x_{n+w}+y_{n+w}}{B^w} + \frac{1}{2}$$

thus

$$\frac{\overline{x+y}_n - 1}{B^n} \leq \frac{x_{n+w}+y_{n+w}-\frac{B}{2}}{B^{n+w}}$$

and

$$\frac{\overline{x+y}_n + w}{B^n} > \frac{x_{n+w}+y_{n+w}+\frac{B}{2}}{B^{n+w}}$$

But according to the definition of w, we have $B^w/2 \geq 2$ and then

$$\frac{\overline{x+y}_n - 1}{B^n} \leq \frac{x_{n+w}+y_{n+w}-2}{B^{n+w}} = \frac{x_{n+w}-1}{B^{n+w}} + \frac{y_{n+w}-1}{B^{n+w}}$$

and

$$\frac{\overline{x+y}_n + w}{B^n} > \frac{x_{n+w}+y_{n+w}+2}{B^{n+w}} = \frac{x_{n+w}+1}{B^{n+w}} + \frac{y_{n+w}+1}{B^{n+w}}.$$

According to the bounds properties of x and y for order $n+w$ satisfied respectively by x_{n+w} and y_{n+w} respectively, we have

$$\frac{x_{n+w}-1}{B^{n+w}} < x < \frac{x_{n+w}+1}{B^{n+w}}$$

and

$$\frac{y_{n+w}-1}{B^{n+w}} < y < \frac{y_{n+w}+1}{B^{n+w}},$$

and we deduce:

$$(\overline{x+y}_n - 1)B^{-n} < x+y < (\overline{x+y}_n + 1)B^{-n}.$$

□

Opposite of a real number

Let x be a real number represented by the sequence $(x_n)_{n\in\mathbb{Z}}$, we represent the opposite of this number $-x$ by the sequence $(\overline{-x}_n)_{n\in\mathbb{Z}}$ such that:

$$\overline{-x}_n = -x_n.$$

Proof (*Correction of the opposite algorithm*).
The correction is obvious by symmetry of the bounds property around 0. □

4.2.3 Multiplication of two real numbers

Let x et y be two real numbers represented by the sequences $(x_n)_{n\in\mathbb{Z}}$ and $(y_n)_{n\in\mathbb{Z}}$ respectively, we represent the product of these two numbers $x+y$ by the sequence $(\overline{x\times y}_n)_{n\in\mathbb{Z}}$ such that:

$$\overline{x\times y}_n = \mathit{sign}(x_{p_x}) \times \mathit{sign}(y_{p_y}) \times \left\lfloor \frac{1+|x_{p_x}\times y_{p_y}|}{B^{p_x+p_y-n}} \right\rfloor$$
$$\text{with } p_x = \max(n-\mathrm{msd}(y)+v, \lfloor (n+w)/2 \rfloor)$$
$$\text{and } p_y = \max(n-\mathrm{msd}(x)+v, \lfloor (n+w)/2 \rfloor)$$
$$\text{and } (v,w) = \begin{cases} (3,2) & \text{if } B \geq 4 \\ (3,3) & \text{if } B = 3 \\ (4,3) & \text{if } B = 2 \end{cases}$$

REMARK.

The computation of $\mathrm{msd}(x)$ is restricted here by the evaluation of the maximum in expression p_y, that is to say that for any integer k from 0 (beginning of the recursion in the computation of $\mathrm{msd}(x)$) to $n+v-\lfloor (n+w)/2 \rfloor$ (maximum value of $\mathrm{msd}(x)$ for which p_y is determined by the first term in the maximum expression) $x_k = 0$, then p_y is determined by the second term and we stop the computation of $\mathrm{msd}(x)$ and in this way multiplication by 0 terminates. This analysis is of course identical for the evaluation of $\mathrm{msd}(y)$ inside the computation of p_x. $\diamond$

Theorem 15

1.. For any integer n, we have

$$(\overline{x\times y}_n - 1)B^{-n} < x\times y < (\overline{x\times y}_n + 1)B^{-n}.$$

2.. If the computation of x and y terminates, then the computation of $x\times y$ terminates too.

Proof (*Correction of the multiplication algorithm*).

First of all, we remark that $p_x + p_y - n \geq 2\times(\lfloor (n+w)/2 \rfloor) - n \geq w-1$.

If $|x_{p_x}| = 0$ and $|y_{p_y}| = 0$, then we have

$$B^n |x\times y| < \frac{(|x_{p_x}|+1)(|y_{p_y}|+1)}{B^{p_x+p_y-n}} \leq \frac{1}{B^{p_x+p_y-n}} \leq \frac{1}{B^{w-1}}$$

since $p_x+p_y-n \geq w-1$. According to the definition of w, we have $\frac{1}{B^{w-1}} \leq 1$ and $-\frac{1}{B^n} < x\times y < \frac{1}{B^n}$. Consequently $\overline{x\times y}_n = 0$ satisfies the bounds property of $x\times y$ for order n.

We will now consider the last case, that is to say, if at least one of the absolute values $|x_{p_x}|$ and $|y_{p_y}|$ is greater or equal to 1. Because of the symmetry of the problem, we can decide, to reduce the combinatorics of this case by case analysis, that x is concerned.

We have:

$$\frac{1+|x_{p_x}\,y_{p_y}| - (|x_{p_x}|+|y_{p_y}|)}{B^{p_x+p_y-n}} < B^n|x\times y| < \frac{1+|x_{p_x}\,y_{p_y}| + (|x_{p_x}|+|y_{p_y}|)}{B^{p_x+p_y-n}}.$$

We will prove that from

$$\frac{|x_{p_x}|+|y_{p_y}|}{B^{p_x+p_y-n}} \leq \frac{1}{2},$$

we can deduce the correction of the algorithm. Let us suppose that this property is satisfied, then we have

$$\frac{1+|x_{p_x}\,y_{p_y}|}{B^{p_x+p_y-n}} - \frac{1}{2} < B^n|x\times y| < \frac{1+|x_{p_x}\,y_{p_y}|}{B^{p_x+p_y-n}} + \frac{1}{2}.$$

According to the definition of $x\mapsto \lfloor x \rceil$, we have:

$$\frac{1+|x_{p_x}\,y_{p_y}|}{B^{p_x+p_y-n}} - \frac{1}{2} < \left\lfloor \frac{1+|x_{p_x}\,y_{p_y}|}{B^{p_x+p_y-n}} \right\rceil \leq \frac{1+|x_{p_x}\,y_{p_y}|}{B^{p_x+p_y-n}} + \frac{1}{2},$$

thus

$$\left\lfloor \frac{1+|x_{p_x}\,y_{p_y}|}{B^{p_x+p_y-n}} \right\rceil - 1 \leq \frac{1+|x_{p_x}\,y_{p_y}|}{B^{p_x+p_y-n}} - \frac{1}{2} < B^n|x\times y| < \frac{1+|x_{p_x}\,y_{p_y}|}{B^{p_x+p_y-n}} + \frac{1}{2} < \left\lfloor \frac{1+|x_{p_x}\,y_{p_y}|}{B^{p_x+p_y-n}} \right\rceil + 1$$

that is to say that

$$\left\lfloor \frac{1+|x_{p_x}\, y_{p_y}|}{B^{p_x+p_y-n}} \right\rfloor$$

satisfies the bounds property of $|x \times y|$ for order n.

According to proposition 10,

$$sign(x \times y) \times \left\lfloor \frac{1+|x_{p_x}\, y_{p_y}|}{B^{p_x+p_y-n}} \right\rfloor$$

satisfies the bounds property of $x \times y$ for order n.

We have to prove that

$$sign(x \times y) \times \left\lfloor \frac{1+|x_{p_x}\, y_{p_y}|}{B^{p_x+p_y-n}} \right\rfloor = sign(x_{p_x}) \times sign(y_{p_y}) \times \left\lfloor \frac{1+|x_{p_x}\, y_{p_y}|}{B^{p_x+p_y-n}} \right\rfloor$$

to be able to deduce that $\overline{x \times y}_n$ satisfies the bounds property of $x \times y$ for order n.

First of all, it is clear that $sign(x \times y) = sign(x) \times sign(y)$. Since $|x_{p_x}| \geq 1$, $sign(x) = sign(x_{p_x})$. Thus we have

$$sign(x \times y) \times \left\lfloor \frac{1+|x_{p_x}\, y_{p_y}|}{B^{p_x+p_y-n}} \right\rfloor = sign(x_{p_x}) \times sign(y) \times \left\lfloor \frac{1+|x_{p_x}\, y_{p_y}|}{B^{p_x+p_y-n}} \right\rfloor .$$

If $y_{p_y} = 0$, then

$$\left\lfloor \frac{1+|x_{p_x}\, y_{p_y}|}{B^{p_x+p_y-n}} \right\rfloor = \left\lfloor \frac{1}{B^{p_x+p_y-n}} \right\rfloor = 0 = \overline{x \times y}_n$$

if $|y_{p_y}| > 0$, $sign(y) = sign(y_{p_y})$ and

$$sign(x \times y) \times \left\lfloor \frac{1+|x_{p_x}\, y_{p_y}|}{B^{p_x+p_y-n}} \right\rfloor = sign(x_{p_x}) \times sign(y_{p_y}) \times \left\lfloor \frac{1+|x_{p_x}\, y_{p_y}|}{B^{p_x+p_y-n}} \right\rfloor = \overline{x \times y}_n.$$

We have now to prove that

$$\frac{|x_{p_x}| + |y_{p_y}|}{B^{p_x+p_y-n}} \leq \frac{1}{2}.$$

Since $|x_{p_x}| \geq 2$, then $p_x \geq \mathrm{msd}(x)$ and $|x_{p_x}| \leq 2B\, B^{p_x-\mathrm{msd}(x)}$.

If $p_y \geq \mathrm{msd}(y)$, we have in the same way $|y_{p_y}| \leq 2B\, B^{p_y-\mathrm{msd}(y)}$ and we have the following inequality:

$$\frac{|x_{p_x}| + |y_{p_y}|}{B^{p_x+p_y-n}} \leq 2B \left(\frac{B^{p_x-\mathrm{msd}(x)} + B^{p_y-\mathrm{msd}(y)}}{B^{p_x+p_y-n}} \right) = 2B \left(\frac{B^n}{B^{p_y+\mathrm{msd}(x)}} + \frac{B^n}{B^{p_x+\mathrm{msd}(y)}} \right).$$

But $p_y + \mathrm{msd}(x) \geq n + v$ and $p_x + \mathrm{msd}(y) \geq n + v$ according to the definition of p_x and p_y, thus

$$\frac{|x_{p_x}| + |y_{p_y}|}{B^{p_x+p_y-n}} \leq \frac{4B}{B^v} = \frac{4}{B^{v-1}}$$

According to the definition of v, we have

$$\frac{4}{B^{v-1}} \leq \frac{1}{2}$$

and the case $p_y \geq \mathrm{msd}(y)$ is terminated.

Finally we consider the case if $p_y < \mathrm{msd}(y)$, then we have $|y_{p_y}| \leq 1$ and

$$\frac{|x_{p_x}| + |y_{p_y}|}{B^{p_x+p_y-n}} \leq \frac{2B\, B^{p_x-\mathrm{msd}(x)} + 1}{B^{p_x+p_y-n}} = \frac{2B\, B^n}{B^{p_y+\mathrm{msd}(x)}} + \frac{1}{B^{p_x+p_y-n}}$$

and we obtain the inequality

$$\frac{2B\, B^n}{B^{p_y+\mathrm{msd}(x)}} + \frac{1}{B^{p_x+p_y-n}} \leq \frac{2B}{B^v} + \frac{1}{B^{w-1}} = \frac{2}{B^{v-1}} + \frac{1}{B^{w-1}}$$

As before, according to the definition of v and w, we have

$$\frac{2}{B^{v-1}} + \frac{1}{B^{w-1}} \leq \frac{1}{2}$$

and the proof is complete. □

4.2.4 Inverse of a real number

Let x be a real number respectively represented by the sequence $(x_n)_{n\in\mathbb{Z}}$, we represent the inverse of this number $1/x$ by the sequence $(\overline{1/x}_n)_{n\in\mathbb{Z}}$ such that:

$$\begin{array}{l}\text{If } n \le -\text{msd}(x) \text{ then } \overline{1/x}_n = 0\\ \quad \text{else } \overline{1/x}_n = \left\lfloor \dfrac{B^{k+n}}{x_k+1} \right\rfloor + 1\\ \text{with } k = n + 2\text{msd}(x) + w \text{ and } w = \begin{cases} 1 & \text{if } B \ge 3\\ 2 & \text{if } B = 2.\end{cases}\end{array}$$

Theorem 16

1.. For any integer n, we have

$$(\overline{1/x}_n - 1)B^{-n} < \frac{1}{x} < (\overline{1/x}_n + 1)B^{-n}.$$

2.. If the computation of x terminates and x is not null, then the computation of $1/x$ terminates.

Proof (*Correction of the inverse algorithm***).**
Let x be a not null real number. According to the bounds property of $|x|$ for order $\text{msd}(x)$ satisfied by $|x_{\text{msd}(x)}|$, we have:

$$0 < \frac{|x_{\text{msd}(x)}| - 1}{B^{\text{msd}(x)}} < |x| < \frac{|x_{\text{msd}(x)}| + 1}{B^{\text{msd}(x)}}$$

thus

$$0 < \left|\frac{1}{x}\right| < \frac{1}{|x_{\text{msd}(x)}| - 1} B^{\text{msd}(x)} \le 1 \times B^{\text{msd}(x)},$$

and 0 satisfies the bounds property of $1/x$ for order n for any $n \le -\text{msd}(x)$.

We will now suppose that $n > -\text{msd}(x)$, then $k > \text{msd}(x) + 1$ and $|x_k| - 1 > 0$. We write the bounds property of $|x|$ for order k satisfied by $|x_k|$:

$$0 < \frac{|x_k| - 1}{B^k} < |x| < \frac{|x_k| + 1}{B^k}$$

then we deduce that

$$\frac{B^{k+n}}{|x_k|+1} B^{-n} < \frac{1}{|x|} < \frac{B^{k+n}}{|x_k|-1} B^{-n}$$

and

$$\left\lfloor \frac{B^{k+n}}{|x_k|+1} \right\rfloor B^{-n} < \frac{1}{|x|} < \left\lceil \frac{B^{k+n}}{|x_k|-1} \right\rceil B^{-n}.$$

Let us define α and β as follows:

$$\begin{aligned} \alpha &= \left\lfloor \frac{B^{k+n}}{|x_k|+1} \right\rfloor \\ \beta &= \left\lceil \frac{B^{k+n}}{|x_k|-1} \right\rceil. \end{aligned}$$

Then the preceding inequality is $\alpha B^{-n} < 1/|x| < \beta B^{-n}$.

