A Scientific Calculator for Exact Real Number Computation Based on LRT, GMP and FC++.

Alejandra Lucatero*, J. Raymundo Marcial-Romero* and J. A. Hernández*

ABSTRACT

Language for Redundant Test (LRT) is a programming language for exact real number computation. Its lazy evaluation mechanism (also called call-by-need) and its infinite list requirement, make the language appropriate to be implemented in a functional programming language such as Haskell. However, a direct translation of the operational semantics of LRT into Haskell as well as the algorithms to implement basic operations (addition subtraction, multiplication, division) and trigonometric functions (sin, cosine, tangent, etc.) makes the resulting scientific calculator time consuming and so inefficient. In this paper, we present an alternative implementation of the scientific calculator using FC++ and GMP. FC++ is a functional C++ library while GMP is a GNU multiple precision library. We show that a direct translation of LRT in FC++ results in a faster scientific calculator than the one presented in Haskell.

INTRODUCTION

During the last decades there have been several paradigms to compute with the real numbers. Among them we can cite floating point number arithmetic [16], interval analysis [15], algebraic manipulation [20], exact real number computation [21], etc.

The exact real number computation paradigm has several advantages compared to the others, for example, it avoids the rounding off errors that occurs in floating point arithmetic. Like interval analysis, it bounds the result, however it guarantees the computation of a smaller interval at each step of the computation which do not occur in interval analysis. Algebraic manipulations can be used in exact real number computation however, when no further reduction can be done a exact method is used to compute the solution compared to algebraic manipulation which has to turn to floating point arithmetic or interval analysis.

There have been several theoretical proposals to exact real number computations [19, 5, 10]. Most of them have succeeded to prove that the theory is sound and complete. However, when they have been implemented, none of them has achieved to be efficient and straightforward to translate from the theory to the practice. On the other hand, implementations such as IRRAM [16], MPFR [6] and RealLib [8] have been developed in C and C++, however, in order to run faster, they have lost the elegancy of functional programming and also.

* Computer Engineering Department, Universidad Autónoma del Estado de México, Cerro de Coatepec s/n, Ciudad Universitaria, C.P. 50100, Toluca, Estado de México, México. E-mail: ritsuka00@gmail.com, jmarcial@uaemex.mx, xoseahernandez@gmail.com

Palabras clave: Cálculo con números reales; programación funcional; semántica operacional

Keywords: Real number computation; functional programming; operational semantics

Redibido: 20 de Diciembre de 2011
Aceptado: 14 de Febrero de 2012
they have slightly deviated from the theory. Although we consider that faster implementations is what is required in practice, we believe that there is increased confidence in the correctness of an implementation the closer it is to the original theory.

A pair of implementations which have a close harmony between theory and practice are Era [2] and a corecursive sign digit representation [3], implemented in Objective Caml and Coq respectively. However, Era, as said by the authors, is slower than IRRAM since C++ generally compiles to more efficient code than Objective Caml. Respect to the corecursive implementation, although an excellent theory is presented, the efficiency is never mentioned.

A further well established theory for exact real number computation is LRT (Language for Redundant Test) [10]. It has been proved that any computable first order function can be defined in LRT [13]. Moreover, an implementation of LRT in Haskell has been presented on Marcial et. al [11]. However, its efficiency compared to a sign digit representation [18] was very poor. In Marcial et. al [12] an implementation of a basic calculator based on LRT and programmed using FC++ and GMP was presented. It was shown that the execution time improved compared to other Haskell implementations.

In this paper, we present the results of an extension of the basic calculator to a scientific calculator, showing that the time also improved in trigonometric functions as well as other operations. FC++ is an extension of C++ which allows to translate functional programs in almost an straightforward way, meaning that the effects of functional programming are there. Since the types of FC++ are the standard of C++ (which implies that numbers are truncated or rounded), we use instead the types of GMP. GMP is a well known library which computes arithmetic operations faster than many other implementations. Additionally, GMP allows to compute to any required precision.

The paper is divided as follows, in Section The LRT Language, the language LRT is defined. In Section The Scientific Calculator a brief explanation of the trigonometric implementations is presented. Finally the conclusions and further work is discussed.

