CR-LIBM: The evaluation of the exponential
C. Daramy, David Defour, Florent de Dinechin, Jean-Michel Muller

To cite this version:

HAL Id: hal-02102084
https://hal-lara.archives-ouvertes.fr/hal-02102084
Submitted on 17 Apr 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
CR-LIBM: The evaluation of the exponential

Catherine Daramy, INRIA
David Defour, LIP
Florent de Dinechin, LIP
Jean-Michel Muller, CNRS

Juillet 2003
CR-LIBM: The evaluation of the exponential

Catherine Daramy, INRIA
David Defour, LIP
Florent de Dinechin, LIP
Jean-Michel Muller, CNRS

Juillet 2003

Abstract

We present a new elementary function library, called CR-LIBM. This library implements the various functions defined by the Ansi99 C standard. It provides correctly rounded functions. When writing this library, our primarily goal was to certify correct rounding, and make it reasonably fast, and with a low utilisation of memory. Hence, our library can be used without any problem on real-scale problems. We are also giving the proof and the elements to understand the implementation of the exponential function of CR-LIBM.

Keywords: Elementary Functions, Exponential, CRlibm, correct rounding.

Résumé

Nous présentons une nouvelle bibliothèque d’évaluation de fonctions élémentaires, appelée CR-LIBM. Cette bibliothèque implémente les différentes fonctions définies par le standard Ansi C99. Sa principale caractéristique est de fournir l’arrondi correct pour la double précision et les quatre modes d’arrondi. Lors de l’écriture de cette bibliothèque, nos principaux objectifs étaient de certifier l’arrondi correct en ne dégradant pas les performances, et en limitant l’utilisation de mémoire. De ce fait, les fonctions de notre bibliothèque peuvent être utilisées sans problèmes dans des applications réelles. Nous donnons également pour l’implémentation de l’exponentielle, la preuve pour certifier l’arrondi correct, et les éléments nécessaires pour comprendre les choix faits.

Mots-clés: Fonctions élémentaires, Exponentielle, CRlibm, arrondi correct.
CR-LIBM: The evaluation of the exponential*

Contents

1 Introduction ................................................. 2

2 A methodology for efficient correctly-rounded functions ·· 2
  2.1 The Table Maker’s Dilemma ........................................... 2
  2.2 The onion peeling strategy ....................................... 3
  2.3 An overview of available mathematical libraries ................. 3

3 The Correctly Rounded Mathematical Library ................ 3
  3.1 Two steps are enough ........................................... 4
  3.2 Portable IEEE-754 FP for fast first step ....................... 4
  3.3 Software Carry-Save for an accurate second step .............. 4
  3.4 Current state of crlibm ......................................... 4

4 Notations and useful results ..................................... 4

5 Overview of the method for the exponential ................. 7

6 Quick phase .................................................. 7
  6.1 Handling special cases .......................................... 8
    6.1.1 Methods to raise IEEE-754 flags ................................. 8
    6.1.2 Avoiding overflows and underflows ......................... 8
    6.1.3 Rounding to nearest ......................................... 8
    6.1.4 Rounding toward $+\infty$ .................................... 9
    6.1.5 Rounding toward $-\infty$ .................................. 10
    6.1.6 Rounding toward 0 .......................................... 11
  6.2 The range reduction ........................................... 11
    6.2.1 First reduction step ........................................ 11
    6.2.2 Second range reduction ................................... 14

7 Polynomial evaluation ........................................ 15

8 Reconstruction ............................................... 16

9 Test if correct rounding is possible .......................... 22
  9.1 Rounding to nearest ........................................... 22
  9.2 Rounding toward $+\infty$ ....................................... 25
  9.3 Rounding toward $-\infty$ ....................................... 26
  9.4 Rounding toward 0 ............................................. 27

* Authors may be reached via e-mail at [Catherine.Daramy, David.Defour, Florent.de.Dinechin, Jean-Michel.Muller]@ens-lyon.fr. This text is also available as a research report of the Institut National de Recherche en Informatique et en Automatique http://www.inria.fr.
1 Introduction

The need for accurate elementary functions is important in many critical programs. Methods for computing these functions include table-based methods[13, 26], polynomial approximations and mixed methods[7]. See the books by Muller[24] or Markstein[22] for recent surveys on the subject.

The IEEE-754 standard for floating-point arithmetic[16] defines the usual floating-point formats (single and double precision). It also specifies the behavior of the four basic operators (+, −, ×, ÷) and the square root in four rounding modes (to the nearest, towards +∞, towards −∞ and towards 0). Its adoption and widespread use have increased the numerical quality of, and confidence in floating-point code. In particular, it has improved portability of such code and allowed construction of proofs on its numerical behavior. Directed rounding modes (towards +∞, −∞ and 0) also enabled efficient interval arithmetic[23, 17].

However, the IEEE-754 standard specifies nothing about elementary functions, which limits these advances to code excluding such functions. Currently, several options exist: on one hand, we can use today’s mathematical libraries that are efficient but without any warranty on the correctness of the results. When strict guarantees are needed, some multiple-precision packages like MPFR [3] offer correct rounding in all rounding modes, but are several orders of magnitude slower than the usual mathematical libraries for the same precision. The recently released IBM Ultimate Math Library[1] claims to offer correct rounding to the nearest, and this library is both portable and fast, if bulky. However, for reasons detailed below, this claim is not proven. Besides, this library still lacks directed rounding modes needed for interval arithmetic, and has other drawbacks that we analyze in the sequel.