We will now prove that $1 \le \beta - \alpha \le 2$.
We have:

$$0 < \frac{B^{k+n}}{|x_k|-1} - \frac{B^{k+n}}{|x_k|+1} = \frac{2B^{k+n}}{|x_k|^2-1} < \frac{2B^{k+n}}{|x_k|^2}\left(1 + \frac{2}{|x_k|^2}\right).$$

But $k > \text{msd}(x)$, thus $|x_k| \ge B^{k-\text{msd}(x)} \ge B$ and

$$0 < \frac{B^{k+n}}{|x_k|-1} - \frac{B^{k+n}}{|x_k|+1} < 2B^{n+2\text{msd}(x)-k}\left(1 + \frac{2}{B^2}\right) = \frac{2}{B^w}\left(1 + \frac{2}{B^2}\right)$$

and according to the definition of w, we have finally

$$0 < \frac{B^{k+n}}{|x_k| - 1} - \frac{B^{k+n}}{|x_k| + 1} < 1. \tag{4.1}$$

Furthermore, according to the definition of $x \mapsto \lfloor x \rfloor$ and $x \mapsto \lceil x \rceil$, we have:

$$\alpha > \frac{B^{k+n}}{|x_k| + 1} - 1$$

and

$$\beta < \frac{B^{k+n}}{|x_k| - 1} + 1.$$

Consequently, we combine these two inequalities with the inequality 4.1 and we obtain

$$0 < \beta - \alpha < 1 + \frac{B^{k+n}}{|x_k| - 1} - \frac{B^{k+n}}{|x_k| + 1} + 1 < 3.$$

This inequality concerns integers so we deduce that $1 \leq \beta - \alpha \leq 2$.

Consequently $\alpha B^{-n} < 1/|x| < (\alpha + 2)B^{-n}$ and $(\beta - 2)B^{-n} < 1/|x| < \beta B^{-n}$, thus $\alpha + 1$ and $\beta - 1$ satisfy the bounds property of $1/|x|$ for order n. We deduce from property 10 that $sign(x) \times (\alpha + 1)$ and $sign(x) \times (\beta - 1)$ satisfy the bounds property of $1/x$ for order n. If $x > 0$, then $sign(x) = 1$ and

$$\alpha + 1 = \left\lfloor \frac{B^{k+n}}{x_k + 1} \right\rfloor + 1,$$

thus $sign(x) \times (\alpha + 1) = \overline{1/x}_n$. On the other hand, if $x < 0$, then $sign(x) = -1$ and

$$\beta - 1 = \left\lceil \frac{B^{k+n}}{-x_k - 1} \right\rceil - 1,$$

thus $sign(x) \times (\beta - 1) = \overline{1/x}_n$ also. Finally $\overline{1/x}_n$ in both cases satisfies the bounds property of $1/x$ for order n. □

4.3 Algorithms for algebraic or transcendental functions

4.3.1 General idea of these algorithms

For the computation of any algebraic or transcendental function f, we will use an intermediate function $\underline{f} : \mathbb{Q} \to \mathcal{R}$ such that for any rational number q for which $f(q)$ is defined, $\underline{f}(q)_n$ satisfies the bounds property of $f(q)$ for any order n. We will then use these $\underline{f}$ functions to define the $\overline{f}$ functions but since the computation of $\underline{f}$ is more well known, even if this computation may lead to many tricks to be efficient, we will only quickly mention guidelines to compute them after this subsection to prove the completeness of our approach.

We present now such algorithms for algebraic and transcendental usual functions.

4.3.2 k-th root

Let x be a real number represented by the sequence $(x_n)_{n \in \mathbb{Z}}$ and k be an integer greater than or equal to 2. We represent the k-th root of this number $\sqrt[k]{x}$ by the sequence $(\overline{\sqrt[k]{x}}_n)_{n \in \mathbb{Z}}$ such that:

If $x_{kn} \geq 0$
  then $\left\lfloor \sqrt[k]{x_{kn}} \right\rfloor$
  else fails.

REMARK.

We choose here to always give a value to $\sqrt[k]{x}$ when it make sense, even if it means that it does not fail for some slightly negatives values of x. We can of course choose to fail for all negative values by replacing the condition $x_{kn} \geq 0$ by $x_{kn} \geq 1$, but in this case for some slightly positive values it will fail also. $\diamond$

Theorem 17 *For any positive real number x and any integer n, we have*

$$(\overline{\sqrt[k]{x}}_n - 1)B^{-n} < \sqrt[k]{x} < (\overline{\sqrt[k]{x}}_n + 1)B^{-n}.$$

Proof.

Let us suppose that $x_{kn} = 0$, then we have $-B^{-kn} < x < B^{-kn}$ and if $\sqrt[k]{x}$ is defined then we have also $\sqrt[k]{x} < B^{-n}$, thus $\lfloor \sqrt[k]{x_{kn}} \rfloor = 0$ satisfies the bounds property of $\sqrt[k]{x}$ for order n.

Will suppose now that $x_{kn} \geq 1$. We have

$$\frac{(x_{kn} - 1)}{B^{kn}} < x < \frac{(x_{kn} + 1)}{B^{kn}}$$

and consequently

$$\frac{\lfloor \sqrt[k]{x_{kn} - 1} \rfloor}{B^n} < \sqrt[k]{x} < \frac{\lceil \sqrt[k]{x_{kn} + 1} \rceil}{B^n}.$$

If there exists an integer p such that $x_{kn} = p^k$, then $\lfloor \sqrt[k]{x_{kn} - 1} \rfloor = p - 1$ and $\lceil \sqrt[k]{x_{kn} + 1} \rceil = p + 1$, thus $\overline{\sqrt[k]{x}}_n = p$ satisfies bounds property of $\sqrt[k]{x}$ for order n.

If there exists an integer p such that $x_{kn} = p^k + 1$, then $\lfloor \sqrt[k]{x_{kn} - 1} \rfloor = p$ and $\lceil \sqrt[k]{x_{kn} + 1} \rceil = p + 1$, thus $\overline{\sqrt[k]{x}}_n = p$ satisfies the bounds property of $\sqrt[k]{x}$ for order n.

If there exists an integer p such that $x_{kn} = p^k - 1$, then $\lfloor \sqrt[k]{x_{kn} - 1} \rfloor = p - 1$ and $\lceil \sqrt[k]{x_{kn} + 1} \rceil = p$, thus $\overline{\sqrt[k]{x}}_n = p - 1$ satisfies the bounds property of $\sqrt[k]{x}$ for order n.

Otherwise $x_{kn} - 1$, x_{kn} and $x_{kn} + 1$ are each one between p^k and $(p+1)^k$ with $p^k + 2 \leq x_{kn} \leq (p+1)^k - 2$ and we have $\lfloor \sqrt[k]{x_{kn} - 1} \rfloor = \lfloor \sqrt[k]{x_{kn}} \rfloor = p$ and $\lceil \sqrt[k]{x_{kn} + 1} \rceil = p + 1$, thus $\overline{\sqrt[k]{x}}_n = p$ satisfies the bounds property of $\sqrt[k]{x}$ for order n. $\square$

4.3.3 Exponential function

Let x be a real number represented by the sequence $(x_n)_{n \in \mathbb{Z}}$, we represent $\exp(x)$ by the sequence $(\overline{\exp(x)}_n)_{n \in \mathbb{Z}}$ such that:

If $\underline{\exp}(x_m / B^m)_p \leq 0$, then $\overline{\exp(x)}_n = 0$ else
If $n > 0$ and $\underline{\log_e}(1 - 1/B^n)_\ell + 2 < x_\ell < \underline{\log_e}(1 + 1/B^n)_\ell - 2$ or $n \leq 0$ and $x_0 \leq \lfloor \log_e(1 + 1/B^n) \rfloor - 1$
then $\overline{\exp(x)}_n = B^n$
else $\overline{\exp(x)}_n = \left\lceil \left(\underline{\exp}(x_k / B^k)_p / B - 1 \right) (1 - 1/B^k) \right\rceil$
with $m = \max\left(0, \lceil \log_B(e/(B-1)) \rceil\right)$, $\ell = n + w$, $d = \max(-p, c \log_B(e) / B^{\mathrm{msd}(x)})$

$$p = \max(0, \ell),\ c = \begin{cases} 2B & \text{if } x_{\mathrm{msd}(x)} \geq 1 \\ -1 & \text{otherwise} \end{cases},\ w = \begin{cases} 2 & \text{if } B = 2 \text{ or } B = 3 \\ 1 & \text{if } B \geq 4 \end{cases}$$

$$k = \begin{cases} (0, \mathrm{msd}(x), p + 1 + \lceil d + \log_B(e+1) \rceil) & \text{if } B = 2 \\ (0, \mathrm{msd}(x), p + 1 + \lceil d + \log_B((e+1)/(B-2)) \rceil) & \text{otherwise} \end{cases}$$

$$w = \begin{cases} 2 & \text{if } B = 2 \text{ or } B = 3 \\ 1 & \text{if } B \geq 4 \end{cases}$$

REMARKS.

1. *The first test aims to determine if x is a sufficiently negative number such that $\exp(x)$ may be approximated by 0 within B^{-n} rather than bother the multiplication of inequalities.*

2. *The second test, corresponding to the double hypothesis, aims to determine if x is close enough to 0 such that $\exp(x)$ may be approximated by 1 rather than not to terminate in 0 because of the computation of $\mathrm{msd}(0)$.*

3. *The Neperian logarithms* $\log_e(z)$*, where* $z = 1 \pm 1/B^n$*, are greater than* $\underline{\log_e}(z)_0 - 1$ *and lesser than* $\pm 1/B^n$*. We compute in the same way* $\log_B(e/(B-1))$*. Function* $\underline{\log_e}$ *refers to the logarithm function for rational arguments that we will suppose it exists in the next algorithm.*

4. *Practically, we substitute the values of* $\log_B e$ *and* $\log_B((e+1)/(B-2))$ *by a simple upper or lower bound in the formula above (for example if* $B = 4$*, we use the fact that* $0.72134 < \log_B(e) < 0.72135$ *and* $\log_B((e+1)/(B-2)) < 0.44732$*).*

5. *Practically, we can improve this algorithm by using the best known approximation of* x *after the determination of* $\mathrm{msd}(x)$ *and replace* $d/B^{\mathrm{msd}(x)}$ *by* $(x_{\mathrm{mpa}(x)} + 1)/B^{\mathrm{mpa}(x)}$*, so we obtain a finer upper bound and may appreciably reduce the value of* k*.*

Proof (*Correction of the exponential algorithm***).**
First of all we will consider the case $\underline{\exp}(x_m/B^m)_p \leq 0$. According to the definition of $\underline{\exp}(x_m/B^m)_p$, we have

$$\exp\left(\frac{x_m}{B^m}\right) < \frac{\underline{\exp}\left(\frac{x_m}{B^m}\right)_p + 1}{B^p}.$$

Furthermore from the hypothesis $\underline{\exp}(x_m/B^m)_p \leq 0$ we deduce that

$$\exp\left(\frac{x_m}{B^m}\right) < \frac{1}{B^p}.$$

But $x < (x_m + 1)/B^m$ and exp is a strictly increasing function thus

$$\exp(x) < \exp\left(\frac{x_m + 1}{B^m}\right) = \exp\left(\frac{x_m}{B^m}\right) \times \exp\left(\frac{1}{B^m}\right).$$

Furthermore the function exp has only positive values and we have

$$\exp(x) < \frac{\exp\left(\frac{1}{B^m}\right)}{B^p}.$$

But $m \geq 0$ according to the hypothesis, thus $0 < 1/B^m \leq 1$ and

$$\exp\left(\frac{1}{B^m}\right) \leq 1 + \frac{e}{B^m}.$$

Consequently we have

$$\exp(x) < \frac{1 + \frac{e}{B^m}}{B^p}.$$

But $m \geq \log_B(e/(B-1))$, thus

$$\frac{e}{B^m} \leq B - 1$$

and

$$\exp(x) < \frac{B}{B^p} = \frac{1}{B^{p-1}}.$$

We have also $p \geq n + 1$, thus

$$\exp(x) < \frac{1}{B^n}$$

and 0 satisfies the bounds property of $\exp(x)$ for order n.

We will now consider the first case of the double hypothesis. First of all we notice that since n is strictly positive each part of the inequality for x_ℓ presented in the description of the algorithm are meaningful.

According to the definition of $\underline{\log_e}(1 - 1/B^n)$ and of $\underline{\log_e}(1 + 1/B^n)$, we have

$$B^\ell \log_e\left(1 - \frac{1}{B^n}\right) < \underline{\log_e}\left(1 - \frac{1}{B^n}\right)_\ell + 1$$

and

$$\underline{\log_e}\left(1+\frac{1}{B^n}\right)_\ell - 1 < B^\ell \log_e\left(1+\frac{1}{B^n}\right).$$

Thus, if

$$\underline{\log_e}\left(1-\frac{1}{B^n}\right)_\ell + 2 < x_\ell < \underline{\log_e}\left(1+\frac{1}{B^n}\right)_\ell - 2$$

then

$$B^\ell \log_e\left(1-\frac{1}{B^n}\right) + 1 < x_\ell < B^\ell \log_e\left(1+\frac{1}{B^n}\right) - 1.$$

Furthermore, we have

$$\frac{x_\ell - 1}{B^\ell} < x < \frac{x_\ell + 1}{B^\ell}$$

thus we have

$$\log_e\left(1-\frac{1}{B^n}\right) < x < \log_e\left(1+\frac{1}{B^n}\right)$$

and

$$1-\frac{1}{B^n} < \exp(x) < 1+\frac{1}{B^n}$$

that is to say $B^n - 1 < B^n \exp(x) < B^n + 1$ and B^n satisfies the bounds property of $\exp(x)$ for order n.

Since the distinction of this case should precisely avoid the computation (and the non-termination) of msd(0), the interval including x_ℓ is not authorized to be empty and the choice of ℓ will ensure this fact.

We will indeed prove that the choice of ℓ ensures that the distance between the two ends of this interval is greater or equal to 1 and consequently there is at least one integer in this interval. So let us evaluate this distance. We have

$$L = \left(\underline{\log_e}\left(1+\frac{1}{B^n}\right)_\ell - 2\right) - \left(\underline{\log_e}\left(1-\frac{1}{B^n}\right)_\ell + 2\right)$$

that is to say

$$L = \underline{\log_e}\left(1+\frac{1}{B^n}\right)_\ell - \underline{\log_e}\left(1-\frac{1}{B^n}\right)_\ell - 4$$

and

$$L = \left(\underline{\log_e}\left(1+\frac{1}{B^n}\right)_\ell + 1\right) - \left(\underline{\log_e}\left(1-\frac{1}{B^n}\right)_\ell - 1\right) - 6$$

so

$$L \geq B^\ell \log_e\left(1+\frac{1}{B^n}\right) - B^\ell \log_e\left(1-\frac{1}{B^n}\right) - 6$$

that is to say that

$$L \geq B^\ell \log_e\left(\frac{B^n+1}{B^n-1}\right) - 6$$

or

$$L \geq B^\ell \log_e\left(1+\frac{2}{B^n-1}\right) - 6.$$

But since $B \geq 2$ and $n \geq 1$ we have

$$\log_e\left(1+\frac{2}{B^n-1}\right) \geq \frac{2}{B^n}$$

thus

$$L \geq B^\ell \frac{2}{B^n} - 6 = 2B^w - 6$$

and according to the definition of w, we have $B^w \geq 4$, consequently $L \geq 1$ and the interval is not empty.

We will now consider the second case of the double hypothesis. We will again have to examine what happens around 0, but when n is negative or null. We have $x < x_0 + 1$ thus if

$$x_0 \leq \left\lfloor \log_e\left(1+\frac{1}{B^n}\right)\right\rfloor - 1,$$

then

$$x < \left\lfloor \log_e \left(1 + \frac{1}{B^n}\right) \right\rfloor$$

and

$$x < \log_e \left(1 + \frac{1}{B^n}\right),$$

so

$$\exp(x) < 1 + \frac{1}{B^n}.$$

Furthermore $\exp(x) > 0$ and $1 - 1/B^n < 0$ so we have the following inequality

$$1 - \frac{1}{B^n} < \exp(x) < 1 + \frac{1}{B^n}$$

and B^n satisfies the bounds property of $\exp(x)$ for order n.