**THE LRT LANGUAGE**

We introduce the LRT language, which is a variant of Real PCF [5]. This is a call-by-name language.

**Syntax**

The language LRT is an extension of PCF (Programming Computable Functions) [17] with a ground type for real numbers and suitable primitive functions for real-number computation. Its raw syntax is given by

\[
x \in \text{Variable},
\]

\[
t ::= \text{nat} \mid \text{bool} \mid \mathbb{I} \mid t \to t,
\]

\[
P ::= x \mid n \mid \text{true} \mid \text{false} \mid (+1)(P) \mid (-1)(P) \mid
\]

\[
(= \theta)(P) \mid \text{if } P \text{ then } P \text{ else } P \mid \text{cons}_{\mathbb{I}}(P) \mid
\]

\[
\text{tail}_{\mathbb{I}}(P) \mid \text{rtest}_{\mathbb{I}}(P) \mid \lambda x : t.P \mid PP \mid YP,
\]

where Variable is a set of variables, \( t \) represents a set of types, in this case the language has three ground types, the natural numbers type (represented by nat), the booleans (represented by bool) and the unit real number type (represented by I which denotes the set of intervals in \([-1,1]\), as it was shown in [9] the complete computable real line can be easily represented in this language, even more the implementation presented here considers the complete real line). The type \( t \to t \) denotes higher order types. The constructs of the language (represented by \( P \)) are the variables (represented by \( x \)), the constants for natural numbers and booleans (represented by \( n, \text{true} \) and \( \text{false} \)), the successor, predecessor and equal test for zero operations for naturals numbers ((+1), (-1) and (=0)), the classical if operator; three operation for exact real number computation (cons, tail and rtest) where the subscripts of the constructs cons and tail are rational intervals (sometime written as \( a \) or \( [a, \bar{a}] \)) and those of rtest are rational numbers. The last three constructors of the languages are those of the lambda calculus (\( \lambda x : t.P \), PP and YP) where the first denotes abstraction, the second application and the third recursion.

Because the intention of this paper is not to present the denotational semantics of the language which is based on powedrdomains [10], we just present the mathematical objects which describe the cons, tail and rtest constructors. The others are the well known PCF constructors and can be consulted at [7, 17].

Let \( D \subseteq [-1,1] \), the function \( \text{cons}_a : D \to D \) is the unique increasing affine map with image the interval \( a \), i.e.,

\[
\text{cons}_a([\bar{a}, \bar{a}]) = \left[ \frac{a - a}{2} \bar{a} + \frac{a + a}{2} , \frac{a - a}{2} + \frac{a + a}{2} \right]
\]

That is, rescale and translate the interval \([-1,1]\) so that it becomes \([a, \bar{a}]\), and define \( \text{cons}_a([\bar{a}, \bar{a}]) \) to be the interval which results from applying the same rescaling and translation to \([\bar{a}, \bar{a}]\). In order to keep the notation simple, when the context permits we use \( x \) to
represent \([5,7]\), meaning that the same operation is applied to both end points of the interval obtained, for example the \(\text{cons}\) function can be written as:

\[
\text{cons}_{\bar{a}}(x) = \frac{\bar{a} - a}{2} x + \frac{\bar{a} + a}{2} \tag{1}
\]

The function \(\text{tail}_{\bar{a}}(x) : D \rightarrow D\) is a left inverse, i.e.

\[
\text{tail}_{\bar{a}}(\text{cons}_{\bar{a}}(x)) = x.
\]

More precisely, the following left inverse is taken, where \(\kappa_a\) is \(\bar{a} - a\) and \(\tau_a\) is \(\bar{a} + a\):

\[
\text{tail}_{\bar{a}}(x) = \max(-1, \min(2x - \tau_a)/\kappa_a, 1)).
\]

This definition guarantees that the range of the tail function is in the interval \([-1,1]\). The details of why this is a convenient definition can be consulted in [5]. It is worthy to mention that an infinite shrinking sequence of cons intervals represent a real number in the interval \([-1,1]\), the operational semantics defined below gives a rule for constructing a real number.