The purpose of this paper is to show that the combination of several recent advances allows us to design a correctly rounded mathematical library which is fast enough to replace the existing libraries, at a minor cost in terms of performance and resources. Section 2 presents the context of this library. Section 3 presents the state of the library. Section 4 give some notations and results use in the sequel of this paper. Section 5 to 10 describe the exponential and prove the correct rounding of the exponential function. Section 11 analyzes this function, and shows that it is comparable in size and speed to other mathematical libraries.

2 A methodology for efficient correctly-rounded functions

2.1 The Table Maker’s Dilemma

With a few exceptions, the image $y$ of a floating-point number $x$ by a transcendental function $f$ is a transcendental number, and can therefore not be represented exactly in standard numeration systems. The only hope is to compute the floating-point number that is closest to (resp. immediately above or immediately below) the mathematical value, which we call the result correctly rounded to the nearest (resp. towards $+\infty$ or towards $-\infty$).

It is only possible to compute an approximation $\hat{y}$ to the real number $y$ with precision $\varepsilon$. This ensures that the real value $y$ belongs to the interval $[\hat{y} - \varepsilon, \hat{y} + \varepsilon]$. Sometimes however, this information is not
enough to decide correct rounding. For example, if \([\hat{y} - \varepsilon, \hat{y} + \varepsilon]\) contains the middle of two consecutive floating-point numbers, it is impossible to decide which of these two numbers is the correctly rounded to the nearest of \(y\). This is known as the Table Maker’s Dilemma (TMD).

2.2 The onion peeling strategy

A method described by Ziv [27] is to increase the precision \(\varepsilon\) of the approximation until the correctly rounded value can be decided. Given a function \(f\) and an argument \(x\), the value of \(f(x)\) is first evaluated using a quick approximation of precision \(\varepsilon_1\). Knowing \(\varepsilon_1\), it is possible to decide if rounding is possible, or if more precision is required, in which case the computation is restarted using a slower approximation of precision \(\varepsilon_2\) greater than \(\varepsilon_1\), and so on. This approach makes sense even in terms of average performance, as the slower steps are rarely taken.

However there was until recently no practical bound on the termination time of such an algorithm. This iteration has been proven to terminate, but the actual maximal precision required in the worst case is unknown. This might prevent using this method in critical application.

2.3 An overview of available mathematical libraries

Many high-quality mathematical libraries are freely available, including fdlibm, written by Sun[2] and libultim written by IBM[1], which are portable assuming IEEE-754 arithmetic, and processor-specific libraries by Intel[15, 4] and HP[22, 21] among other. Operating systems often include several mathematical libraries, some of which are derivatives of one of the previous.

Among these libraries, two offer correct correct rounding:

- The libultim library also called MathLib, is developed at IBM by Ziv and others [1]. It provides correct rounding, under the assumption that 800 bits are enough in all case. This approach suffers two weaknesses. The first is the absence of proof that 800 bits are enough: all there is is a very high probability. The second is that, as we will see in the sequel, for challenging cases, 800 bits are much of an overkill, which can increase the execution time up to 20,000 times a normal execution. This will prevent such a library from being used in real-time applications. Besides, to prevent this worst case from degrading average performance, there is usually some intermediate levels of precision in MathLib’s elementary functions, which makes the code larger, more complex, and more difficult to prove.

In addition this library provides correct rounding only to nearest. This is the most used rounding mode, but it might not be the most important as far as correct rounding is concerned: correct rounding provides a precision improvement over current mathematical libraries of only a fraction of a unit in the last place (ulp). Conversely, the three other rounding modes are needed to guarantee intervals in interval arithmetic. Without correct rounding in these directed rounding modes, interval arithmetic looses up to one ulp of precision in each computation.

- MPFR is a multiprecision package safer than libultim as it uses arbitrary multiprecision. It provides most of elementary functions for the four rounding modes defined by the IEEE-754 standard. However this library is not optimized for double precision arithmetic. In addition, as its exponent range is much wider than that of IEEE-754, the subtleties of denormal numbers are difficult to handle properly using such a multiprecision package.

3 The Correctly Rounded Mathematical Library

We have designed our own library called crlibm (correctly rounded mathematical library). It is based on the work of Lefèvre[20, 19] who computed the worst-case \(\varepsilon\) required for correctly rounding several functions in double-precision over selected intervals in the four IEEE-754 rounding modes. For example, he proved that 157 bits are enough to ensure correct rounding of the exponential function on all of its domain for the four IEEE-754 rounding modes.
3.1 Two steps are enough

Thanks to such results, we are able to guarantee correct rounding in two iterations only, which we may then optimize separately. The first of these iterations is relatively fast and provides between 60 and 80 bits of accuracy (depending on the function), which is sufficient in most cases. It will be referred throughout the paper as the Quick phase of the algorithm. The second phase, referred to as the Accurate phase, is dedicated to challenging cases. It is slower but has a reasonably bounded execution time, tightly targeted at Lefèvre’s worst cases.

Having a proven worst-case execution time lifts the last obstacle to a generalization of correctly rounded transcendental. Besides, having only two steps allows us to publish, along with each function, a proof of its correctly rounding behavior.

3.2 Portable IEEE-754 FP for fast first step

The computation of a tight bound on the approximation error of the first step ($\varepsilon_1$) is crucial for the efficiency of the onion peeling strategy: overestimating $\varepsilon_1$ means going more often than needed through the second step. As we want the proof to be portable as well as the code, our first steps are written in strict IEEE-754 arithmetic. On some systems, this means preventing the compiler/processor combination to use advanced floating-point features such as fused multiply-and-add or extended double precision. It also means that the performance of our portable library will be lower than optimized libraries using these features.