We will now consider the general case. We have

$$\frac{x_k - 1}{B^k} < x < \frac{x_k + 1}{B^k}$$

and $x \mapsto \exp(x)$ is an increasing function

$$\exp\left(\frac{x_k - 1}{B^k}\right) < \exp(x) < \exp\left(\frac{x_k + 1}{B^k}\right). \tag{4.2}$$

But $k \geq 0$, thus $0 < 1/B^k \leq 1$, and for $z \in]0, 1]$, we have $\exp(-z) \geq 1 - z$ and $\exp(z) \leq 1 + e\,z$, thus

$$\exp\left(-\frac{1}{B^k}\right) \geq \left(1 - \frac{1}{B^k}\right) \tag{4.3}$$

and

$$\exp\left(\frac{1}{B^k}\right) \leq \left(1 + \frac{e}{B^k}\right). \tag{4.4}$$

Furthermore $\exp\left(\frac{x_k}{B^k}\right) \geq 0$, so

$$\exp\left(\frac{x_k - 1}{B^k}\right) = \exp\left(\frac{x_k}{B^k}\right) \times \exp\left(-\frac{1}{B^k}\right) \geq \exp\left(\frac{x_k}{B^k}\right) \times \left(1 - \frac{1}{B^k}\right) \tag{4.5}$$

and

$$\exp\left(\frac{x_k + 1}{B^k}\right) = \exp\left(\frac{x_k}{B^k}\right) \times \exp\left(\frac{1}{B^k}\right) \leq \exp\left(\frac{x_k}{B^k}\right) \times \left(1 + \frac{e}{B^k}\right). \tag{4.6}$$

Moreover, according to the definition of the $\underline{\exp}$ function, we have

$$\frac{\underline{\exp}\left(\frac{x_k}{B^k}\right)_p - 1}{B^p} < \exp\left(\frac{x_k}{B^k}\right) < \frac{\underline{\exp}\left(\frac{x_k}{B^k}\right)_p + 1}{B^p}. \tag{4.7}$$

But according to the hypothesis of the general case

$$\underline{\exp}\left(\frac{x_k}{B^k}\right)_p > 0 \tag{4.8}$$

and any term of the two preceding inequalities are positive. So we can combine the inequalities (4.2, 4.3, 4.4, 4.5, 4.6, 4.7, 4.8) by multiplication and we obtain

$$B^{n-p}\left(\underline{\exp}\left(\frac{x_k}{B^k}\right)_p - 1\right)\left(1 - \frac{1}{B^k}\right) < B^n \exp(x) < B^{n-p}\left(\underline{\exp}\left(\frac{x_k}{B^k}\right)_p + 1\right)\left(1 + \frac{e}{B^k}\right).$$

Let us name $v = \underline{\exp}(x_k/B^k)_p$, we have

$$B^{n-p}(v-1)\left(1-\frac{1}{B^k}\right) < B^n \exp(x) < B^{n-p}(v+1)\left(1+\frac{e}{B^k}\right)$$

We define α, α', β, β' as follows:

$$\begin{aligned}
\alpha &= (v-1)\left(1-\frac{1}{B^k}\right) \\
\alpha' &= B^{n-p}\alpha \\
\beta &= (v+1)\left(1+\frac{e}{B^k}\right) \\
\beta' &= B^{n-p}\beta.
\end{aligned}$$

We will now prove that $\beta' - \alpha' \leq 1$ to prepare application of property 11. We have

$$\beta - \alpha = 2 + \frac{e+1}{B^k}\,v + \frac{e-1}{B^k} \leq 2 + \frac{e+1}{B^k}\,(v+1)$$

But according to the definition of k we have for $B > 2$

$$k \geq p + 1 + \left\lceil d + \log_B\left(\frac{e+1}{B-2}\right)\right\rceil,$$

thus

$$\frac{e+1}{B^k} \leq \frac{B-2}{B^{d+p+1}} \tag{4.9}$$

and for $B = 2$

$$k \geq p + 1 + \lceil d + \log_B(e+1)\rceil,$$

thus

$$\frac{e+1}{B^k} \leq \frac{1}{B^{d+p+1}}. \tag{4.10}$$

Furthermore, according to the definition of $\underline{\exp}$,

$$v = \underline{\exp}\left(\frac{x_k}{B^k}\right)_p < B^p\,\exp\left(\frac{x_k}{B^k}\right) + 1.$$

But $k \geq \mathrm{msd}(x)$ so according to the property 13, we have

$$B^{k-\mathrm{msd}(x)} \leq |x_k| \leq 2B\,B^{k-\mathrm{msd}(x)}.$$

If $x_{\mathrm{msd}(x)} \geq 1$, then

$$\frac{x_k}{B^k} \leq \frac{2B}{B^{msd(x)}}$$

and if $x_{\mathrm{msd}(x)} \leq -1$, then

$$\frac{x_k}{B^k} \leq -\frac{1}{B^{\mathrm{msd}(x)}}.$$

In both case and according to the definition of d, we have

$$\exp\left(\frac{x_k}{B^k}\right) \leq B^d.$$

Consequently, we have

$$v < B^{p+d} + 1. \tag{4.11}$$

We combine 4.11 and 4.10 for $B = 2$ and we obtain

$$\beta - \alpha \leq 2 + \frac{1}{B^{p+d+1}}\left(B^{p+d} + 2\right) \leq 2 + \frac{1}{B}\,\left(1 + \frac{2}{B^{p+d}}\right).$$

But $p+d \geq 0$, thus

$$\frac{2}{B^{p+d}} \leq 2$$

and

$$\beta - \alpha \leq 2 + \frac{1}{B}(1+2) = 2 + \frac{3}{B} = \frac{5}{2} \leq 4 = B^2$$

and $\beta' - \alpha' \leq B^2 \times B^{n-p}$. But $p \geq n+2$ thus $\beta' - \alpha' \leq 1$ for $B = 2$.

If $B > 2$, we combine 4.11 and 4.9 to obtain

$$\beta - \alpha \leq 2 + \frac{B-2}{B^{p+d+1}}\left(B^{p+d}+2\right) \leq 2 + \frac{B-2}{B}\left(1 + \frac{2}{B^{p+d}}\right)$$

But $p+d \geq 0$, thus

$$\frac{2}{B^{p+d}} \leq 2$$

and

$$\beta - \alpha \leq 2 + 3 \times \frac{B-2}{B} = \frac{5B-6}{B} = 5 - \frac{6}{B} \leq B$$

because any integer $B \geq 2$ is outside (or equal to the ends) of the interval between the roots of the second degree equation $B^2 - 5B + 6 = 0$. So we deduce that

$$\beta - \alpha \leq B$$

and $\beta' - \alpha' \leq B \times B^{n-p}$. But $p \geq n+1$ thus $\beta' - \alpha' \leq 1$.

Consequently $\alpha' < B^n \exp(x) < \beta'$ and $\beta' - \alpha' \leq 1$ independently of the value of B. We obtain the final result by application of property 11, then $\lceil \alpha' \rceil$ satisfies the bounds property of $\exp(x)$ for order n. □

4.3.4 Logarithm to base B'

Let B' be a real number greater or equal to 2 and x be a strictly positive real number represented by the sequence $(x_n)_{n\in\mathbb{Z}}$, $\log_{B'}(x)$ is represented by the sequence $(\overline{\log_{B'} x}_n)_{n\in\mathbb{Z}}$ such that:

$$\overline{\log_{B'} x}_n = \left\lfloor (\underline{\log_{B'}}(x_k/B^k)_{n+w} + 1)/B^w + \log_{B'}(e)\, B^n/x_k \right\rfloor$$

with $k = n + \mathrm{msd}(x) + w$, $w = c - \min(0,n)$ and $c = \begin{cases} 2 & \text{if } B \geq 3 \\ 3 & \text{if } B = 2. \end{cases}$

REMARKS.

1. *The computation terminates if x is a finite strictly positive real number.*
2. *Practically, we implement $\underline{\log_{B'}}$ only for the case $B' = e$ and if we want to compute the logarithm in another base B' of a real number, we deduce $\underline{\log_{B'}}$ from $\underline{\log_e}$ and then the corresponding function $\overline{\log_{B'}}$ or we deduce $\overline{\log_e}$ and then the function $\overline{\log_{B'}}$.*

Theorem 18 *For any integer n, we have*

$$(\overline{\log_{B'} x}_n - 1)B^{-n} < \log_{B'}(x) < (\overline{\log_{B'} x}_n + 1)B^{-n}.$$

Proof.

We have

$$\frac{x_k - 1}{B^k} < x < \frac{x_k + 1}{B^k}.$$

The function $x \mapsto \log_{B'}(x)$ is an increasing one on $\mathbb{R}_+^*$, thus

$$\log_{B'}\left(\frac{x_k - 1}{B^k}\right) < \log_{B'}(x) < \log_{B'}\left(\frac{x_k + 1}{B^k}\right)$$

so

$$\log_{B'}\left(\frac{x_k}{B^k}\right)+\log_{B'}\left(1-\frac{1}{x_k}\right)<\log_{B'}(x)<\log_{B'}\left(\frac{x_k}{B^k}\right)+\log_{B'}\left(1+\frac{1}{x_k}\right)$$

and

$$\log_{B'}\left(\frac{x_k}{B^k}\right)+\log_{B'}\left(1-\frac{1}{x_k}\right)\leq\log_{B'}\left(\frac{x_k}{B^k}\right)+\frac{\log_{B'}(e)}{x^k}$$

since for any positive real number z we have $\log_{B'}(z)\leq\log_{B'}(e)\times z$.

Furthermore, according to the definition of $\underline{\log_{B'}}$, we have

$$\frac{\underline{\log_{B'}}\left(\frac{x_k}{B^k}\right)_{n+w}-1}{B^{n+w}}<\log_{B'}\left(\frac{x_k}{B^k}\right)<\frac{\underline{\log_{B'}}\left(\frac{x_k}{B^k}\right)_{n+w}+1}{B^{n+w}}.$$

We combine these two inequalities and multiply each term by B^n, we obtain

$$\frac{\underline{\log_{B'}}\left(\frac{x_k}{B^k}\right)_{n+w}-1}{B^w}+B^n\log_{B'}(1-\frac{1}{x_k})<B^n\log_{B'}(x)<\frac{\underline{\log_{B'}}\left(\frac{x_k}{B^k}\right)_{n+w}+1}{B^w}+\frac{\log_{B'}(e)B^n}{x_k}.$$

Let us name

$$v=\frac{\underline{\log_{B'}}\left(\frac{x_k}{B^k}\right)_{n+w}}{B^w},$$

then the preceding inequality can be rewritten as

$$v-\frac{1}{B^w}+B^n\ \log_{B'}\left(1-\frac{1}{x_k}\right)<B^n\ \log_{B'}(x)<v+\frac{1}{B^w}+\frac{\log_{B'}(e)\,B^n}{x_k}.$$

We define α and β as follows:

$$\begin{aligned}\alpha &= v-\frac{1}{B^w}+B^n\ \log_{B'}\left(1-\frac{1}{x_k}\right)\\ \beta &= v+\frac{1}{B^w}+\frac{\log_{B'}(e)\,B^n}{x_k}.\end{aligned}$$

We will prove that $\beta-\alpha<1$. We have

$$\beta-\alpha=\frac{2}{B^w}+B^n\ \left(\frac{\log_{B'}(e)}{x_k}-\log_{B'}\left(1-\frac{1}{x_k}\right)\right)<\frac{2}{B^w}+B^n\ \log_{B'}(e)\ \left(\frac{1}{x_k}+\frac{1}{x_k-1}\right),$$

so

$$\beta-\alpha<\frac{2}{B^w}+\frac{2\,B^n\ \log_{B'}(e)}{x_k-1}.$$

But $w=c-\min(0,n)$ thus $n+w\geq c$ and $k\geq\mathrm{msd}(x)$. Consequently, according to property 13, we have $x_k\geq B^{k-\mathrm{msd}(x)}$ and

$$\beta-\alpha<\frac{2}{B^w}+\frac{2\,B^n\ \log_{B'}(e)}{B^{k-\mathrm{msd}(x)}-1}=\frac{2}{B^w}+\frac{2\ \log_{B'}(e)}{B^{k-n-\mathrm{msd}(x)}-B^{-n}}$$

and, according to the definition of k

$$\beta-\alpha<\frac{2}{B^w}+\frac{2\ \log_{B'}(e)}{B^w-B^{-n}}=\frac{2}{B^w}\left(1+\frac{\log_{B'}(e)}{1-\frac{1}{B^{n+w}}}\right).$$

But $n+w\geq c$ and $w\geq c$ thus

$$\beta-\alpha<\frac{2}{B^c}\left(1+\frac{\log_{B'}(e)}{1-\frac{1}{B^c}}\right).$$

We have also $\frac{1}{B^c}\geq\frac{1}{8}$ (minimum reached for $B=2$) and thus

$$\beta-\alpha<\frac{2}{8}\left(1+\frac{8}{7}\ \log_{B'}(e)\right)=\frac{1}{4}+\frac{\log_{B'}(e)}{7}<\frac{1}{4}+\frac{\log_2(e)}{7}<1.$$

Consequently $0<\beta-\alpha<1$ and we obtain the final result by application of property 11. □

4.3.5 Inverse trigonometric functions: the arctangent function

Let x be a real number represented by the sequence $(x_n)_{n\in\mathbb{Z}}$, $\arctan(x)$ is represented by the sequence $(\overline{\arctan(x)}_n)_{n\in\mathbb{Z}}$ such that:

If $x_k = 0$
then $\overline{\arctan(x)}_n = 0$
else $\overline{\arctan(x)}_n = \left\lfloor \frac{\underline{\arctan}(x_k/B^k)_{n+w}+1)}{B^w} + \frac{B^{n+k}}{B^{2n+2}+x_k^2+x_k} \right\rfloor$
with $k = \max(0, n+w)$ and $w = \begin{cases} 1 & \text{if } B \geq 4 \\ 2 & \text{if } B = 2 \text{ or } B = 3. \end{cases}$

Theorem 19 *For any integer n, we have*

$$(\overline{\arctan(x)}_n - 1)B^{-n} < \arctan(x) < (\overline{\arctan(x)}_n + 1)B^{-n}.$$

We will have to use the following lemma in our proof:

Lemma 2 *For any real numbers a and b, we have:*

– *If $1 + a(a+b) > 0$, then*

$$\arctan(a+b) = \arctan(a) + \arctan\left(\frac{b}{1+a(a+b)}\right).$$

– *If $1 + a(a+b) < 0$, else*

$$\arctan(a+b) = \arctan(a) + \arctan\left(\frac{b}{1+a(a+b)}\right) - sign(a) \times \pi.$$