The definition of the function \(\text{rtest}_{\bar{a}} : D \rightarrow \{\text{true}, \text{false}\}\), where \(l < r\) are rational numbers, can be formulated as

\[
\text{rtest}_{\bar{a}}(x) = \begin{cases} 
\text{true}, & \text{if } x \subseteq (-\infty, l], \\
\text{true or false}, & \text{if } x \subseteq (l, r), \\
\text{false}, & \text{if } x \subseteq [r, \infty).
\end{cases} \tag{2}
\]

The function \(\text{rtest}_{\bar{a}}\) is operationally computable because, for any argument \(x\) given intentionally as a shrinking sequence of \(\text{cons}\) intervals, the computational rules systematically establish one of the semidecidable conditions \(l < \bar{x} < r\) where \(l, r\) are rational numbers.

**Operational Semantics**

We consider a small-step style operational semantics for our language. We define the one-step reduction relation \(\rightarrow\) to be the least relation containing the one-step reduction rules for evaluation of PCF [17] together with those given below.

We first need some preliminaries. For intervals \(a\) and \(b\) in \([-1,1]\), we define

\[
ab = \text{cons}_{\bar{a}}(b),
\]

where \(\text{cons}\) is the function defined previously. This operation is associative, and has the interval \([-1,1]\) (denoted by \(\bot\)) as its neutral element [5]:

\[
(ab)c = a(bc), \quad a\bot = \bot a = a.
\]

In the interval domain literature [1], \(a \subseteq b\) iff \(b \leq a\). Moreover,

\[
a \subseteq b \iff \exists c \in D. \ ac = b,
\]

and this \(c\) is unique if \(a\) has non-zero length. In this case we denote \(c\) by

\[
b \setminus a.
\]

For intervals \(a\) and \(b\), we define

\[
a \leq b \iff \bar{a} \leq \bar{b}
\]

and

\[
a \uparrow b \iff \exists c. \ a \leq c \text{ and } b \leq c.
\]

With this notation, the rules for Real PCF as defined in [5] are:

1. \(\text{cons}_{\bar{a}}(\text{cons}_{\bar{a}}(M)) \rightarrow \text{cons}_{\bar{a}}(\text{cons}_{\bar{a}}(M))\)
2. \(\text{cons}_{\bar{a}}(M) \rightarrow \text{cons}_{\bar{a}}(M')\)
3. \(\text{tail}_{\bar{a}}(\text{cons}_{\bar{a}}(M)) \rightarrow Y\text{cons}_{\bar{a} - 1,0}\) if \(b \leq a\)
4. \(\text{tail}_{\bar{a}}(\text{cons}_{\bar{a}}(M)) \rightarrow Y\text{cons}_{\bar{a} + 1,0}\) if \(a \geq b\)
5. \(\text{tail}_{\bar{a}}(\text{cons}_{\bar{a}}(M)) \rightarrow \text{cons}_{\bar{a} + 1,0}\) if \(a \leq b\) and \(a \neq b\)
6. \(\text{tail}_{\bar{a}}(M) \rightarrow \text{tail}_{\bar{a}}(M')\)
7. \(\text{if } true\ M\ N \rightarrow M\)
8. \(\text{if } false\ M\ N \rightarrow N\)
9. \(\text{if } M\ N_1\ N_2 \rightarrow M'\ N_1\ N_2\)

For our language \(LRT\), we add:

10. \(\text{rtest}_{\bar{a}}(\text{cons}_{\bar{a}}(M)) \rightarrow \text{true}\) if \(\bar{x} < r\)
11. \(\text{rtest}_{\bar{a}}(\text{cons}_{\bar{a}}(M)) \rightarrow \text{false}\) if \(l < \bar{x}\)
12. \(\text{rtest}_{\bar{a}}(M) \rightarrow \text{rtest}_{\bar{a}}(M')\) if \(M \rightarrow M'\).

**Remarks:**

1. Rule (1) plays a crucial role and amounts to the associativity law. The idea is that both \(a\) and \(b\) give partial information about a real number, and \(ab\) is the result of gluing the partial information together in an incremental way. See [5] for a further discussion including a geometrical interpretation.