To ease these proofs, our first steps make wide use of classical, well proven techniques. In particular, when a result is needed in a precision higher than double precision (as is the case of $\hat{y}_1$, the result of the first step), it is represented as as the sum of two floating-point numbers. There are well-known algorithms for computing on such sums (for instance Sterbenz’ lemma, the Fast2Sum algorithm, the Dekker algorithm[18]) with mechanically checked proofs.

A sequence of simple tests on $\hat{y}_1$ allows to decide whether to go for the second step. The sequence corresponding to each rounding mode is shared by most functions and has also been carefully proven.

3.3 Software Carry-Save for an accurate second step

For the second step, we designed an ad-hoc multiple-precision library called Software Carry-Save library (scslib) which is lighter and faster than other available libraries for this specific application [11, 10]. This choice is motivated by considerations of code size and performance, but also by the need to be independent of other libraries: Again, we need a library on which we may rely at the proof level. This library is independent from the mathematical library and distributed separately [5].

3.4 Current state of crlibm

The library crlibm (correctly rounded mathematical library) currently offers accurate parts for the exponential, logarithm in radix 2, 10 and $\pi$, sine, cosine, tangent, arctangent, plus trigonometric argument reduction. The first quick part and its proof have only been written for the exponential thus far. The difficulty is to prove both the algorithm and the C program. The proof relies heavily on several shared lemmas, assuming the good behavior of the system composed of the compiler and the processor. Another difficulty is that performance is important. Therefore many parts of this proof could be done only by hand.

4 Notations and useful results

Throughout the paper, we will note $+, -, \times$ the usual mathematical operations, and $\oplus, \ominus$ and $\otimes$ the corresponding floating-point operations in IEEE-754 double precision, in the IEEE-754 round to nearest mode. To characterize the error we will use the following definition :

**Definition** $1 (\varepsilon_n)$ For any integer $n$, we will define by $\varepsilon_n$ a value $\alpha$ such that:

$$|\alpha| \leq 2^n$$
For a floating-point number \( x \), we will classically denote \( \text{ulp}(x) \) the value of the least significant bit of its mantissa.

We will make use of the following well-known results:

**Theorem 1 (Sterbenz Lemma [25, 14])** If \( x \) and \( y \) are floating-point numbers, and if \( y/2 \leq x \leq 2y \) then \( x \oplus y \) is computed exactly, without any rounding error.

A double precision floating-point number is coded on 64 bits, that is two times the size of an integer, usually represented with 32 bits in current processors. The order in which the two 32 bits words are stored in memory depends on the architecture. An architecture is said **Little Endian** if the lower part of the number is stored in memory at the smallest address; x86 processor use this representation. Conversely, an architecture with the higher part of the number stored in memory at the smallest address is said **Big Endian**; PowerPC processor use this representation.

The following code extracts the upper and lower parts of a double precision number \( x \) in a classical and relatively portable way.

**Listing 1: Extract upper and lower part of a double precision number \( x \)**

```c
// LITTLE_ENDIAN/BIG_ENDIAN are define by the user or */
// automatically by tools such as autoconf/automake. */

#define LITTLE_ENDIAN
#define HI(x) *(1+(int*)&x)
#define LO(x) *(int*)&x

#define BIGEndian
#define HI(x) *(int*)&x
#define LO(x) *(1+(int*)&x)

```

The previous code is also efficient on many architectures where FP and integer pipelines are hermetic, requiring conversions from one format to the other to be done through memory. There are probably architectures where a more efficient implementation can be found. Our code uses exclusively the previous two function for converting FP to integer and back, which ensures a quick implementation of such an architecture-specific optimization.

**Theorem 2 (Fast2sum algorithm [18])** For \( a \) and \( b \) two floating-point numbers, the following method computes two floating-point numbers \( s \) and \( r \), such that \( s + r = a + b \) exactly, and \( s \) is the floating-point number which is closest to \( a + b \).

**Listing 2: Fast2SumCond**

```c
#define Fast2SumCond(s, r, a, b) 
{ double z, _a=a, _b=b; 
  s = _a + _b; 
  if ( ((HI(_a)&0x7FF00000) > (HI(_b)&0x7FF00000)) ) { 
    z = s - _a; 
    r = _b - z; 
  } else { 
    z = s - _b; 
    r = _a - z; 
  } 
}
```

This algorithm requires 3 floating-point additions, 2 masks and 1 test over integer.

In the case we know that the exponent of \( a \) is greater than the one of \( b \), then the previous algorithm to perform an exact addition of 2 floating-point numbers becomes:

**Listing 3: Fast2Sum**

```c
#define Fast2Sum(s, r, a, b) 
{ double z, _a=a, _b=b; 
  s = _a + _b; 
  z = s - _a; 
  r = _b - z; 
}
```

The cost of this algorithm is 3 floating-point additions.
Theorem 3 (double multiplication[12, 18]) Let \( a \) and \( b \) be two floating-point numbers, with \( p \geq 2 \) the size of their mantissa. Let \( c = 2^{52} + 1 \). The following method computes the two floating-point numbers \( r_1 \) and \( r_2 \) such that \( r_1 + r_2 = a + b \) with \( r_1 = a \otimes b \) in the case \( p = 53 \) (double precision):

### Listing 4: DekkerCond
```c
void inline DekkerCond (double *r1, double *r2, double a, double b) {
    double two_m53 = 1.1102230246251565404e-16; /* 0x5CA00000, 0x00000000 */
    double two_53 = 2.8217054991008458516e-01; /* 0x24100000, 0x00000000 */
    double c = 134217728.; /* 0x41A00000, 0x02000000 */
    double u, up, u1, u2, v, vp, v1, v2, r1, r2;

    if (HI(a)>0x7C900000) u = a * two_m53;
    else u = a;

    if (HI(b)>0x7C900000) v = b * two_m53;
    else v = b;

    up = u * c; vp = v * c;
    u1 = (u - up) + up; v1 = (v - vp) + vp;
    u2 = u - u1; v2 = v - v1;

    *r1 = u * v;
    *r2 = (((u1 * v1 * r1) + (u1 * v2)) + (u2 * v1)) + (u2 * v2);
}
```

We have to test \( a \) and \( b \) before and after the core of the algorithms in order to avoid overflow by multiplying by \( c \). The global cost in the worst case is 4 tests over integers, 10 floating-point additions and 13 floating-point multiplications.