Proof (*Lemma*).
We will apply the tan function to each part of these equalities. We have

$$\tan(\arctan(a+b)) = a+b$$

and let T be the tangent of the second expression:

$$T = \tan\left(\arctan(a) + \arctan\left(\frac{b}{1+a(a+b)}\right)\right),$$

we have

$$T = \frac{\tan(\arctan(a)) + \tan\left(\arctan\left(\frac{b}{1+a(a+b)}\right)\right)}{1 - \tan(\arctan(a)) \times \tan\left(\arctan\left(\frac{b}{1+a(a+b)}\right)\right)} = \frac{a + \frac{b}{1+a(a+b)}}{1 - a \times \frac{b}{1+a(a+b)}}$$

and we multiply the numerator and the denominator by $1 + a(a+b)$, so we obtain

$$T = \frac{a(1+a(a+b)) + b}{(1+a(a+b)) - ab} = \frac{a + a^2(a+b) + b}{1 + a^2 + ab - ab} = \frac{(a+b)(1+a^2)}{1+a^2} = a+b.$$

Furthermore, the tan function is periodic with period π, thus the tangent of each expression is always equal to $a+b$. It remains to be proved that in both cases the right part of the equality is in the interval $]-\pi/2, \pi/2[$ to conclude that this expression is equal to the principal determination of $\arctan(a+b)$ and to ends the proof of this lemma. Consequently we will study for a fixed value of a, the f_a function of variable b defined as follows:

$$f_a(b) = \arctan(a) + \arctan\left(\frac{b}{1+a(a+b)}\right) - \arctan(a+b)$$

for $b \neq -\frac{1+a^2}{a}$. The derivative f'_a of this function is

$$f'_a(b) = \frac{\frac{1}{1+a(a+b)} - \frac{ab}{(1+a(a+b))^2}}{1+\left(\frac{b}{1+a(a+b)}\right)^2} - \frac{1}{1+(a+b)^2}$$

and we multiply the numerator and the denominator by $(1+a(a+b))^2$ and we obtain

$$f'_a(b) = \frac{(1+a(a+b)-ab}{(1+a(a+b))^2+b^2} - \frac{1}{1+(a+b)^2}$$

and we develop the numerator and denominator of the first fraction

$$f'_a(b) = \frac{1+a^2}{1+2a(a+b)+a^2(a+b)^2+b^2} - \frac{1}{1+(a+b)^2}$$

then

$$f'_a(b) = \frac{1+a^2}{1+2a^2+2ab+a^4+2a^3b+a^2b^2+b^2} - \frac{1}{1+(a+b)^2}$$

and we regroup the terms of the denominator

$$f'_a(b) = \frac{1+a^2}{(1+2ab+a^2+b^2)\times(1+a^2)} - \frac{1}{1+(a+b)^2}$$

then we simplify the numerator and the denominator of the first fraction

$$f'_a(b) = \frac{1}{1+2ab+a^2+b^2} - \frac{1}{1+(a+b)^2}$$

and we regroup the terms of the denominator of the first fraction

$$f'_a(b) = \frac{1}{1+(a+b)^2} - \frac{1}{1+(a+b)^2} = 0.$$

Thus the f'_a function is null in any point of the definition domain of f_a. Furthermore we have

$$\lim_{b\to-\infty} f_a(b) = \arctan(a) + \lim_{b\to-\infty} \arctan\left(\frac{\frac{1}{1+a^2}}{\frac{b}{}+a}\right) - \lim_{b\to-\infty} \arctan(a+b)$$

so

$$\lim_{b\to-\infty} f_a(b) = \arctan(a) + \arctan\left(\frac{1}{a}\right) - \frac{\pi}{2} = sign(a)\times\frac{\pi}{2} - \frac{\pi}{2} = (sign(a)-1)\times\frac{\pi}{2}$$

and

$$\lim_{b\to+\infty} f_a(b) = \arctan(a) + \lim_{b\to+\infty} \arctan\left(\frac{\frac{1}{1+a^2}}{\frac{b}{}+a}\right) + \lim_{b\to+\infty} \arctan(a+b)$$

thus

$$\lim_{b\to+\infty} f_a(b) = \arctan(a) + \arctan\left(\frac{1}{a}\right) + \frac{\pi}{2} = sign(a)\times\frac{\pi}{2} + \frac{\pi}{2} = (sign(a)+1)\times\frac{\pi}{2}.$$

Consequently f_a varies as follows:

b	-∞	$-\frac{1+a^2}{a}$	∞
$f'_a(b)$	0	0	
$f_a(b)$	$(sign(a)+1)\times\frac{\pi}{2}$		$(sign(a)-1)\times\frac{\pi}{2}$

Thus we have

– If $a > 0$, then $\arctan(a) + \arctan(\frac{1}{a}) = \frac{\pi}{2}$ and $f_a(b) = 0$ if $b > -\frac{1+a^2}{a}$ and π otherwise.

– If $a < 0$, then $\arctan(a) + \arctan(\frac{1}{a}) = -\frac{\pi}{2}$ and $f_a(b) = 0$ if $b < -\frac{1+a^2}{a}$ and $-\pi$ otherwise.

– If $a = 0$, then $\arctan(a) = 0$, $\arctan(\frac{b}{1+a(a+b)}) = \arctan(b)$ and $\arctan(a+b)$=$\arctan(b)$, thus $f_a(b) = 0$.

That is to say that $f_a(b) = 0$ if $a > 0$ and $b > -\frac{1+a^2}{a}$, if $a < 0$ and $b < -\frac{1+a^2}{a}$ or if $a = 0$, and in the other cases, if $b \neq -\frac{1+a^2}{a}$, then $f_a(b) = sign(a) \times \pi$. All the conditions to ensure $f_a(b) = 0$ are equivalent to the inequality $1 + a(a+b) > 0$ and the proof of this lemma ends. □

Proof (*Correction of the* arctan *algorithm*).
We have

$$\frac{x_k - 1}{B^k} < x < \frac{x_k + 1}{B^k}$$

and arctan is a strictly increasing function thus

$$\arctan\left(\frac{x_k - 1}{B^k}\right) < \arctan(x) < \arctan\left(\frac{x_k + 1}{B^k}\right).$$

If $x_k = 0$, then we have:

$$-B^n \arctan\left(\frac{1}{B^k}\right) < B^n \arctan(x) < B^n \arctan\left(\frac{1}{B^k}\right)$$

and, since $k \geq 0$, thus $0 < 1/B^k \leq 1$, and $\arctan(z) < z$ is true on this interval so

$$B^n \arctan\left(\frac{1}{B^k}\right) < B^{n-k} \leq \frac{1}{B} < 1$$

since $k \geq n+1$. Consequently 0 satisfies the bounds property of $\arctan(x)$ for order n.

We consider now the case $x_k \neq 0$. We will apply the preceding lemma to the computation of $\arctan\left(\frac{x_k-1}{B^k}\right)$ and $\arctan\left(\frac{x_k+1}{B^k}\right)$ with $a = x_k/B^k$, $b = \varepsilon/B^k$ and $\varepsilon \in \{1, -1\}$. We will first determine in what case of this lemma we are before to apply it:

$$1 + a\,(a+b) = 1 + \frac{x_k}{B^k}\frac{x_k + \varepsilon}{B^k} = \frac{B^{2k} + x_k^2 + \varepsilon x_k}{B^k}.$$

But the discriminant of the equation $X^2 + \varepsilon X + B^{2k} = 0$ is equal to $\Delta = 1 - 4\,B^{2k} < 1 - 4\,B^0 = -3 < 0$ since $k \geq 0$. Consequently the polynom $X^2 + \varepsilon X + B^{2k}$ is strictly positive for any real value of X and particularly in x_k thus we apply the first part of the lemma and obtain

$$\arctan\left(\frac{x_k - 1}{B^k}\right) = \arctan\left(\frac{x_k}{B^k}\right) + \arctan\left(\frac{-\frac{1}{B^k}}{1 + \frac{x_k}{B^k}\frac{x_k - 1}{B^k}}\right)$$

that is to say

$$\arctan\left(\frac{x_k - 1}{B^k}\right) = \arctan\left(\frac{x_k}{B^k}\right) - \arctan\left(\frac{B^k}{B^{2k} + x_k^2 - x_k}\right)$$

and

$$\arctan\left(\frac{x_k + 1}{B^k}\right) = \arctan\left(\frac{x_k}{B^k}\right) + \arctan\left(\frac{\frac{1}{B^k}}{1 + \frac{x_k}{B^k}\frac{x_k + 1}{B^k}}\right)$$

so

$$\arctan\left(\frac{x_k + 1}{B^k}\right) = \arctan\left(\frac{x_k}{B^k}\right) + \arctan\left(\frac{B^k}{B^{2k} + x_k^2 + x_k}\right).$$

Consequently, we have

$$\arctan\left(\frac{x_k}{B^k}\right) - \arctan\left(\frac{B^k}{B^{2k}+x_k^2-x_k}\right) < \arctan(x)$$

and

$$\arctan(x) < \arctan\left(\frac{x_k}{B^k}\right) + \arctan\left(\frac{B^k}{B^{2k}+x_k^2+x_k}\right).$$

Furthermore, according to the definition of $\underline{\arctan}$, we have

$$\frac{\underline{\arctan}\left(\frac{x_k}{B^k}\right)_{n+w} - 1}{B^{n+w}} < \arctan\left(\frac{x_k}{B^k}\right) < \frac{\underline{\arctan}\left(\frac{x_k}{B^k}\right)_{n+w} + 1}{B^{n+w}}.$$

We combine these two inequalities and multiply each part by B^n, we obtain

$$B^n\left(\frac{\underline{\arctan}\left(\frac{x_k}{B^k}\right)_{n+w} - 1}{B^{n+w}} - \arctan\left(\frac{B^k}{B^{2k}+x_k^2-x_k}\right)\right) < B^n\arctan(x)$$

and

$$B^n\arctan(x) < B^n\left(\frac{\underline{\arctan}\left(\frac{x_k}{B^k}\right)_{n+w} + 1}{B^{n+w}} + \arctan\left(\frac{B^k}{B^{2k}+x_k^2+x_k}\right)\right).$$

But for any positive real number z we have $\arctan(z) < z$, and we obtain the following inequality:

$$B^n\left(\frac{\underline{\arctan}\left(\frac{x_k}{B^k}\right)_{n+w} - 1}{B^{n+w}} - \frac{B^k}{B^{2k}+x_k^2-x_k}\right) < B^n\arctan(x)$$

and

$$B^n\arctan(x) < B^n\left(\frac{\underline{\arctan}\left(\frac{x_k}{B^k}\right)_{n+w} + 1}{B^{n+w}} + \frac{B^k}{B^{2k}+x_k^2+x_k}\right).$$

We define α and β as follows:

$$\alpha = B^n\left(\frac{\underline{\arctan}\left(\frac{x_k}{B^k}\right)_{n+w} - 1}{B^{n+w}} - \frac{B^k}{B^{2k}+x_k^2-x_k}\right)$$

$$\beta = B^n\left(\frac{\underline{\arctan}\left(\frac{x_k}{B^k}\right)_{n+w} + 1}{B^{n+w}} + \frac{B^k}{B^{2k}+x_k^2+x_k}\right).$$

We will prove that $\beta - \alpha < 1$ to apply property 11 and conclude. We have

$$\beta - \alpha = \frac{2}{B^w} + B^{n+k}\left(\frac{1}{B^{2k}+x_k^2+x_k} + \frac{1}{B^{2k}+x_k^2-x_k}\right) = \frac{2}{B^w} + \frac{2\,B^{n+k}\,(B^{2k}+x_k^2)}{(B^{2k}+x_k^2)^2 - x_k^2}.$$

The function $x \mapsto \frac{b+x}{(b+x)^2-x}$ is strictly increasing on $\mathbb{R}_+$ with maximum value $\frac{1}{b}$. Consequently, if $b = B^{2k}$ we have

$$\beta - \alpha < \frac{2}{B^w} + \frac{2\,B^{n+k}}{B^{2k}} = \frac{2}{B^w} + 2\,B^{n-k}.$$

But $k \geq n + w$, thus

$$\beta - \alpha \leq \frac{4}{B^w} \leq 1$$

and we conclude by application of property11. □

REMARKS.

1. *We can deduce π for example by the following formula due to Gauss:*

$$\frac{\pi}{4} = 12 \arctan\left(\frac{1}{18}\right) + 8 \arctan\left(\frac{1}{57}\right) - 5 \arctan\left(\frac{1}{239}\right).$$

2. *Jean-Christophe Filliâtre suggests an improvement for arguments of absolute value greater than or equal to 1 in [20].*

4.3.6 Direct trigonometric functions: the sine function

Choice

We can compute the direct trigonometric functions in two different ways:

- The first one consist in computing the sine function for real numbers between 0 and $\pi/2$, deduce it elsewhere and deduce the cosine and tangent functions ultimately.

  The initial interval guarantee the monotony of the sine function and a reasonable combination of inequalities. The $\underline{\sin}$ function is easy to compute using sin Taylor expansion. The computation of sin for arguments out of the interval $[0, \pi/2]$ is a little complicated.

- The second one consist in computing the tangent between 0 and $\pi/2$, deduce its value elsewhere and deduce the sine and cosine functions ultimately.

  The initial interval guarantee not only the monotony of the tangent function but also the definition of the function. The sine and cosine functions are rational functions of the half-arc tangent.

  The computation of $\sin(\pi)$ or $\cos(\pi)$, implies the computation of $\tan(\pi/2)$ so it is difficult to forecast the behavior of the algorithm for values near to π: maybe the computation will not terminate or its duration will be prohibitive.

  Moreover, $\underline{\tan}$ is uneasy to compute, we have to deduce it by division of a limited Taylor expansion of the sine and cosine functions.

  Finally, if you compute $\overline{\sin(x)}$, you will have to deduce it from $\overline{\tan(x/2)}$ that you deduce from $\underline{\tan(x/2)}$ that is finally deduced from $\underline{\sin(x/2)}$ and $\underline{\cos(x/2)}$, this seems less efficient that the first approach.

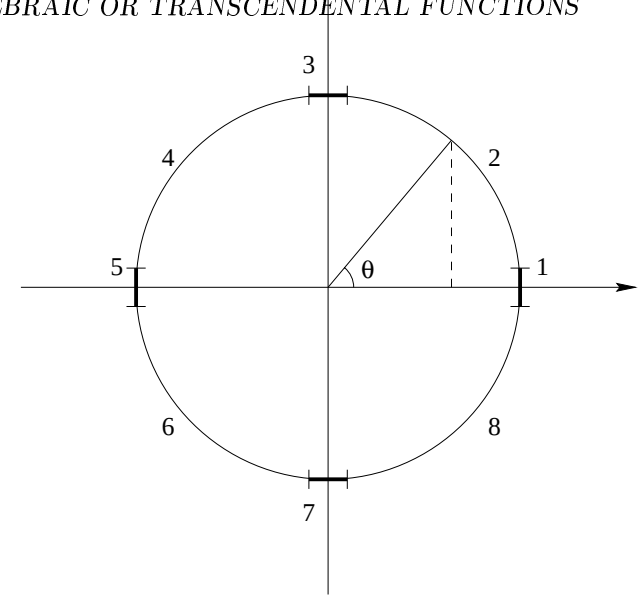
The second approach seems to have several major defaults and we have chosen the first approach to describe the sine function.