2. Rules (2), (6), (9) and (12) are applied whenever any of the other rules are matched.

3. Rule (3) represents the fact that we already know that the rest of the real number we are looking for is an infinite sequence of the interval \([-1,0]\), i.e.

\[
Y\text{cons}_{\bar{a} - 1,0} = \text{cons}_{\bar{a} - 1,0}(\text{cons}_{\bar{a} - 1,0}(\ldots)))
\]

4. Rule (4) is similar to rule (3).

5. Rule (5) is applied when the partial information accumulated at some point contains the interval of the next input.
6. Rules (7) and (8) are the classical conditional rules.

7. Notice that if the interval \(a\) is contained in the interval \([l, r]\), rules (10) and (11) can be applied.

8. Rules (10)-(12) cannot be made deterministic given the particular computational adequacy formulation which is proved in [10].

9. In practice, one would like to avoid divergent computations by considering a strategy for application of the rules. In [10] total correctness of basic algorithms and in [14] total correctness of first order functions are shown, hence any implementation of any strategy will be correct.

For a deeper discussion of the relation between the operational and denotational semantics of LRT, the reader is referred to [10, 14].

**THE SCIENTIFIC CALCIULATOR**

The scientific calculator consists of basic operations (addition, subtraction, multiplication and division), trigonometric function (sin, cosine, tangent, cotangent, etc.) and additional function like \(x^2\), \(\sqrt{x}\), \(\pi\), among others.

In order to describe how an algorithm in LRT works we present a particular example. The average function defined by:

\[
x \oplus y = \frac{x + y}{2}
\]

can be implemented in LRT as follows:

```c
faverage(x, y) =
    if rtestC1 (x)
        then ConsC(faverage(TailLx, TailLy))
    else if rtestC1 (y)
        then ConsC(faverage(Taillx, Tailly))
    else Cons(faverage(Taillx, Tailly))

where
l = -1/2, c = 0, r = 1/2,
L = [-1, 0], C = [-1/2, 1/2],
R = [0, 1], C1 = [-3/4, 1/4], C2[-1/4, 3/4].
```

The intuition behind this program is the following. If both \(x\) and \(y\) are in the interval \(L\), then we know that \(x \oplus y\) is in the interval \(L\). If both \(x\) and \(y\) are in the interval \(R\), then we know that \(x \oplus y\) is in the interval \(R\), and so on. The boundary cases are taken care of by the rtest conditional.

What is interesting is that, despite the use of the multi-valued construction \(rtest\), the overall result of the computation is single valued. In other words, different computation paths will give different shrinking sequences of intervals, but all of them will shrink to the same number. A proof of this fact and of correctness of the program is provided in [9]. This can be seen as follows: an unfolding of \(1/2 \oplus 1/2\) gives the expression \(ConsC(faverage(0 \oplus 0))\). This means that the result of the operation is in the interval \(R = [0, 1]\). A second unfolding gives \(ConsRConsC(faverage(1 \oplus 1))\), due to it is a call by need language, the first two conses are reduced using Equation 1. This means that the result is in the interval \([1/3, 2/3]\). A repeated unfolding gives the required result 1/2.

We present and explain a pair of GMP-FC++ implementations of the operational semantics described in the previous section. The idea is to illustrate the straightforward translation of the algorithms presented in [10] to our framework and present an implementation of the trigonometric functions comparing its efficiency with previous functional programming implementations [18, 11].

**Example 1** An easy example is the representation of the real number 1 which can be coded as follows:

```c
struct InfiniteListOne :
    public CFunType<List<Intervalo>> { 
    List<Intervalo> const {
        Interval i1;
        mpf_init_set_ui(i1.lower, 0);
        mpf_init_set_ui(i1.upper, 1);
        return cons(i1, curry(InfiniteListOne));
    }
} listainfinita;
```
An unfolding of the program gives the interval \texttt{cons}(0, 1). Since the procedure calls itself, a second unfolding gives the intervals \texttt{cons}(0, 1)\texttt{cons}(0, 1). This procedure does not have a basic case, so a potential infinite list of intervals of the form \texttt{cons}(0, 1) is generated. Since FC++ can be used as a call-by-need language, a call to the procedure \texttt{InfiniteListOne} returns the reduction of intervals from left to right to the required precision applying Equation 1.