If we are know that \( a \) and \( b \) are less then \( 2^{970} \) we can skip this test, and get the following algorithm:

### Listing 5: Dekker
```c
void inline Dekker (double *r1, double *r2, double u, double v) {
    double c = 134217728.; /* 0x41A00000, 0x02000000 */
    double up, u1, u2, vp, v1, v2;

    up = u * c; vp = v * c;
    u1 = (u - up) + up; v1 = (v - vp) + vp;
    u2 = u - u1; v2 = v - v1;

    *r1 = u * v;
    *r2 = (((u1 * v1 * r1) + (u1 * v2)) + (u2 * v1)) + (u2 * v2);
}
```

which reduces the cost of this algorithm to 10 floating-point additions and 7 floating-point multiplications.

It should be noted that the availability of fused multiply-and-add (FMA), with only one rounding, on architectures like PowerPC and IA-64, allows the implementation of the Dekker algorithm in only two operations: \(*r1 = u*v; *r2 = FMA(u*v-r1)\); Again, this is an architecture-dependent optimization.

Theorem 4 (Conversion from floating-point to integer [6]) The following algorithm convert a floating-point number \( d \) into an integer \( i \) with rounding to nearest mode.

### Listing 6: Solution 2
```c
#define DOUBLE2INT(i, d) \
    { double t = (d + 6755399441055744.0); i = LO(t); }
```

This algorithm add the constant \( 2^{52} + 2^{51} \) to the floating-point number to put the integer part of \( x \), in the lower part of the floating-point number. We use \( 2^{52} + 2^{51} \) and not \( 2^{52} \), because the value \( 2^{51} \) is used to contain possible carry propagations with negative numbers.
5 Overview of the method for the exponential

We are now going to present and prove the correct rounding of the evaluation scheme chosen for the exponential within crlibm. We will use a property deduced from the enumeration of worst cases done for the exponential function by Lefèvre [19].

Property 1 (Correct rounding of the exponential) Let \( y \) be the result of the exponential of a floating-point number \( x \) in double precision. Let \( y^* \) be an approximation of \( y \) such that the distance between mantissa of \( y \) and \( y^* \) is bounded by \( \varepsilon \).

If \( \varepsilon \leq 2^{157} \) then rounding \( y^* \) is equivalent to rounding \( y \) for the four rounding modes.

We have done the evaluation of the exponential in two steps. First, we use the quick phase of the algorithm to get an approximation good to 68 bits of the result. Then we perform a test to check whether we need to use the accurate phase, based on multiprecision operators from SCSlib [5].

To increase the trust in the code, with have included constants in hexadecimal format (big endian only for concision). However to help the reader we are giving the corresponding decimal values.

6 Quick phase

Here is the general scheme chosen for the first step of the evaluation of the exponential:

1. “Mathematical” range reduction
   We compute the reduced argument
   \[
   (r_{hi} + r_{lo}) \in [-\ln(2)/2, +\ln(2)/2]
   \]
   such that:
   \[
   x = k.\ln(2) + (r_{hi} + r_{lo})
   \]
   therefore
   \[
   \exp(x) = \exp(r_{hi} + r_{lo}).2^k
   \]

2. Tabular range reduction
   Let \( \text{index}_{flt} \) be the first 8 bits of \( (r_{hi} + r_{lo}) \) and \( (rp_{hi} + rp_{lo}) = (r_{hi} + r_{lo}) - \text{index}_{flt} \), such that \( (rp_{hi} + rp_{lo}) \in [-2^{-9}, +2^9] \). We have
   \[
   \exp(r_{hi} + r_{lo}) = \exp(index_{flt}) \times \exp(rp_{hi} + rp_{lo})
   \]
   where \( \exp(index_{flt}) = (ex_{hi} + ex_{lo}) \) will be looked up in a table.

3. Polynomial evaluation
   We evaluate the polynom \( P_r \) of degree 3 such that:
   \[
   \exp(rp_{hi} + rp_{lo}) \approx 1 + (rp_{hi} + rp_{lo}) + \frac{1}{2}(rp_{hi} + rp_{lo})^2 + (rp_{hi} + rp_{lo})^3.\text{(P}_r\text{)}
   \]
   with \( P_r = c_0 + c_1 rp_{hi} + c_2 rp_{hi}^2 + c_3 rp_{hi}^3 \) and \( rp_{hi} \in [-2^{-9}, +2^9] \)

4. Reconstruction
   \[
   \exp(x) = 2^k.(ex_{hi} + ex_{lo}).(1 + (rp_{hi} + rp_{lo}) + \frac{1}{2}(rp_{hi} + rp_{lo})^2 + (rp_{hi} + rp_{lo})^3.\text{P}_r\text{)}.(1 + \varepsilon_{-68})
   \]
6.1 Handling special cases

6.1.1 Methods to raise IEEE-754 flags

The IEEE standard requires, in certain cases, to raise flags and exceptions for the operators $+,$ $\times,$ $\div,$ $\sqrt{}$. Therefore, it is legitimate to require the same for elementary functions. For portability, these exceptions and flags will be generated using the following techniques:

- **underflow**: The multiplication $\pm smallest \times smallest$ where $smallest$ correspond to the smallest denormal number,
- **overflow**: The multiplication $\pm largest \times largest$ where $largest$ correspond to the largest normal number,
- **division by zero**: The division $\pm 1.0/0.0$,
- **inexact**: The addition $(x + smallest) - smallest$ where $x$ is the result,
- **invalid**: The division $\pm 0.0/0.0$.