Description

Let x be a real number represented by the sequence $(x_n)_{n\in\mathbb{Z}}$. We note $p = \overline{\left(\frac{x}{\pi}\right)}_0 - 1$, $\theta = x - p\pi$ and $z = \frac{\pi}{2}$. We represent $\sin(x)$ by the sequence $(\overline{\sin(x)}_n)_{n\in\mathbb{Z}}$ such that:

If $0 \leq \theta_k \leq 1$, $4z_k - 4 \leq \theta_k \leq 4z_k + 4$ or $2z_k - 2 \leq \theta_k \leq 2z_k + 2$,
then $\overline{\sin(x)}_n = 0$
else if $2 \leq \theta_k \leq z_k - 2$ or $z_k + 2 \leq \theta_k \leq 2z_k - 3$,
then $\overline{\sin(x)}_n = (-1)^p \left\lfloor \dfrac{\underline{\sin}\left(\dfrac{\theta_k}{B^k}\right)_{n+w} + 1}{B^w} + B^{n-k} \right\rfloor$
else if $z_k - 1 \leq \theta_k \leq z_k + 1$,
then $\overline{\sin(x)}_n = (-1)^p B^n$
else if $2z_k + 3 \leq \theta_k \leq 3z_k - 4$ or $3z_k + 4 \leq \theta_k \leq 4z_k - 5$,
then $\overline{\sin(x)}_n = (-1)^p \left\lceil \dfrac{\underline{\sin}\left(\dfrac{\theta_k}{B^k}\right)_{n+w} - 1}{B^w} - B^{n-k} \right\rceil$
else if $3z_k - 3 \leq \theta_k \leq 3z_k + 3$,
then $\overline{\sin(x)}_n = (-1)^{p+1} B^n$

with $k = \max(c, n + w)$ and $(c, w) = \begin{cases} (2,2) & \text{if } B \geq 3 \\ (3,4) & \text{if } B = 2. \end{cases}$



Legend for the zones:

zone 1 : $0 \leq \theta_k \leq 1$ or $4z_k - 4 \leq \theta \leq 4z_k + 4$, $\sin(\theta)$ is "closed" to 0

zone 2 : $2 \leq \theta_k \leq z_k - 2$, the sine function is increasing and positive "around" θ

zone 3 : $z_k - 1 \leq \theta_k \leq z_k + 1$, $\sin(\theta)$ is "closed" to 1

zone 4 : $z_k + 2 \leq \theta_k \leq 2z_k - 3$, the sine function is decreasing and positive "around" θ

zone 5 : $2z_k - 2 \leq \theta_k \leq 2z_k + 2$, $\sin(\theta)$ is "closed" to 0

zone 6 : $2z_k + 3 \leq \theta_k \leq 3z_k - 4$, the sine function is decreasing and negative "around" θ

zone 7 : $3z_k - 3 \leq \theta_k \leq 3z_k + 3$, $\sin(\theta)$ is "closed" to -1

zone 8 : $3z_k + 4 \leq \theta_k \leq 4z_k - 5$, the sine function is increasing and negative "around" θ.

Figure 4.1: The eight intervals of the trigonometric circle for the computation of the sine function.

REMARK.
To manage the difficulty of the non uniform monotonicity of the sine function, we reduce the number to the equivalent between 0 and 2π, and then we carve the trigonometric circle in eight pieces numbered 1 to 8 on figure 4.1. $\diamond$

Theorem 20 *For any integer n, we have*

$$(\overline{\sin(x)}_n - 1)B^{-n} < \sin(x) < (\overline{\sin(x)}_n + 1)B^{-n}.$$

Proof.
First of all, we will prove that $0 < \theta < 2\pi$. We have

$$p = \overline{\left(\frac{x}{\pi}\right)}_0 - 1 < \frac{x}{\pi} < \overline{\left(\frac{x}{\pi}\right)}_0 + 1 = p + 2$$

thus $p\pi < x < p\pi + 2\pi$ and $0 < \theta = x - p\pi < 2\pi$. More precisely, $0 < \theta < 4z$ so $0 < \theta_k + 1$ and $\theta_k - 1 < 4(z_k + 1)$ and finally the inequality $0 \leq \theta_k \leq 4z_k + 4$.

Furthermore, if p is an even integer we have $\sin(x) = \sin(\theta)$ and $\sin(x) = \sin(\pi + \theta) = -\sin(\theta)$ if p is an odd integer thus $\sin(x) = (-1)^p \sin(\theta)$ and according to the property 10 we can transpose this equality with the *ad hoc* sign for the sequences that represent x and θ.

We will use throughout this demonstration the following facts:

1.. The inequalities $0 < \sin(1/B^k) \leq 1/B^k$ and $1 - 1/(2\,B^{2k}) \leq cos(1/B^k) \leq 1$.

2.. The formulas

$$\sin\left(\frac{\theta_k + \varepsilon}{B^k}\right) = \sin\left(\frac{\theta_k}{B^k}\right)\cos\left(\frac{1}{B^k}\right) + \varepsilon\cos\left(\frac{\theta_k}{B^k}\right)\sin\left(\frac{1}{B^k}\right)$$

where $\varepsilon \in \{1, -1\}$.

3.. The choice for k, w and c implies that $d/B^k \leq 1/B^n$ and $d/B^k < \pi/2$ for any number $d \leq 9$, $d/2B^{n-2k} < 1$ for $d \leq 49$ and $2/B^w + 2B^{n-k} + B^{n-2k}/2 < 1$.

Let us consider each of these properties:

(a) If $d \leq 9$, then $d/B^k \leq 9/B^k$ thus it is sufficient to prove that $9/B^k < 1/B^n$ et $d/B^k < \pi/2$. According to the definition of k, we have $k \geq n + w$ and $k \geq c$ thus it is sufficient to prove that $9/B^w < 1$ and $9/B^c < \pi/2$. We will now distinguish according to whether B is equal to 2 or not. If $B = 2$, then $c = 3$ and $w = 4$, thus $9/B^w = 9/16 < 1$ and $9/B^c = 9/8 < \pi/2$. If $B \geq 3$, then $c = 2$ et $w = 2$, thus $9/B^w \leq /3^2 \leq 1$ and $9/B^c = 1 < \pi/2$.

(b) In the same way, if $d \leq 49$, then $d/2B^{n-2k} \leq 49/2B^{n-2k}$ thus it is sufficient to prove that $49/2B^{n-2k} < 1$. According to the definition of k, we have $k \geq n+w$ and $k \geq c$, thus it is sufficient to prove that $49/2 < B^{w+c}$. Let us distinguish according to whether B is equal to 2 or not. If $B = 2$, then $c = 3$ and $w = 4$, and then $B^{w+c} = 2^7 = 128 > 49/2$. If $B \geq 3$, then $c = 2$ and $w = 2$, then $B^{w+c} \geq 3^4 = 81 > 49/2$.

(c) We will now consider the final inequality. According to the definition of k, we have $k \geq n + w$ and $k \geq c$, thus $2/B^w + 2B^{n-k} + B^{n-2k}/2 \leq 2/B^w + 2/B^w + 1/(2B^{w+c}) = (4 + 2/B^c)/B^w$. Let us distinguish according to whether B is equal to 2 or not. If $B = 2$, then $c = 3$ and $w = 4$ thus $(4 + 2/B^c)/B^w = 17/64 < 1$. If $B \geq 3$, then $c = 2$ and $w = 2$, so $(4 + 2/B^c)/B^w \leq (4 + 2/3^c)/3^w = 38/81 < 1$.□

We will now distinguish according to the position of θ on the trigonometric circle.

Case $0 \leq \theta_k \leq 1$**:** We have

$$-\frac{\pi}{2} < -\frac{1}{B^k} \leq \frac{\theta_k - 1}{B^k} < \theta < \frac{\theta_k + 1}{B^k} \leq \frac{2}{B^k} < \frac{\pi}{2}$$

according to the preceding remark, thus θ is in the right part of the trigonometric circle and the sine function is strictly increasing on the interval $](\theta_k - 1)B^{-k}, (\theta_k + 1)B^{-k}[$. We have

$$\sin\left(\frac{\theta_k - 1}{B^k}\right) < \sin(\theta) < \sin\left(\frac{\theta_k + 1}{B^k}\right).$$

But, we have

$$\sin\left(\frac{\theta_k + 1}{B^k}\right) \leq \frac{\theta_k + 1}{B^k} \leq \frac{2}{B^k} \leq \frac{1}{B^n}$$

and

$$\sin\left(\frac{\theta_k - 1}{B^k}\right) \geq \sin\left(-\frac{1}{B^k}\right) = -\sin\left(\frac{1}{B^k}\right) \geq -\frac{1}{B^k} \geq -\frac{1}{B^n},$$

thus

$$-\frac{1}{B^n} < \sin(\theta) < \frac{1}{B^n}$$

and 0 satisfies the bounds property of $\sin(\theta)$ for order n.

Case $2 \leq \theta_k \leq z_k - 2$**:** We have

$$\frac{1}{B^k} \leq \frac{\theta_k - 1}{B^k} < \theta < \frac{\theta_k + 1}{B^k} \leq \frac{z_k - 1}{B^k} < \frac{\pi}{2},$$

thus θ is in the upper right quarter of the trigonometric circle and the sine function is strictly increasing on the interval $](\theta_k - 1)B^{-k}, (\theta_k + 1)B^{-k}[$. We have

$$\sin\left(\frac{\theta_k - 1}{B^k}\right) < \sin(\theta) < \sin\left(\frac{\theta_k + 1}{B^k}\right).$$

According to the fact that $0 < \theta_k/B^k < \pi/2$ and that $k \geq 0$, we have $\sin\left(\theta_k/B^k\right) > 0$ and $0 < \cos\left(\theta_k/B^k\right) < 1$ thus

$$\sin\left(\frac{\theta_k - 1}{B^k}\right) = \sin\left(\frac{\theta_k}{B^k}\right)\cos\left(\frac{1}{B^k}\right) - \cos\left(\frac{\theta_k}{B^k}\right)\sin\left(\frac{1}{B^k}\right)$$

and

$$\sin\left(\frac{\theta_k + 1}{B^k}\right) = \sin\left(\frac{\theta_k}{B^k}\right)\cos\left(\frac{1}{B^k}\right) + \cos\left(\frac{\theta_k}{B^k}\right)\sin\left(\frac{1}{B^k}\right)$$

verify respectively

$$\sin\left(\frac{\theta_k - 1}{B^k}\right) > \sin\left(\frac{\theta_k}{B^k}\right)\left(1 - \frac{1}{2B^{2k}}\right) - \frac{1}{B^k}$$

and

$$\sin\left(\frac{\theta_k + 1}{B^k}\right) < \sin\left(\frac{\theta_k}{B^k}\right) + \frac{1}{B^k}.$$

Consequently

$$\sin\left(\frac{\theta_k}{B^k}\right)\left(1 - \frac{1}{2B^{2k}}\right) - \frac{1}{B^k} < \sin(\theta) < \sin\left(\frac{\theta_k}{B^k}\right) + \frac{1}{B^k}.$$

Furthermore, according to the definition of $\underline{\sin}$, we have

$$\frac{\underline{\sin}\left(\frac{\theta_k}{B^k}\right)_{n+w} - 1}{B^{n+w}} < \sin\left(\frac{\theta_k}{B^k}\right) < \frac{\underline{\sin}\left(\frac{\theta_k}{B^k}\right)_{n+w} + 1}{B^{n+w}}$$

and we combine these inequalities and multiply each term by B^n, we obtain

$$\frac{\underline{\sin}\left(\frac{\theta_k}{B^k}\right)_{n+w} - 1}{B^w}\left(1 - \frac{1}{2B^{2k}}\right) - B^{n-k} < B^n \sin(\theta) < \frac{\underline{\sin}\left(\frac{\theta_k}{B^k}\right)_{n+w} + 1}{B^w} + B^{n-k}.$$

We define α and β as follows:

$$\begin{aligned} \alpha &= \frac{\underline{\sin}\left(\frac{\theta_k}{B^k}\right)_{n+w} - 1}{B^w}\left(1 - \frac{1}{2B^{2k}}\right) - B^{n-k} \\ \beta &= \frac{\underline{\sin}\left(\frac{\theta_k}{B^k}\right)_{n+w} + 1}{B^w} + B^{n-k}. \end{aligned}$$

We will prove that $\beta - \alpha < 1$ in order to apply the property 11. We have

$$\beta - \alpha = \frac{2}{B^w} + 2\,B^{n-k} + \frac{1}{2B^{2k}}\frac{\underline{\sin}\left(\frac{\theta_k}{B^k}\right)_{n+w} - 1}{B^w}.$$

Consequently

$$\beta - \alpha < \frac{2}{B^w} + 2\,B^{n-k} + \frac{B^n}{2B^{2k}}\sin\left(\frac{\theta_k}{B^k}\right) \leq \frac{2}{B^w} + 2\,B^{n-k} + \frac{B^{n-2k}}{2} < 1$$

according to the facts established at the beginning of this proof. We conclude this case by application of property 11.

Case $z_k - 1 \leq \theta_k \leq z_k + 1$**:** We have

$$1 \geq \sin(\theta) = \cos\left(\theta - \frac{\pi}{2}\right) > 1 - \left(\frac{1}{2}\left(\theta - \frac{\pi}{2}\right)^2\right).$$

But $\theta - \frac{\pi}{2} = \theta - z$ and

$$|\theta - z| < \frac{|\theta_k - z_k| + 2}{B^k}$$

and $|\theta_k - z_k| \leq 1$, consequently

$$\left|\theta - \frac{\pi}{2}\right| < \frac{3}{B^k}$$

thus

$$1 \geq \sin(\theta) > 1 - \left(\frac{1}{2}\left(\frac{3}{B^k}\right)^2\right) = 1 - \frac{9}{2B^{2k}}.$$

Thus we have

$$B^n - \frac{9}{2}B^{n-2k} < B^n \sin(\theta) \leq B^n.$$

But $9/2B^{n-2k} < 1$ according to the facts established at the beginning of this proof and then

$$B^n - 1 < B^n \sin(\theta) \leq B^n$$

and finally B^n satisfies the bounds property of $\sin(\theta)$ for order n.