The \texttt{cons} operation presented in Equation 1 is implemented as follows:

```c
struct Conz : public CFunType<Interval, Interval, Interval>
{
    Interval operator()(Interval a, Interval x) const
    {
        mpf_t aux;
        mpf_init2(aux, Prec);
        mpf_init2(iC.lower, Prec);
        mpf_init2(iC.upper, Prec);
        mpf_sub(aux, a.upper, a.lower);
        mpf_div_ui(aux, aux, 2);
        mpf_mul(iC.upper, aux, x.upper);
        mpf_mul(iC.lower, aux, x.lower);
        mpf_add(aux, a.upper, a.lower);
        mpf_div_ui(aux, aux, 2);
        mpf_add(iC.lower, iC.lower, aux);
        mpf_add(iC.upper, iC.upper, aux);
        mpf_clear(aux);
        return iC;
    }
};
```

According to Equation 1 \texttt{cons} is a lineal function which takes two intervals as inputs and returns a single interval as output stated in the code by \texttt{< Interval, Interval, Interval>}. The initialization of variables in GMP is done by the function \texttt{mpf_init2}. This function takes two arguments, the variable to be initialized and its precision in terms of bits. In this case an auxiliary temporal variable and a global interval variable, in which the result is returned, are initialized. Basic operations like addition, subtraction, etc. are computed in GMP with especial procedures. These operations begin with the word \texttt{mpf}. The comments included in the code, indicate which operation is performed. The reader can compare the operations against Equation 1.

Finally, the procedure \texttt{mpf_clear}, free dynamic memory allocation used by GMP, in this procedure the unique local variable is aux. A similar implementation is coded for the \texttt{tail} function.

To approximate a real number, the first rule of the operational semantics is applied to the elements on the \texttt{mantissa} as many times as precision is required. If the first rule is not applied, a further evaluation of the input list should be done.

It is worth to note that the implementation of the operational semantics only works with real numbers in the interval \([-1, 1]\). The final result to the desired precision is calculated multiplying both interval end points at the head of the \texttt{mantissa} by 2 to the power of the exponent.

We hope that the discussion of the previous codes allows the reader to understand the implementation.

The trigonometric functions

In order to define the trigonometric functions an algorithm which compute the limit of Cauchy sequences is defined. The trigonometric functions can be defined as the limits of such sequences. The algorithm which compute the limit of Cauchy sequence was first presented in [18].

To define the limits function, an auxiliary function which takes three real number \((x, y, z)\) as input (in the form of infinite lists) is defined. The numbers have the properties that \(x \leq y \leq z\). The function examine \(x\) and \(z\) and generates a stream of conses whose range contains all possible values of \(y\) and then uses \(y\) directly to return the reminder of the output such that the whole output represent the number \(y\). Importantly, this function determines information about \(y\) from the bounds of \(x\) and \(z\) before examining \(y\) itself.

This methods seems weird since \(y\) is an input of the function, however the real number \(y\) is passed as a pointer to other functions which computes its value, hence it is better to use \(x\) and \(z\) to compute the value of \(y\) before examining \(y\). To compute limits it is required that this function has the property that if \(x\) and \(z\) are closer
than some finite amount a cons is generated, and that if $x$ and $z$ are equal, $y$ is not examined at all.

The function may be used to convert a real number represented as a stream of nested intervals into a single stream of conses. The function tries to find common digits from the first interval in the stream of nested intervals when possible, then end uses the remaining intervals to determine the rest of the conses in the stream. Because the nested intervals converge to a single value, we can get arbitrarily many output conses by examining a finite number of intervals. In order to compute the limit of a given Cauchy sequence, we can express the sequence as a stream of nested lower and upper bounds of the limit. Many useful functions can be defined as the limits of Cauchy sequences.

Suppose we have a stream of nested intervals represented as $(x_1, z_1), (x_2, z_2), (x_3, z_3), \ldots$ with the following properties:

$$[x_1, z_1] \supseteq [x_2, z_2] \supseteq [x_3, z_3] \supseteq \cdots$$

Let $f$ be the function that takes the infinite representation of three real numbers $x, y, z$ as input and outputs the stream $y$ by examining $x$ and $z$. We can compute the stream $y$ by the stream of nested intervals $(x_1, z_1), (x_2, z_2), (x_3, z_3), \ldots$ as the result of:

$$y = \lim [x_1, z_1), (x_2, z_2), (x_3, z_3), \ldots] = f(x_1, \lim [(x_2, z_2), (x_3, z_3), \ldots, z_3])$$

Computing trigonometric functions is performed as follows: suppose we wish to compute the value $f(x)$, if we find a series in terms of $x$ which either tends to a limit $f(x)$ with a known rate of convergence or converges to but oscillates round the limit $f(x)$. Once we have this series we can generate a sequence of upper and lower bounds on this limit at each term of the original sequence. We know that the sequence converges so we can use this fact to generate an infinite and strictly nested stream of intervals containing the limit.