6.1.2 Avoiding overflows and underflows

In the sequel of this paper, we will consider input numbers in the range $[u_{\text{bound}}, o_{\text{bound}}]$, where $u_{\text{bound}}$ and $o_{\text{bound}}$ are:

\[
\begin{align*}
u_{\text{bound}} &= \triangle (\ln ((1 - 2^{-53}) .2^{-1075})) = -745.1332 \ldots \\
o_{\text{bound}} &= \triangledown (\ln ((1 - 2^{-53}) .2^{1024})) = 709.7827 \ldots
\end{align*}
\]

where $\triangle(x)$ and $\triangledown(x)$ respectively correspond to the rounding toward $+\infty$ and $-\infty$ of $x$. In the rounding mode to nearest, the exponential of a number greater than $o_{\text{bound}}$ is an overflow, whereas the exponential of a number less than $u_{\text{bound}}$ is rounded to 0, and raises an inexact flag.

However, subtler under/overflow situations may arise in two cases, which we should avoid:

- An intermediate computation may raise an overflow although the final result is representable as an IEEE-754 floating-point number.
- In IEEE-754 arithmetic, when a result is between $2^{-1023}$ and $2^{-1074}$, a gradual underflow exception arises to signal that the precision of the result is reduced in a drastic way.

In both cases, as we will show in the following, it is possible to avoid the exception by predicting that it will occur, and appropriately scaling the input number in the range reduction phase.

6.1.3 Rounding to nearest

```
static const union {int i[2]; double d;}
#endif BIG_ENDIAN
#define largest _largest.d
#define smallest _smallest.d
#define u_bound _u_bound.d
#define o_bound _o_bound.d
unsigned int hx;
hx = HI(x);
hx &= 0x7fffffff;
```
/* Filter special cases */
if (hx == 0x40862E42) {
    if (hx >= 0x7ff00000) {
        return x+x; /* NaN */
    } else return ( (hx&0x80000000) ==0) ? x : 0.0 ; /* exp(+/- inf) = inf ,0 */
} else if (hx <= 0x3C900000) return 1.; /* if (hx <= 2^(-54)) */

∥ Proof. ∥

line 17 Put the high part of x in hx. (cf. prog. 1)
line 18 Remove the sign information within hx. It will make tests on special cases simpler.
line 21 Test equivalent to if(|x| > 709.7822265625). This test is true if x > u_bound, x < o_bound, x = ±inf or x = NaN. This test is performed with integers to make it faster.
line (22-24) Test if x = ±inf or x = NaN and give the corresponding results (exact +∞ or 0).
line 27 Under the assumption that the compiler correctly translates the floating-point number we have o_bound = 390207173010335/549755813888. If x > o_bound then exp(x) = + inf. The multiplication largest*largest leaves to the compiler the generation of an overflow and the corresponding flags.
line 28 Under the assumption that the compiler correctly translates the floating-point number we have u_bound = -3277130554578985/4398046511104. If x < u_bound then exp(x) = +0. The multiplication largest*largest leaves to the compiler the generation of an underflow and the corresponding flags.
line 32 Test equivalent to if(|x| ≤ 2^(-54)). This test is performed with integers to make it faster and is valid because x ≠ {NaN,∞}. In addition, this test allows to handle cases when x is a denormal number. We have the following property:

<1> |x| > 2^(-54) and x ≠ {NaN,∞} •

Indeed, in rounding to nearest, if |x| ≤ 2^-54 then exp(x) = 1.0. This test prevents to encounter a denormal number in the rest of the program.

6.1.4 Rounding toward +∞

Listing 8: Handling special cases in rounding toward +∞

static const union {int i[2]; double d;}
#eliff BIG_ENDIAN
    _largest = {0x7fefefff , 0xfffefff },
    _smallest = {0x00000000 , 0x00000001 },
    _u_bound = {0xC0874910 , 0xD52D3052 } , /* -7.4513321910194122122107e+02 */
    _o_bound = {0x40862E42 , 0xFEFA39F0 } , /* 7.09782712893384086783e+02 */
    two_m52_56 = {0x3CB10000 , 0x00000000 }; /* 2.35922392732845764840e-16 */
#else
    ...
#endif
#define largest _largest.d
#define smallest _smallest.d
#define u_bound _u_bound.d
#define o_bound _o_bound.d
#define two_m52_56 _two_m52_56.d