Case $z_k + 2 \leq \theta_k \leq 2z_k - 3$**:** We have

$$\frac{\pi}{2} < (z_k + 1)B^{-k} \leq (\theta_k - 1)B^{-k} < \theta < (\theta_k + 1)B^{-k} < 2(z_k - 1)B^{-k} < \pi,$$

thus θ is in the upper left quarter of the trigonometric circle and the sine function is strictly decreasing on the interval $](\theta_k - 1)B^{-k}, (\theta_k + 1)B^{-k}[$. We have

$$\sin\left(\frac{\theta_k + 1}{B^k}\right) < \sin(\theta) < \sin\left(\frac{\theta_k - 1}{B^k}\right).$$

Since $\pi/2 < \theta_k/B^k < \pi$ and $k \geq 0$, we have $\sin\left(\theta_k/B^k\right) > 0$ and $-1 < \cos\left(\theta_k/B^k\right) < 0$ thus

$$\sin\left(\frac{\theta_k + 1}{B^k}\right) = \sin\left(\frac{\theta_k}{B^k}\right)\cos\left(\frac{1}{B^k}\right) + \cos\left(\frac{\theta_k}{B^k}\right)\sin\left(\frac{1}{B^k}\right)$$

and

$$\sin\left(\frac{\theta_k - 1}{B^k}\right) = \sin\left(\frac{\theta_k}{B^k}\right)\cos\left(\frac{1}{B^k}\right) - \cos\left(\frac{\theta_k}{B^k}\right)\sin\left(\frac{1}{B^k}\right)$$

verifies respectively

$$\sin\left(\frac{\theta_k + 1}{B^k}\right) > \sin\left(\frac{\theta_k}{B^k}\right)\left(1 - \frac{1}{2B^{2k}}\right) - \frac{1}{B^k}$$

and

$$\sin\left(\frac{\theta_k - 1}{B^k}\right) < \sin\left(\frac{\theta_k}{B^k}\right) + \frac{1}{B^k}.$$

Consequently

$$\sin\left(\frac{\theta_k}{B^k}\right)\left(1 - \frac{1}{2B^{2k}}\right) - \frac{1}{B^k} < \sin(\theta) < \sin\left(\frac{\theta_k}{B^k}\right) + \frac{1}{B^k}.$$

Furthermore, according to the definition of $\underline{\sin}$, we have

$$\frac{\underline{\sin}\left(\frac{\theta_k}{B^k}\right)_{n+w} - 1}{B^{n+w}} < \sin\left(\frac{\theta_k}{B^k}\right) < \frac{\underline{\sin}\left(\frac{\theta_k}{B^k}\right)_{n+w} + 1}{B^{n+w}}$$

and we combine these two inequalities and multiply each term by B^n, we obtain

$$\frac{\underline{\sin}\left(\frac{\theta_k}{B^k}\right)_{n+w} - 1}{B^w}\left(1 - \frac{1}{2B^{2k}}\right) - B^{n-k} < B^n \sin(\theta) < \frac{\underline{\sin}\left(\frac{\theta_k}{B^k}\right)_{n+w} + 1}{B^w} + B^{n-k}.$$

We define α and β as follows:

$$\begin{aligned}\alpha &= \frac{\underline{\sin}\left(\frac{\theta_k}{B^k}\right)_{n+w} - 1}{B^w}\left(1 - \frac{1}{2B^{2k}}\right) - B^{n-k} \\ \beta &= \frac{\underline{\sin}\left(\frac{\theta_k}{B^k}\right)_{n+w} + 1}{B^w} + B^{n-k}.\end{aligned}$$

We will now prove that $\beta - \alpha < 1$ in order to apply property 11. We have

$$\beta - \alpha = \frac{2}{B^w} + 2\,B^{n-k} + \frac{1}{2B^{2k}}\frac{\underline{\sin}\left(\frac{\theta_k}{B^k}\right)_{n+w} - 1}{B^w}.$$

Consequently

$$\beta - \alpha < \frac{2}{B^w} + 2\,B^{n-k} + \frac{B^n}{2B^{2k}}\sin\left(\frac{\theta_k}{B^k}\right) \leq \frac{2}{B^w} + 2\,B^{n-k} + \frac{B^{n-2k}}{2} < 1$$

according to the facts established at the beginning of this proof and we conclude for this case by application of the property 11.

Case $2z_k - 2 \leq \theta_k \leq 2z_k + 2$**:** We have

$$\frac{\pi}{2} < \pi - \frac{5}{B^k} < \frac{2z_k - 3}{B^k} \leq \frac{\theta_k - 1}{B^k} < \theta < \frac{\theta_k + 1}{B^k} \leq \frac{2z_k + 3}{B^k} < \pi + \frac{5}{B^k} < \frac{3\pi}{2}$$

according to the facts established at the beginning of the proof. Consequently θ, $\pi - 5/B^k$ and $\pi + 5/B^k$ are each one in the left part of the trigonometric circle and the sine function is strictly decreasing on the interval $]\pi - 5/B^k, \pi + 5/B^k[$, thus

$$\sin\left(\pi + \frac{5}{B^k}\right) < \sin(\theta) < \sin\left(\pi - \frac{5}{B^k}\right).$$

But

$$\sin\left(\pi + \varepsilon\frac{5}{B^k}\right) = -\varepsilon\sin\left(\frac{5}{B^k}\right)$$

for $\varepsilon \in \{-1, 1\}$. Thus we have

$$-\sin\left(\frac{5}{B^k}\right) < \sin(\theta) < \sin\left(\frac{5}{B^k}\right).$$

But

$$\sin\left(\frac{5}{B^k}\right) < \frac{5}{B^k} \leq \frac{1}{B^n}$$

according to the facts established at the beginning of this proof and consequently

$$-\frac{1}{B^n} < \sin(\theta) < \frac{1}{B^n}$$

so 0 satisfies the bounds property of $\sin(\theta)$ for order n.

Case $2z_k + 3 \leq \theta_k \leq 3z_k - 4$**:** We have

$$\pi < 2(z_k + 1)B^{-k} \leq (\theta_k - 1)B^{-k} < \theta < (\theta_k + 1)B^{-k} < 3(z_k - 1)B^{-k} < \frac{3\pi}{2},$$

thus θ is in the left part of the trigonometric circle and the sine function is strictly decreasing on the interval $](\theta_k - 1)B^{-k}, (\theta_k + 1)B^{-k}[$. We have

$$\sin\left(\frac{\theta_k + 1}{B^k}\right) < \sin(\theta) < \sin\left(\frac{\theta_k - 1}{B^k}\right).$$

Since $\pi < \theta_k/B^k < 3\pi/2$ et $k \geq 0$, we have $\sin\left(\theta_k/B^k\right) < 0$ and $-1 < \cos\left(\theta_k/B^k\right) < 0$, thus

$$\sin\left(\frac{\theta_k}{B^k}\right) - \frac{1}{B^k} < \sin(\theta) < \sin\left(\frac{\theta_k}{B^k}\right)\left(1 - \frac{1}{2B^{2k}}\right) + \frac{1}{B^k}.$$

Furthermore, according to the definition of $\underline{\sin}$, we have

$$\frac{\underline{\sin}\left(\frac{\theta_k}{B^k}\right)_{n+w} - 1}{B^{n+w}} < \sin\left(\frac{\theta_k}{B^k}\right) < \frac{\underline{\sin}\left(\frac{\theta_k}{B^k}\right)_{n+w} + 1}{B^{n+w}}$$

and we combine these two inequalities and multiply each term by B^n, and we obtain

$$\frac{\underline{\sin}\left(\frac{\theta_k}{B^k}\right)_{n+w} - 1}{B^w} - B^{n-k} < B^n \sin(\theta) < \frac{\underline{\sin}\left(\frac{\theta_k}{B^k}\right)_{n+w} + 1}{B^w}\left(1 - \frac{1}{2B^{2k}}\right) + B^{n-k}.$$

We define α and β as follows:

$$\begin{aligned}
\alpha &= \frac{\underline{\sin}\left(\frac{\theta_k}{B^k}\right)_{n+w} - 1}{B^w} - B^{n-k} \\
\beta &= \frac{\underline{\sin}\left(\frac{\theta_k}{B^k}\right)_{n+w} + 1}{B^w}\left(1 - \frac{1}{2B^{2k}}\right) + B^{n-k}.
\end{aligned}$$

We will now prove that $\beta - \alpha < 1$ in order to apply property 11. We have

$$\beta - \alpha = \frac{2}{B^w} + 2B^{n-k} - \frac{1}{2B^{2k}} \frac{\underline{\sin}\left(\frac{\theta_k}{B^k}\right)_{n+w} + 1}{B^w}.$$

But

$$\frac{\underline{\sin}\left(\frac{\theta_k}{B^k}\right)_{n+w} + 1}{B^w} > B^n \sin\left(\frac{\theta_k}{B^k}\right) \geq -B^n$$

thus

$$\beta - \alpha < \frac{2}{B^w} + 2B^{n-k} + \frac{B^{n-2k}}{2} < 1$$

according to the facts established at the beginning of this proof and we conclude by application of property 11.

Case $3z_k - 3 \leq \theta_k \leq 3z_k + 3$**:** we have

$$-1 \leq \sin(\theta) = -\cos\left(\frac{3\pi}{2} - \theta\right) < -1 + \left(\frac{1}{2}\left(\frac{3\pi}{2} - \theta\right)^2\right)$$

But $\frac{3\pi}{2} - \theta = 3z - \theta$ and

$$|3z - \theta| < \frac{|3z_k - \theta_k| + 4}{B^k}$$

and $|3z_k - \theta_k| \leq 3$, consequently

$$\left|\frac{3\pi}{2} - \theta\right| < \frac{7}{B^k}$$

thus

$$-1 \leq \sin(\theta) < -1 + \left(\frac{1}{2}\left(\frac{7}{B^k}\right)^2\right) = -1 + \frac{49}{2B^{2k}}$$

and

$$-B^n \leq B^n \sin(\theta) < -B^n + \frac{49}{2}B^{n-2k} < -B^n + 1$$

according to the facts established at the beginning of this proof and $-B^n$ satisfies the bounds property of $\sin(\theta)$ for order n.

Case $3z_k + 4 \leq \theta_k \leq 4z_k - 5$: We have

$$\frac{3\pi}{2} < 3\frac{z_k+1}{B^k} \leq \frac{\theta_k - 1}{B^k} < \theta < \frac{\theta_k + 1}{B^k} \leq 4\frac{z_k - 1}{B^k} < 2\pi$$

thus θ is in the lower right quarter of the trigonometric circle and the sine function is strictly increasing on the interval $](\theta_k - 1)B^{-k}, (\theta_k + 1)B^{-k}[$. We have

$$\sin\left(\frac{\theta_k - 1}{B^k}\right) < \sin(\theta) < \sin\left(\frac{\theta_k + 1}{B^k}\right).$$

Since $3\pi/2 < \theta_k/B^k < 2\pi$ and $k \geq 0$, we have $\sin\left(\theta_k/B^k\right) < 0$ and $0 < \cos\left(\theta_k/B^k\right) < 1$ thus

$$\sin\left(\frac{\theta_k + 1}{B^k}\right) = \sin\left(\frac{\theta_k}{B^k}\right)\cos\left(\frac{1}{B^k}\right) + \cos\left(\frac{\theta_k}{B^k}\right)\sin\left(\frac{1}{B^k}\right)$$

and

$$\sin\left(\frac{\theta_k - 1}{B^k}\right) = \sin\left(\frac{\theta_k}{B^k}\right)\cos\left(\frac{1}{B^k}\right) - \cos\left(\frac{\theta_k}{B^k}\right)\sin\left(\frac{1}{B^k}\right)$$

verify respectively

$$\sin\left(\frac{\theta_k + 1}{B^k}\right) < \sin\left(\frac{\theta_k}{B^k}\right)\left(1 - \frac{1}{2B^{2k}}\right) + \frac{1}{B^k}$$

and

$$\sin\left(\frac{\theta_k - 1}{B^k}\right) > \sin\left(\frac{\theta_k}{B^k}\right) - \frac{1}{B^k}.$$

Consequently

$$\sin\left(\frac{\theta_k}{B^k}\right) - \frac{1}{B^k} < \sin(\theta) < \sin\left(\frac{\theta_k}{B^k}\right)\left(1 - \frac{1}{2B^{2k}}\right) + \frac{1}{B^k}.$$

Furthermore, according to the definition of $\underline{\sin}$, we have

$$\frac{\underline{\sin}\left(\frac{\theta_k}{B^k}\right)_{n+w} - 1}{B^{n+w}} < \sin\left(\frac{\theta_k}{B^k}\right) < \frac{\underline{\sin}\left(\frac{\theta_k}{B^k}\right)_{n+w} + 1}{B^{n+w}}$$

and we combine these two inequalities and multiply each term by B^n, we obtain

$$\frac{\underline{\sin}\left(\frac{\theta_k}{B^k}\right)_{n+w} - 1}{B^w} - B^{n-k} < B^n \sin(\theta) < \frac{\underline{\sin}\left(\frac{\theta_k}{B^k}\right)_{n+w} + 1}{B^w}\left(1 - \frac{1}{2B^{2k}}\right) + B^{n-k}.$$

We define α and β as follows

$$\begin{aligned}
\alpha &= \frac{\underline{\sin}\left(\frac{\theta_k}{B^k}\right)_{n+w} - 1}{B^w} - B^{n-k} \\
\beta &= \frac{\underline{\sin}\left(\frac{\theta_k}{B^k}\right)_{n+w} + 1}{B^w}\left(1 - \frac{1}{2B^{2k}}\right) + B^{n-k}.
\end{aligned}$$

We will now prove that $\beta - \alpha < 1$ in order to apply the property 11. We have

$$\beta - \alpha = \frac{2}{B^w} + 2B^{n-k} - \frac{1}{2B^{2k}}\frac{\underline{\sin}\left(\frac{\theta_k}{B^k}\right)_{n+w} + 1}{B^w}.$$

But

$$\frac{\underline{\sin}\left(\frac{\theta_k}{B^k}\right)_{n+w} + 1}{B^w} > B^n \sin\left(\frac{\theta_k}{B^k}\right) \geq -B^n$$

thus

$$\beta - \alpha < \frac{2}{B^w} + 2B^{n-k} + B^{n-2k} < 1$$

according to the facts established at the beginning of this proof and we conclude by application of property 11.

Case $4z_k - 4 \leq \theta_k \leq 4z_k + 4$**:** We have

$$\frac{3\pi}{2} \leq 2\pi - \frac{9}{B^k} < \frac{4z_k - 5}{B^k} \leq \frac{\theta_k - 1}{B^k} < \theta < \frac{\theta_k + 1}{B^k} \leq \frac{4z_k + 5}{B^k} < 2\pi + \frac{9}{B^k} \leq \frac{5\pi}{2}$$

according to the facts established at the beginning of this proof. Consequently, θ, $2\pi - 9/B^k$ and $2\pi + 9/B^k$ are each one in the right part of the trigonometric circle and the sine function is strictly increasing on the interval $]2\pi - 9/B^k, 2\pi + 9/B^k[$, so

$$\sin\left(2\pi - \frac{9}{B^k}\right) < \sin(\theta) < \sin\left(2\pi + \frac{9}{B^k}\right).$$

But

$$\sin\left(2\pi + \varepsilon\frac{9}{B^k}\right) = \varepsilon \sin\left(\frac{9}{B^k}\right)$$

for $\varepsilon \in \{-1, 1\}$. Thus we have

$$-\sin\left(\frac{9}{B^k}\right) < \sin(\theta) < \sin\left(\frac{9}{B^k}\right).$$

But

$$\sin\left(\frac{9}{B^k}\right) < \frac{9}{B^k} \leq \frac{1}{B^n}$$

according to the facts established at the beginning of this proof and consequently

$$-\frac{1}{B^n} < \sin(\theta) < \frac{1}{B^n}$$

and finally 0 satisfies the bounds property of $\sin(\theta)$ for order n.

□

4.3.7 Other elementary functions

We deduce the other usual elementary functions from the preceding algorithms using the following formulas:

$$\sinh(x) = \frac{\exp(x) - \exp(-x)}{2}, \cosh(x) = \frac{\exp(x) + \exp(-x)}{2}, \tanh(x) = \frac{\sinh(x)}{\cosh(x)},$$
$$x^y = \exp\left(\frac{y \log_B(x)}{\log_B(\exp(1))}\right), \log_x y = \frac{\log_B y}{\log_B x},$$
$$\operatorname{arcsinh}(x) = \log(x + \sqrt{x^2 + 1}), \operatorname{arccosh}(x) = \log(x + \sqrt{x^2 - 1}),$$
$$\operatorname{arctanh}(x) = \tfrac{1}{2}\log\left(\tfrac{1+x}{1-x}\right),$$
$$\arcsin(x) = \arctan\left(\frac{x}{\sqrt{1 - x^2}}\right), \arccos(x) = \arctan\left(\frac{\sqrt{1 - x^2}}{x}\right),$$
$$\cos(x) = \sin(\frac{\pi}{2} - x), \tan(x) = \frac{\sin(x)}{\cos(x)}.$$

4.4 Comparison algorithms for real numbers

We use the expression *absolute comparison* for comparison that may loop for equal numbers but returns always exact results and *relative comparison* for comparison that never loops but returns only results within a given precision.