For example, the sine function can be defined using the following series:

$$\sin(x) = \sum_{n=0}^{\infty} \frac{(-1)^{n+1} x^{2n+1}}{(2n+1)!}$$

Using this sum, we can easily construct a sequence $S_n(x)$ such that $S_n(x) \to \sin(x)$ as $n \to \infty$. In other words

$$\lim_{n \to \infty} S_n(x) = \sin(x)$$

which is equivalent to compute the limit of the nested intervals $(S_0, S_1), (S_1, S_2), \ldots$. The FC++ code to compute this function can be consulted at http://fi.uaemex.mx/rmacial/FC++. 

<table>
<thead>
<tr>
<th>Operation</th>
<th>Time reported in seconds for each implementation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Time reported by both the Haskell Glasgow and the FC++ compilers. All operations were calculated at precision $1E-5$</td>
</tr>
<tr>
<td></td>
<td>Operation</td>
</tr>
<tr>
<td></td>
<td>Haskell</td>
</tr>
<tr>
<td>$\sum_{i=1}^{n} \frac{i}{2^i+1}$</td>
<td>15.999878</td>
</tr>
<tr>
<td>$\sum_{i=1}^{n} \frac{i}{3^i+1}$</td>
<td>30.45</td>
</tr>
<tr>
<td>$\sum_{i=1}^{20} \frac{i}{7^i+1}$</td>
<td>111.85</td>
</tr>
<tr>
<td>$\sum_{i=1}^{20} \frac{i}{3^i+1}$</td>
<td>1583.99</td>
</tr>
<tr>
<td>$\sin(13 + \cos(7/9))(\cos(13 + \sin(7/11)))$</td>
<td>12.54</td>
</tr>
<tr>
<td>$\cos(13 + \tan(1/3))$</td>
<td>9.27</td>
</tr>
<tr>
<td>$\pi + \arctan(1/3) + (\cos(2/3) + \tan(13/15))$</td>
<td>23.97</td>
</tr>
<tr>
<td>$\pi + \arctan(1/3) + \pi$</td>
<td>5.33</td>
</tr>
<tr>
<td>$\cos(4/13 - \pi^2/31)$</td>
<td>27.109</td>
</tr>
<tr>
<td>$\arctan(1/3) + \pi$</td>
<td>12.88</td>
</tr>
<tr>
<td>$\sin(1/3 + \pi)/(\cos(1/3 + \sin(7/11))$</td>
<td>35.88</td>
</tr>
<tr>
<td>$\pi + \arctan(1/3) + (\cos(2/3) + \tan(2/3))$</td>
<td>43.23</td>
</tr>
</tbody>
</table>

Comparing with the three digits representation

The Haskell implementation we use in our comparison was presented at Marcial et al. [11]. We will not discuss the implementation in this paper, instead we refer the interested reader to the cited reference. We can say, however, that this implementation is more efficient than the Haskell in each operation. These results are shown at Table 1.

CONCLUSIONS

We have presented an implementation of LRT in the FC++ programming language using the GMP library.
Although C++ is an imperative language, FC++ is a functional C++ implementation, meaning that it allows a call by need evaluation and the definition of infinite lists. The algorithms presented in [11] were straightforward translated to this setting and the time reported is considerably improved compared to an implementation based on a pure functional programming language. In order to show that this implementation is faster, we used the logistic map which is chaotic function. However, our implementation is still slower than at least another C++ implementation called iRAM. A first further work is the implementation of trigonometric functions using Taylor series, e.g. the limit function has to be defined. A second further work is the improvement of the efficiency of the implementations in order to be as competitive as the C++ based.

REFERENCES