unsigned int hx;
David Defour and Florent de Dinechin

10

10 David Defour and Florent de Dinechin

20 /∗ Filter special cases ∗/
21 if (hx >= 0x40862E42)
22 "```c
23 if ((hx&0x000fffff)|LO(x))!=0)
24 "```c
25 "```c
26 return x+x; /* NaN */
27 "```c
28 if (hx >= 0x7ff00000 ) {
29 if (((hx&x000fffff)|LO(x))!=0)
30 return x+x; /* NaN */
31 else return ((hx&60x80000000)==0) ? x:0.0; /* exp(+/- inf) = inf,0 */
32 }
33 if (x > o_bound) return largest.largest ; /* overflow */
34 if (x < u_bound) return smallest.smallest; /* 2^(−1074) */
35 }
36 if (hx < 0x3CA00000) {
37 if (HI(x) < 0)
38 return 1. + smallest ; /* 1 and inexact */
39 else
40 return 1. + two_m52_56 ; /* 1 + 2^(-52) and inexact */
41 }
42 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 "```c
43 “/∗ Proof. This program is similar to the one used in rounding to nearest mode with the following exceptions:
44 “/∗
45 “/∗ When (x < u_bound), in rounding toward +∞, we have to return as result the smallest representable number (2^{−1074}) with the inexact flag raised.
46 “/∗
47 “/∗ When (|x| < 2^{−53}), in rounding toward +∞, we have to return as result 1.0 if x < 0 with the inexact flag raised or 1 + 2^{−52} with the inexact flag raised if x > 0.
48 "*/
49
50 6.1.5 Rounding toward −∞
51
52 Listing 9: Handling special cases in rounding toward −∞
53
54 static const union {int i[2]; double d;}
55 #ifdef BIG_ENDIAN
56 _largest = {0x7fefffff , 0xffffffff },
57 _smallest = {0x00000000 , 0x00000001} ,
58 _u_bound = {0xC0874910 , 0xD52D3052} , /* -7.45133219101941222107e+02 */
59 _o_bound = {0x40862E42 , 0xFEFA39F0} , /* -7.09782712893384086783 e+02 */
60 _two_m52_56 = { 0x3CB10000 , 0x00000000} ; /* 2.35922392732845764840 e-16 */
61 #else ...
62 #endif
63 #define largest _largest.d
64 #define smallest _smallest.d
65 #define u_bound _u_bound.d
66 #define o_bound _o_bound.d
67 #define two_m52_56 _two_m52_56.d
68
69 unsigned int hx ;
70
71 hx = HI(x);
72 hx &= 0x7fffffff ;
73
74 /* Filter special cases */
75 if (hx >= 0x40862E42)
76 if (hx >= 0x7ff00000 ){
77 return x+x; /* NaN */
78 else return ((hx&60x80000000)==0) ? x:0.0; /* exp(+/- inf) = inf,0 */
79 }
80 if (x > o_bound) return largest.largest ; /* overflow */
81 if (x < u_bound) return smallest.smallest; /* 2^(−1074) */
82 }
83 if (hx < 0x3CA00000) {
84 if (HI(x) < 0)
85 return 1. + smallest ; /* 1 and inexact */
86 else
87 return 1. + two_m52_56 ; /* 1 + 2^(-52) and inexact */
88 }
Proof. This program is similar to the one used in rounding to nearest mode with the following exceptions:

- When \((x > o\_bound)\), in rounding toward \(-\infty\), we have to return as result the largest representable number \((1 - 2^{-53}) \times 2^{1024}\) with the inexact flag raised.

- When \(|x| < 2^{-53}\), in rounding toward \(-\infty\), we have to return as result \(1.0 - 2^{-52}\) if \(x < 0\) with the inexact flag raised or 1.0 with the inexact flag raised if \(x > 0\).

\[\square\]

6.1.6 Rounding toward 0

The exponential function is continuous and positive, therefore rounding toward 0 is equivalent to rounding toward \(-\infty\).

6.2 The range reduction

6.2.1 First reduction step

The purpose of this first range reduction is to replace the input number \(x \in [u\_bound, o\_bound]\) with two floating-point numbers \(r\_hi, r\_lo\) and an integer \(k\) such that:

\[x = k \cdot \ln(2) + (r\_hi + r\_lo) \cdot (1 + \varepsilon)\]

with \(|r\_hi + r\_lo| < \frac{1}{2} \ln(2)\)

This “additive” range reduction may generate a cancellation if \(x\) is close to a multiple of \(\ln(2)\). A method from Kahan based on continuous fractions (see Muller [24] pp 154) allows us to compute the worst cases for the range reduction. Examples of results are given in Table 2.

<table>
<thead>
<tr>
<th>Interval</th>
<th>Worst cases</th>
<th>Number of bits lost</th>
</tr>
</thead>
<tbody>
<tr>
<td>([-2^{1024}, 2^{1024}])</td>
<td>5261692873635770 × 2^{499}</td>
<td>66, 8</td>
</tr>
<tr>
<td>([-1024, 1024])</td>
<td>7804143460206699 × 2^{51}</td>
<td>57, 5</td>
</tr>
</tbody>
</table>

Table 2: Worst cases corresponding to the closest number multiple to \(\ln(2)\), for the additive range reduction of the exponential. The maximum number of bits lost by cancellation is also indicated.

The interval \([u\_bound, o\_bound]\) on which we are evaluating the exponential is included within \([-1024, 1024]\). Therefore at most 58 bits can be cancelled during the subtraction of the closest multiple of \(\ln(2)\) to the input number \(x\).

Theorem 5 The sequence of instructions of the program 10 computes two floating-point numbers in double precision \(r\_hi, r\_lo\) and an integer \(k\) such that

\[r\_hi + r\_lo = (x - k \times \ln 2) + \varepsilon_{-69}\]

with \(k\) the closest integer to \(x / \ln 2\).