4.4.1 Absolute comparison between two real numbers

Let x and y be two real numbers represented respectively by the sequences $(x_n)_{n\in\mathbb{Z}}$ and $(y_n)_{n\in\mathbb{Z}}$, then the result $\mathrm{cmp}(x,y)$ of the comparison between x and y is determined as follows:

$n = 0$
While $x_n + 1 > y_n - 1$ and $y_n + 1 > x_n - 1$ do $n \leftarrow n + 1$
If $x_n + 1 \leq y_n - 1$ then $\mathrm{cmp}(x,y) = -1$ else $\mathrm{cmp}(x,y) = 1$

Theorem 21

1.. *The algorithm terminates if and only if* $x \neq y$

2.. $x < y$ *if and only if* $\mathrm{cmp}(x,y) = -1$

3.. $x > y$ *if and only if* $\mathrm{cmp}(x,y) = 1$

REMARK.
This theorem supposes that the computation of x_n and y_n terminates. $\diamond$

Proof.
If $x_n+1 \leq y_n-1$, then $\mathrm{cmp}(x,y) = -1$ and $x < (x_n+1)\,B^{-n} \leq (y_n-1)\,B^{-n} < y$ and by symmetry if $y_n+1 \leq x_n-1$, then $\mathrm{cmp}(x,y) = 1$ et $y < x$. In both cases the algorithm terminates.

We suppose now that x and y are distinct real numbers, we can by symmetry of the algorithm suppose that $x < y$. Then these exists an integer n such that $y - 2/B^n \geq x + 2/B^n$ and consequently

$$\frac{y_n - 1}{B^n} > y - \frac{2}{B^n} \geq x + \frac{2}{B^n} > \frac{x_n + 1}{B^n},$$

thus $y_n - 1 > x_n + 1$, the algorithm terminates and $\mathrm{cmp}(x,y) = -1$. $\square$

4.4.2 Relative comparison between two real numbers

Let x and y be two real numbers represented respectively by the sequences $(x_n)_{n\in\mathbb{Z}}$ and $(y_n)_{n\in\mathbb{Z}}$, let k be an integer, then the result $\mathrm{cmp}_\varepsilon(x,y,k)$ of the comparison between x and y within a precision of B^{-k} is determined as follows

$n = 0$
While $x_n + 1 > y_n - 1$ and $y_n + 1 > x_n - 1$ and $n \leq k + 2$ do $n \leftarrow n + 1$
If $x_n + 1 \leq y_n - 1$ then $\mathrm{cmp}_\varepsilon(x,y,k) = -1$
If $x_n - 1 \geq y_n + 1$ then $\mathrm{cmp}_\varepsilon(x,y,k) = 1$
else $\mathrm{cmp}_\varepsilon(x,y,k) = 0$

Theorem 22

1.. *The algorithm always terminates.*

2.. *We have* $\mathrm{cmp}_\varepsilon(x,y,k) = -1$ *only if* $x < y$.

3.. *We have* $\mathrm{cmp}_\varepsilon(x,y,k) = 1$ *only if* $x > y$.

4.. *We have* $\mathrm{cmp}_\varepsilon(x,y,k) = 0$ *if and only if* $|x - y| < \frac{1}{B^n}$.

REMARK.
This theorem supposes, as the previous one, that the computation of x_n and y_n terminates. $\diamond$

Proof.

This algorithm terminates necessarily since we stop after at most k iterations and k is a finite number.

If $\mathrm{cmp}_\varepsilon(x,y,k) = -1$, then $x_n + 1 \leq y_n - 1$ and $x < (x_n+1)\,B^{-n} \leq (y_n-1)\,B^{-n} < y$ and by symmetry if $\mathrm{cmp}_\varepsilon(x,y,k) = 1$ then $y_n + 1 \leq x_n - 1$ and $y < x$.

If $\mathrm{cmp}_\varepsilon(x,y,k) = 0$, then $x_{k+2} + 1 > y_{k+2} - 1$ and $y_{k+2} + 1 > x_{k+2} - 1$, thus

$$x < \frac{x_{k+2}+1}{B^{k+2}} < \frac{y_{k+2}+3}{B^{k+2}} < y + \frac{4}{B^{k+2}}$$

and

$$x > \frac{x_{k+2}-1}{B^{k+2}} > \frac{y_{k+2}-3}{B^{k+2}} > y - \frac{4}{B^{k+2}}.$$

Consequently

$$y - \frac{4}{B^{k+2}} < x < y + \frac{4}{B^{k+2}}$$

and

$$|x-y| < \frac{4}{B^{k+2}}.$$

But $B \geq 2$, thus $\frac{4}{B^2} \leq 1$ and $|x-y| < \frac{1}{B^n}$. □

4.5 Existence of the $\underline{f}$ functions for all the f functions mentioned above

We assume at the beginning of the previous section that for any "basic" transcendental function f there exists a function $\underline{f} : \mathbb{Q} \to \mathcal{R}$, that maps any rational number r to a sequence $\underline{f}(r,n)/B^n$ ($\underline{f}(r,n) \in \mathbb{Z}$) such that

$$\frac{\underline{f}(r,n) - 1}{B^n} < f(r) < \frac{\underline{f}(r,n) + 1}{B^n}.$$

k-root is a special case since the function $x \mapsto \sqrt[k]{x}$ is defined directly on $\mathbb{N}$ and maps any integer x to $\lfloor \sqrt[k]{x} \rfloor$. Such a function is computed by Newton's method applied to the function $z \mapsto z^n - x$ (see [37] for more details).

4.5.1 Basic case for transcendental functions

The main argument of our proof for the existence of such a function $\underline{f}$ is that this function is defined by an alternate series converging on a non empty interval

$$f(r) = s = \sum_{i \in \mathbb{N}} (-1)^i a_i$$

where $(a_i)_{i\in\mathbb{N}}$ (If the general term of the Taylor expansion of f is b_n, $a_n = b_n r^n$) is a sequence of rational numbers with same common sign.

This approach has two advantages: first of all, we have a very simple stop condition for the series converging according to the alternate series criterion since it is sufficient that $|a_{n+1}|$ is lesser than the required error ε to ensure that the sum of this series up to rank n

$$s_k = \sum_{i=0}^{i=k} (-1)^i a_i$$

is an approximation of the limit s within ε. Furthermore, two consecutive terms of such a series supply an interval with rational bounds containing the limit: if all terms of the sequence $(a_i)_{i\in\mathbb{N}}$ are positive, then $s_{2i+1} < s < s_{2i}$ for any $i \in \mathbb{N}$ and $s_{2i} < s < s_{2i-1}$ for any $i \in \mathbb{N}^*$, if all terms of the sequence $(a_i)_{i\in\mathbb{N}}$ are negative.

So we have a recursively enumerable sequence of nested intervals with rational bounds including the real number to compute and with length vanishing to 0, that is to say that this real number is computable according to the first definition of the notion of computable real number and the equivalence of this notion with the notion of B-approximable real number supply us an algorithm to compute an integer λ (and a value for k to compute λ) such that λ satisfies the bound property of s for order n. Essentially we use the fact that $s_k = p_k/q_k$ and $q_k \geq B^n$ so $\lambda = \lfloor s_k B^n \rfloor$.

Consequently, we will come to this simple case for each basic function and prove that this transformation preserves the bounds property.

4.5.2 Exponential function

Let r be a rational number, we represent $\exp(r)$ by the product of $e^{\lfloor r \rfloor + 1}$ by $\exp(r - (\lfloor r \rfloor + 1))$ if r is not null and 1 otherwise. The first term of this product is computed as a power of e, that is to say by successive multiplications (and inversion if r is negative). The computation of the second term uses the Taylor expansion of $\exp(x)$ for $-1 \leq x < 0$. The basic case can be directly applied to this term. The computation of e may be performed either by using directly the series of general term $1/n!$ and the inequality

$$\sum_{k \geq n+1} 1/k! < 1/(n\, n!)$$

for the stop condition and an interval with rational bounds including e: $s_n < e < s_n + 1/(n\, n!)$ with the same usage as above, or by inversion of $\exp(-1)$, computed according to the second case. Whatever the choice, we use the multiplication of two real numbers and possibly the inversion of a real number, and we obtain function $\underline{\exp}$.

4.5.3 Logarithm function

Let r be a rational number, we compute $\ln(r)$ as follows: if $r \leq 0$, then the computation fails; if $r < 1$, then we take $\ln(r) = -\ln(1/r)$; if $r = 1$, then the result is the null sequence; if $r > 1$, then we use the formula

$$\ln(r) = 2 \operatorname{arctanh}\left(\frac{r-1}{r+1}\right)$$

with $y = (r-1)/(r+1)$, we have

$$\operatorname{arctanh}(y) = \sum_{k \geq 0} \frac{y^{2k+1}}{2k+1}$$

and

$$\sum_{k \geq n+1} \frac{y^{2k+1}}{2k+1} \leq \frac{y^{2n+3}}{(2n+3)\,(1-y^2)}.$$

Thus we have an interval with rational bounds $s_n < \ln(r) < s_n + y^{2n+3}/((2n+3)\,(1-y^2))$ and we use it as for above.

4.5.4 Arctangent function

The Taylor expansion of $\arctan(r)$ is an alternate series for any rational r.

4.5.5 Sine function

The Taylor expansion of $\cos(x)$ is an alternate series for any rational x. The Taylor expansion of $\sin(x)$ is an alternate series for any positive rational x and the sine function is odd so we can always use come to the alternate series.

4.5.6 Remark

This is a proof of the existence of at least one set of functions $\underline{f}$ but it is only one possible way to compute one such set. For example, we can also use ideas similar to those described by Brent in [10] to compute these functions.

Chapter 5

Implementation

5.1 The choice of the Caml language

The use of this language was of course a natural choice insofar as we began the study on the subject of arithmetic for a modern and reliable programming language about this language. Furthermore the first step of this study leads us to implement a very efficient exact rational arithmetic for this language, that relies on the Bignum package [30, 46].

It is obvious that (almost) infinite integers are absolutely necessary, but an exact rational arithmetic (see [36]) is also necessary to compute the transcendental functions on rational parameters underlying the transcendental functions on real arguments.

Moreover functions in this language are easy to use as arguments or results of functions and since real numbers (and more generally infinite objects) are naturally represented by functions, it is easier to deal with real numbers in this language.

5.2 Choices of implementation

We choose as Boehm to represent real numbers as finite B-adic numbers and furthermore these particular finite B-adic numbers, instead as general rational numbers. This choice leads us to a rougher granularity and a slightly lesser flexibility for our representation. For instance, if an accuracy under $1/B^n$ is required, this choice of implementation leads to a computation with an accuracy of $1/B^{n+1}$ and induce a greater running time than a computation to the real precision of $1/B^n - \varepsilon$ where ε is a rational number as small as possible.

Boehm's implementation used rational numbers at the beginning and it turned out that with the library of rational arithmetic used by Boehm, the computations with rational numbers were much slower than those performed with finite B-adic numbers, so he finally choose B-adic numbers to represent real numbers. But this choice deprive us of the natural incrementality of the representation and of a slightly simpler expression of our algorithms. Indeed this representation, if we don't use the mpa functionality, is not incremental, that is to say that if we have computed a result to n digits, if we want to compute its value to $n+1$ digits we need to compute it again from scratch. But we adapt the representation to lessen this drawback by the following choices.

The implementation includes the storage of the most precise approximation already computed for each real number and we choose to work with the base $B = 4$. We will now justify our choice.

For efficiency, the representation includes for each real number x represented by the sequence $(x_n)_{n\in\mathbb{Z}}$, not only the functional closure but also the most precise approximation already computed $x_{\mathrm{mpa}(x)}$ to the order $\mathrm{mpa}(x)$ as mentioned above in page 23. This choice leads us to redo only partly the computation: for example with a function that consider the size of its argument(s) with the msd function, these arguments have been computed to a precision sufficient to compute their msd and the computation of msd is reduced after this first computation to some shifts operations and maybe the already computed approximations for part of the arguments will be sufficient.

Practically a good sharing of the expressions will increase the effect of this information storage and reduce the time of computation.

Concerning the choice of the base, it is preferable to choose 2 to some power, since

$$\left\lfloor \frac{x_{\mathrm{mpa}(x)}}{B^{\mathrm{mpa}(x)-n}} \right\rfloor$$

can be computed by a simple computer shift of $x_{\mathrm{mpa}(x)}$ of $\mathrm{mpa}(x) - n$ digits to the right in this case, that is to say a basic operation of rational arithmetic and then for the underlying hardware arithmetic. It may seems worthwhile that a digit for the base B corresponds exactly to a computer word. However we have also to consider that the smaller the base is and the less we pay to compute an additional digit. Hans Boehm choose to work with $B = 4$ and me too. This is a good compromise between on the one hand the trend to perform computations on computer words and on the other hand the fact that if we need to compute one digit more for a number the cost should not be very different.

5.3 Realisations

Boehm implemented a similar arithmetic, so one can read his commentaries in [8, 7]. Moreover we have currently a complete prototype for this representation. Tests are in progress. The chosen representation has the advantage of using algorithms on integers that are well understood and very efficient.

Furthermore in 2001, Jean-Christophe Filliâtre released an independant implementation from my PhD thesis [37] available at `http://www.lri.fr/~filliatr/software.en.html`.

Finally, XR by Keith Briggs and Yannis Smaragdakis [11, 12] in python and C++ seems to be inspired of this work with $B = 16$ to extend Victor Shoup's arbitrary-precision arithmetic package NTL.

Chapter 6

Conclusion

We have a description of a representation of $\mathcal{R}$ and proved algorithms for this representation for all elementary functions.

We have implemented a complete prototype of this description so we have a complete chain for reliable arithmetic in the Caml language.

In the future, it may be also interesting to study the influence of the radix B on the efficiency of real computations. Furthermore we hope to improve the efficiency of this prototype by an optimized computation of the $\underline{f}$ functions with optimizations like those developed in [3] and by a balancement of the abstract syntax tree during the compilation of expressions such as $x_1 + \ldots + x_n$ to compute each x_i with a well-balanced precision.

It seems to be interesting to combine our arithmetic with floating point analysis method such as interval analysis [41, 42, 1, 18] or the CESTAC method [50, 13, 49] in the spirit of the lazy rational arithmetic by Michelucci [4]: it consists in computing the functional closure of the result and at the same time to compute the interval result according to interval analysis. If a result (maybe an intermediary result) is not precise enough we compute it with exact real arithmetic at the needed precision. This method has the advantage that it is efficient and precise: generally speaking, big floating point applications accept to pay in time only when necessary so this solution is well adapted to this need. However this approach is limited by the fact that IEEE floating point standard arithmetic concerns only rational operations currently.

Our goal is not to substitute our arithmetic to floating-point arithmetic, that is often sufficient and very efficient, but to make available an alternative arithmetic for specific needs. We want to be able to compute a reliable result even if it takes a while rather than to obtain a wrong result immediately.

An idea consists to consider this arithmetic as a static analysis of the needed precision for a floating point computation according to the required precision on the result.

Finally, it would be interesting to build real analysis on top of this real arithmetic.

Chapter 7

Current state of the art

We present here briefly the later works with or without proofs.

7.1 Continued fractions, Möbius transformations, LFT

David Lester in [31] describes a type of continued fractions for which Gosper'algorithms are correct.

Peter Potts and Abbas Edalat in [19, 43, 44] represent real numbers as Linear Fractional Transformations (LFT) and show how to encode continued fractions using LFT and deduce algorithms to compute with LFT. Reinhold Heckmann in [27, 26, 25] shows how to manage computations with LFT according to the expected precision.

7.2 Computable Cauchy sequences

David Lester and Paul Gowland in [23] presents an arithmetic using effective Cauchy sequences (sequences of finite 2-adic numbers) with algorithms similar to ours for rational operations (including iterators), square root and simplistic transcendental functions using power series.

7.3 Adaptive computations

This approach consists in an iterative bottom-up analysis. The computation starts with a predefined precision on all inputs and at each step of the computation, if the required precision is not obtained, the computation is performed with increased precision.

MPFR (Polka team at INRIA Loria, directed by Paul Zimmermann [58]) computes with floating-point representations.

The iRRAM work of Norbert Müller [40] relies on the REAL RAM by Vasco Brattka and Peter Bretling [9].

7.4 Implementations

Jean Vuillemin have carried out a small implementation of continued fractions in Lisp, but it never was available in no way at all.

Hans Boehm have implemented a pocket calculator and a version in Java is currently available at `http://www.hpl.hp.com/personal/Hans_Boehm/crcalc/CRCalc.html`.

We mentioned in subsection 5.3 three implementations of our work.

Peter Potts have made a small prototype for LFT in Caml named Calathea, available at `http://www.purplefinder.com/~potts/calathea.zip`. A complete prototype named IC-reals in C is available at `http://www.doc.ic.ac.uk/~ae/ic-reals-6.2-beta.tar.gz`.

David Lester in [31] mentions an Haskell very slow implementation using continued fractions and a more classical computable Cauchy sequences representation used in the implementation MAP presented in [23, 6], available at `http://www.cs.man.ac.uk/arch/dlester/exact.html` (Haskell version, C version announced).

Norbert Müller has written the C++ very efficient package iRRAM available at `http://www.informatik.uni-trier.de/iRRAM/`.

Paul Zimmermann and al. make MPFR [34] available at `http://www.loria.fr/projets/mpfr/`.

Paul Gowland and David Lester have surveyed exact real arithmetic implementations in [24] and Jens Blanck have compared them in [6].

7.5 Mechanically checked proofs

In [32], David Lester and Paul Gowland have proved in PVS the correctness of their algorithms on computable Cauchy sequences described in [23], relying on the NASA Langley PVS real library for axiomatic definitions of the transcendental functions. The complete proof is available at `http://www.cs.man.ac.uk/arch/dlester/exact.html`.

David Lester in [31] mentions machine-assisted proofs for the central algorithms for rational operations on continued fractions.

Jérôme Créci in [16] has defined our representation and proved our addition, subtraction and multiplication algorithms in Coq, relying on the Reals library axiomatized in the Coq system. When all our algorithms will be proved in Coq, we will be able to combine this work with the real analysis available in the axiomatization of real numbers.

Paul Zimmermann proved some algorithms of MPFR in Coq with the help of Lemme team of Inria Sophia-Antipolis.

Acknowledgements

This work took place essentially at INRIA Rocquencourt during the PhD of the author and minorly at Université d'Évry-Val d'Essonne.

The author thanks the Theory and Formal Methods at Imperial College for his warm hospitality during the major step of the writing of this article, supported at Imperial College by EPSRC project "Techniques for real number computation".

Jean-Michel Muller advises me all along this publication work. Moreover his relevant questions at the time of my PhD defense made me aware of the additional work to do for transcendal functions on rational numbers.

The formal proof in Coq of Jérôme Créci reveals a small error in multiplication and now this algorithm is correct and formaly proved. I hope that this work of formalisation and automated proof will be achieved.

Jean-Christophe Filliâtre suggests an improvement for the computation of the sin function, make its CREAL implementation of my work available and trained Jérôme Créci during his master thesis's work.

Bibliography

[1] O. Aberth. *Precise numerical analysis.* Wm. C. Brown Publishers, 1988.

[2] A. Avizienis. Signed-digit number representations for fast parallel arithmetic. *IRE Transactions on electronic computers*, 10 (1961), 389–400.

[3] C. Batut. *Aspects algorithmiques du système de calcul arithmétique en multiprécision PARI.* Thèse de doctorat, Université de Bordeaux I, Feb. 1989.

[4] M. Benouamer, P. Jaillon, D. Michelucci, and J.-M. Moreau. A Lazy Solution to Imprecision in Computational Geometry. In *Proceedings of the 5th Canadian Conference on Computational Geometry* (Aug. 1993), pp. 73–78. Available from World Wide Web: `ftp://ftp.emse.fr/pub/papers/LAZY/LazyCG.ps.gz`.

[5] E. Bishop and D. Bridges. *Constructive Analysis*, vol. 279 of *Grundlehren der mathematischen Wissenschaften, A series of Comprehensive Studies in Mathematics.* Springer Verlag, 1985.

[6] J. Blanck. Exact real arithmetic systems: Results of competition. In *Computability and Complexity in Analysis* (2001), vol. 2064 of *Lecture Notes in Computer Science*, Springer, pp. 389–393. 4th International Workshop, CCA 2000, September 2000.

[7] H. J. Boehm. Constructive Real Interpretation of numerical Programs. In *Proceedings of the 1987 ACM conference on Interpreters and Interpretives Techniques* (1987), ACM.

[8] H. J. Boehm, R. Cartwright, M. J. O'Donnel, and M. Riggle. Exact real arithmetic: a case study in higher order programming. In *Proceedings of the 1986 ACM conference on Lisp and Functional Programming* (1986), ACM.

[9] V. Brattka and P. Hertling. Feasible real random access machines. *Journal of Complexity 14*, 4 (1998), 490–526.

[10] R. P. Brent. Fast Multiple-Precision Evaluation of Elementary Functions. *Journal of the ACM 23*, 2 (Apr. 1976), 243–251.

[11] K. Briggs. Exact real computation. Talk at University of Warwick, May 2001. Available from World Wide Web: `http://www.btexact.com/people/briggsk2/xr-talk.ps`.

[12] K. Briggs and Y. Smaragdakis. XR — exact real arithmetic. World-Wide Web document and software package, Mar. 2001. Available from World Wide Web: `http://www.btexact.com/people/briggsk2/XR.html`.

[13] J.-M. Chesneaux. Study of the computing accuracy by using probabilistic approach. In *Contribution to Computer arithmetic and Self-Validating Numerical Methods* (Oct. 1990), C. Ullrich, Ed., Academic Press, pp. 19–30.

[14] J.-M. Chesneaux. L'approche probabiliste des erreurs d'arrondi. Talk to a workshop on the quality of computers results organized by Paris 6 University, Apr. 1997.

[15] R. Cori and D. Lascar. *Logique mathématique*, vol. 2. Masson, 1993.

[16] J. Créci. Certification d'algorithmes d'arithmétique rélle exacte dans le système Coq. DEA Logique et Fondements de l'Informatique, Université Paris 11, 2002.

[17] N. Cutland. *Computability: an introduction to recursive function theory.* Cambridge University Press, 1980.

[18] M. Daumas, C. Mazenc, and J.-M. Muller. User transparent interval arithmetic. In *Proceedings of IMACS/GAMM International Symposium SCAN-93* (1993). Available from World Wide Web: `ftp://ftp.lip.ens-Lyon.fr/pub/Rapports/RR/RR94/RR94-02.ps.Z`. Also available as research report number 94-02 of École Normale Supérieure de Lyon, January 1994.

[19] A. Edalat and P. J. Potts. A new representation for exact real numbers. *Electronic notes in Theoretical Computer Science 6*, 14 (1997).

[20] J.-C. Filliâtre. Description of the Creal module for Objective Caml. World-Wide Web document, Nov. 2001. Available from World Wide Web: `http://www.lri.fr/~filliatr/ftp/ocaml/ds/creal.ps.gz`.

[21] D. Goldberg. What every computer scientist should know about floating point arithmetic. *ACM Computing Surveys 23*, 1 (Mar. 1991), 5–47. Available from World Wide Web: `http://docs.sun.com/htmlcoll/coll.648.2/iso-8859-1/NUMCOMPGD/ncg_goldbe%rg.html`.

[22] B. Gosper. Continued Fraction Arithmetic. HAKMEM Item 101B, MIT AI MEMO 239, Feb. 1972. Available from World Wide Web: `ftp://ftp.netcom.com/pub/hb/hbaker/hakmem/cf.html\#item101b`.

[23] P. Gowland and D. Lester. The correctness of an implementation of exact arithmetic. In *Proceedings of the fourth conference on real numbers and computers* (2000).

[24] P. Gowland and D. Lester. A survey of exact arithmetic implementations. In *Computability and Complexity in Analysis* (2001), vol. 2064 of *Lecture Notes in Computer Science*, pp. 30–47. 4th International Workshop, CCA 2000, September 2000.

[25] R. Heckmann. Translation of Taylor series into LFT expansions. submitted to Proceedings of Dagstuhl Seminar "Symbolic Algebraic Methods and Verification Methods", Nov. 1999.

[26] R. Heckmann. How many argument digits are needed to produce n result digits. *Electronic Notes in Theoretical Computer Science 24* (2000). RealComp '98 Real Number Computation, Indianapolis (June 1998).

[27] R. Heckmann. Contractivity of linear fractional transformations. *Theoretical Computer Science 279*, 1 (2002), 65 – 82. Third Real Numbers and Computers Conference (1998).

[28] F. C. Hennie. *Introduction to computability.* Addison-Wesley series in computer science and information processing. Addison-Wesley, 1977.

[29] H. Hermes. *Enumerability Decidability Computability. An introduction to the theory of recursive functions.* Springer, 1965.

[30] J.-C. Hervé, F. Morain, D. Salesin, B. Serpette, J. Vuillemin, and P. Zimmermann. BigNum: A Portable and Efficient Package for Arbitrary-Precision Arithmetic. Tech. Rep. 1016, INRIA, Domaine de Voluceau, 78153 Rocquencourt, FRANCE, Apr. 1989.

[31] D. Lester. Effective continued fractions. In *Proceedings of the 15th IEEE Symposium on Computer Arithmetic* (2001), pp. 163–170.

[32] D. Lester and P. Gowland. Using PVS to validate the algorithms of an exact arithmetic. *Theoretical Computer Science 291*, 2 (Nov. 2002), 203–218. Available from World Wide Web: `http://www.cs.man.ac.uk/arch/dlester/exact.html`.

[33] H. X. Lin and H. J. Sips. On-Line Cordic Algorithms. IEEE Transactions on Computers *39*, 8 (Aug. 1990), 1038–1052.

[34] M. T. LORIA / INRIA Lorraine. *MPFR: Multiple Precision Floating-Point Reliable Library*, 2.0.1 ed., Apr. 2002. Available from World Wide Web: `http://www.loria.fr/projets/mpfr/mpfr-current/documentation.html`.

[35] P. Martin-Löf. *Notes on Constructive Mathematics.* Almqvist et Wiksell, Stockholm, 1970.

[36] V. Ménissier-Morain. The CAML Numbers Reference Manual. Tech. Rep. 141, INRIA, July 1992. Available from World Wide Web: `http://calfor.lip6.fr/~vmm/documents/doc_arith_without_camels.ps.gz`.

[37] V. Ménissier-Morain. *Arithmétique exacte, conception, algorithmique et performances d'une implémentation informatique en précision arbitraire.* Thèse, Université Paris 7, Dec. 1994. Available from World Wide Web: `http://calfor.lip6.fr/~vmm/documents/these94.ps.gz`.

[38] J.-M. Muller. *Arithmétique des ordinateurs.* Etudes et recherches en informatique. Masson, 1989.

[39] J.-M. Muller. Ordinateurs en quête d'arithmétique. *La Recherche 26*, 278 (Juillet-Août 1995), 772–777.

[40] N. T. Müller. The iRRAM: Exact arithmetic in C++. In *Computability and Complexity in Analysis* (2001), vol. 2064 of *Lecture Notes in Computer Science*, pp. 222–252. Available from World Wide Web: `http://www.informatik.uni-trier.de/~mueller/`. Proceedings of the 4th International Workshop, CCA 2000, September 2000.

[41] K. Nickel, Ed. *Interval mathematics* (May 1975). LNCS 29.

[42] K. Nickel, Ed. *Proceedings of the international symposium on interval mathematics* (Sept. 1985), Springer-Verlag. LNCS 212.

[43] P. J. Potts. *Exact real arithmetic using Möbius transformations.* PhD thesis, Imperial College, 1999.

[44] P. J. Potts, A. Edalat, and P. Sünderhauf. Lazy computation with exact real numbers. In *Proceedings of the Third ACM SIGPLAN International Conference on Functional Programming* (1998), pp. 185–194.

[45] H. G. Rice. Recursive real numbers. *Proceedings of the American Mathematical Society 5*, 5 (1954).

[46] B. Serpette, J. Vuillemin, and J. Hervé. BigNum: A Portable and Efficient Package for Arbitrary-Precision Arithmetic. Tech. Rep. 2, Digital PRL, May 1989.

[47] G. Stolzenberg. A new Course in Analysis. Note of a course given between 1972 and 1987 at Northeastern University, Boston MA 02115 USA, 1972-1987.

[48] A. M. Turing. On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society 2*, 42 (1936).

[49] J. Vignes. A stochastic arithmetic for reliable scientific computation. *Mathematics and computers in simulation 35* (1993), 233–261.

[50] J. Vignes and M. La Porte. Error analysis in computing. In *Information Processing 74* (Aug. 1974), North-Holland.

[51] J. Vuillemin. Exact real computer arithmetic with continued fractions. Research report 760, INRIA, 1987.

[52] J. Vuillemin. Exact real computer arithmetic with continued fractions. In *Proceedings ACM conference on Lisp and Functional Programming* (1988), ACM. Extended version as INRIA research report 760, 1987.

[53] J. Vuillemin. Exact real computer arithmetic with continued fractions. *IEEE Transactions on computers 39*, 8 (Aug. 1990), 1087–1105.

[54] K. Weihrauch. *Computability.* EATCS monographs on theoretical computer science. Springer Verlag, 1987.

[55] E. Wiedmer. *Exaktes Rechnen mit reellen Zahlen und anderen unendlichen Objekten.* PhD thesis, ETH, Zurich, 1977. Diss. ETH 5975.

[56] E. Wiedmer. Calculs avec des fractions décimales et d'autres objets infinis. Compte-rendu d'un exposé à l'Institut de Programmation de l'Université Paris VII, Jan. 1979.

[57] E. Wiedmer. Computing with infinite objects. *Theoretical Computer Science 10* (1980).

[58] P. Zimmermann. MPFR: a library for multiprecision floating-point arithmetic with exact rounding. In *Proceedings of the fourth conference on real numbers and computers* (2000), pp. 8–90.