Listing 10: First range reduction

```c
static const union {int i[2]; double d;}

#ifdef BIG_ENDIAN
    _ln2_hi = {0x3FE62E42, 0xFEFA3800} /* 6.93147180559890330187 * 10^-01 */
    _ln2_me = {0x3D2EF357, 0x93C76000} /* 5.4979230187072095198 * 10^-14 */
    _ln2_lo = {0x3A8CC01F, 0x97B57A08} /* 1.16122272293625324218 * 10^-26 */
    _inv_ln2 = {0x3FF71547, 0x6533245F} /* 1.44269504088896338700 * 10^0 */
#else
    ...
#endif
#define ln2_hi _ln2_hi.d
#define ln2_me _ln2_me.d
```
#define ln2_lo _ln2_lo
#define inv_ln2 _inv_ln2

double r_hi, r_lo, rp_hi, rp_lo;
double u, tmp;
int k;

DOUBLE2INT(k, x * inv_ln2)

if (k != 0) {
    * r_hi+r_lo = x - (ln2_hi + ln2_me + ln2_lo)*k */
    rp_hi = x-ln2_hi*k;
    rp_lo = -ln2_me*k;
    Fast2SumCond(r_hi, u, rp_hi, rp_lo);
    r_lo = u - ln2_lo*k;
} else {
    r_hi = x; r_lo = 0.;
}

\begin{proof}

\begin{enumerate}
\item By construction: \( ln2_hi + ln2_me + ln2_lo = \ln(2)(1 + \varepsilon_{-140}) \)
\item \(|ln2_hi| \leq 2^0 \quad |ln2_me| \leq 2^{-44} \quad |ln2_hi| \leq 2^{-86} \)
\item \(ln2_hi\) and \(ln2_me\) hold at most 42 bits of precision
\item Put in \(k\) the closest integer of \(x * inv\_ln2\). We use the property of DOUBLE2INT that converts a floating-point number in rounding to nearest mode (program 6, page 6). In addition \(k\) satisfies the following property:
\item \(|x \times inv\_ln2| \leq k \leq \lceil x \times inv\_ln2 \rceil \quad \text{et} \quad |k| \leq \frac{2}{2} \times inv\_ln2 \)
\item We have seen in Section 6.1.2: \(-745.1332 \ldots < x < 709.7827 \ldots\), then:
\item \(-1075 \leq k \leq 1025 \quad \text{et} \quad |k| \text{ is an integer on at most 11 bits}\)
\end{enumerate}
\end{proof}
Properties <4> and <6> give us:

\[ \ln_2 \text{hi} \otimes k = \ln_2 \text{hi} \times k \quad \text{and} \quad \ln_2 \text{me} \otimes k = \ln_2 \text{me} \times k \text{ exactly} \]

By property <5> we have:

\[
(x \times \text{inv}_2 - 1) \times \ln_2 \text{hi} \leq k \times \ln_2 \text{hi} \leq (x \times \text{inv}_2 + 1) \times \ln_2 \text{hi}
\]

\[ x/2 \leq k \times \ln_2 \text{hi} \leq 2x \]

By the Sterbenz theorem (theorem 1, page 5), we have

\[ x \otimes (\ln_2 \text{hi} \otimes k) = x - (\ln_2 \text{hi} \otimes k) \]

Combined with property <7> we have:

\[ x \otimes (\ln_2 \text{hi} \otimes k) = x - (\ln_2 \text{hi} \times k) \text{ exactly} \]

We use conditional Fast2Sum algorithm (with tests on entries), because \( x - \ln_2 \text{hi} \times k \) can be equal to zero (due to the 58 bits of cancellation). The conditional Fast2Sum algorithm (program 2, page 5) leads to

\[ r_{hi} + u = (x \otimes \ln_2 \text{hi} \otimes k) + (-\ln_2 \text{me} \otimes k) \]

With properties <7> and <8> we have:

\[ r_{hi} + u = (x - \ln_2 \text{hi} \times k) + (-\ln_2 \text{me} \times k) \text{ exactly} \]

By the property <6> we have:

\[ |\ln_2 \text{lo} \times k| \leq 2^{-75}, \]

\[ \ln_2 \text{lo} \otimes k = (\ln_2 \text{lo} \times k). (1 + \varepsilon_{-54}) \]

\[ r_{lo} = u \otimes (\ln_2 \text{lo} \otimes k) = (u - (\ln_2 \text{lo} \otimes k)). (1 + \varepsilon_{-54}) = (u - (\ln_2 \text{lo} \times k). (1 + \varepsilon_{-54})). (1 + \varepsilon_{-54}) = (u - (\ln_2 \text{lo} \times k)). (1 + \varepsilon_{-54}) + \varepsilon_{-129} + \varepsilon_{-183} \]

That gives us:

\[ r_{lo} = (u - (\ln_2 \text{lo} \times k)). (1 + \varepsilon_{-54}) + \varepsilon_{-129} + \varepsilon_{-183} \]

We have:

\[ r_{hi} + r_{lo} = r_{hi} + (u - (\ln_2 \text{lo} \times k)). (1 + \varepsilon_{-54}) + \varepsilon_{-129} + \varepsilon_{-183} \]

\[ = (x - \ln_2 \text{hi} \times k) + (-\ln_2 \text{me} \times k) - (\ln_2 \text{lo} \times k) + (u - (\ln_2 \text{lo} \times k)). \varepsilon_{-54} + \varepsilon_{-129} + \varepsilon_{-183} \]

\[ = (x - k \ln(2)) + k. \varepsilon_{-140} + (u - (\ln_2 \text{lo} \times k)). \varepsilon_{-54} + \varepsilon_{-129} + \varepsilon_{-183} \]

\[ = (x - k \ln(2)) + (u - (\ln_2 \text{lo} \times k)). \varepsilon_{-54} + \varepsilon_{-128} + \varepsilon_{-183} \]

In the worst case, we are losing at most 58 bits by cancellation (Table 2). By property <9>, we deduce that \( u = 0 \) in this case, property <10> \(|\ln_2 \text{lo} \times k| \leq 2^{-75}\) gives us:

\[ r_{hi} + r_{lo} = (x - k \times \ln 2) + \varepsilon_{-127}+58=-69 \]

In addition 69 bits is a precision that can be represented as the sum of 2 floating-point numbers in double precision.
If $k = 0$ then no subtraction is necessary, then $r_{hi} + r_{lo} = x$ exactly.

At the end of this first range reduction we have:

\[
\exp(x) = 2^k \cdot \exp(r_{hi} + r_{lo} + \epsilon_{-69}) = 2^k \cdot \exp(r_{hi} + r_{lo}) (1 + \epsilon_{-69})
\]

### 6.2.2 Second range reduction

The number $(r_{hi} + r_{lo})$ is still too big to be used in a polynomial evaluation. A second range reduction needs to be done. This second range reduction is based on the additive property of the exponential $e^a + e^b = e^{a+b}$, and on the tabulation of some values of the exponential.

Let $index_{flt}$ be the first bits of $(r_{hi} + r_{lo})$, then we have:

\[
\exp(r_{hi} + r_{lo}) = \exp(index_{flt}) \cdot \exp(r_{hi} + r_{lo} - index_{flt}) \\
\approx (ex_{hi} + ex_{lo}) \cdot \exp(rp_{hi} + rp_{lo})
\]

where $ex_{hi}$ and $ex_{lo}$ are double precision floating-point numbers extracted from a table addressed by $index_{flt}$, such that $ex_{hi} + ex_{lo} \approx \exp(index_{flt})$. The input argument after this reduction step will be represented as the sum of two double precision floating-point numbers $rp_{hi}$ and $rp_{lo}$ such that

\[
rp_{hi} + rp_{lo} = r_{hi} + r_{lo} - index_{flt}
\]

Tests show that the optimal table size for the range reduction is 4KBytes [8]. If we want to store these values and keep enough precision (at least 69 bits), we need two floating-point numbers (16 bytes) per value.

Let $\ell$ be the parameter such that $[-2^{-\ell-1}, 2^{-\ell-1}]$ is the range after reduction, we want:

\[
\lceil \ln 2 \cdot 2^\ell \rceil 16 \leq (2^{12} = 4096)
\]

With $\ell = 8$ we have $\lceil \ln(2) \cdot 2^8 \rceil$ 16 bytes = 2848 bytes, and the evaluation range is reduced to $[-2^{-9}, 2^{-9}]$. After this reduction step, we have $|rp_{hi} + rp_{lo}| \leq 2^{-9}$.

The corresponding sequence of instructions performing this second range reduction is:

Listing 11: Second range reduction

```c
/* Constants definition */
static const union {int i[2]; double d;}
#ifdef BIG_ENDIAN
	_two_44_43 = {0x42B80000, 0x00000000}; /* 26388279066624. */
#else
...
#endif
#define two_44_43 _two_44_43.d
#define bias 89;
#define ex_hi, ex_lo, index_flt;
int index;

index_flt = (r_hi + two_44_43);
index = LO(index_flt);
index_flt -= two_44_43;
index += bias;
index_flt -= index_flt;

/* Results normalization */
Fast2Sum(rp_hi, rp_lo, r_hi, r_lo)

/* Table lookup */
ex_hi = tab_exp[index][0];
ex_lo = tab_exp[index][1];
```

Proof.
The constant \( \text{two}_44_43 = 2^{44} + 2^{43} \) is used in rounding to nearest mode to extract the \( \ell = 8 \) leading bits of \( r_{hi} + r_{lo} \).

In the C language, tables have positive indices. We consider positive values as well as negative ones for \( index \), therefore we need to use a bias equal to \( 178/2 = 89 \).

This sequence of instructions is similar to the one used within DOUBLE2INT (program 6, page 6). It puts in \( index \) variable, bits of weight \( 2^0 \) to \( 2^{-8} \), minus the value of the bias. Meanwhile, it puts in \( index_{flt} \) the floating-point value corresponding to the first 8 bits of \( r_{hi} \).

In line 85 we have:

\[
\texttt{<13> } r_{hi} = r_{hi} - index_{flt} \text{ exactly } \bullet
\]

The Fast2Sum algorithm guarantee:

\[
\texttt{<14> } |r_{hi}| \leq 2^{-9} \text{ and } |r_{lo}| \leq 2^{-63},
\]

\[
r_{hi} + r_{lo} = r_{hi} + r_{lo} \text{ exactly } \bullet
\]

We perform table lookup of the 2 values \( ex_{hi} \) and \( ex_{lo} \). The table is built such that only one cache miss can be encountered in these two table lookups.

By construction of the table \( tab_{exp} \) we have:

\[
\texttt{<15> } |ex_{hi}| \leq 2^{-1} \text{ and } |ex_{lo}| \leq 2^{-55}
\]

\[
ex_{hi} + ex_{lo} = \exp(index_{flt}).(1 + \varepsilon_{-109}) \bullet
\]

At the end of this second range reduction we have:

\[
\exp(x) = 2^k.(ex_{hi} + ex_{lo}).\exp(r_{hi} + r_{lo}).(1 + \varepsilon_{-69}).(1 + \varepsilon_{-109})
\]

7 Polynomial evaluation

Let \( r = (r_{hi} + r_{lo}) \), we need to evaluate \( \exp(r) \) with \( r \in [-2^{-9}, 2^{-9}] \). We will evaluate \( f(r) = \frac{(\exp(r) - 1 - r - \frac{r^2}{2})}{r^3} \) with the following polynom of degree 3:

\[
P(r) = c_0 + c_1 r + c_2 r^2 + c_3 r^3
\]

where

- \( c_0 = 6004799503160629/36028797018963968 \leq 2^{-2} \)
- \( c_1 = 750599937895079/18014398509481984 \leq 2^{-4} \)
- \( c_2 = 300240009245077/36028797018963968 \leq 2^{-6} \)
- \( c_3 = 3202560062254639/2305843009213693952 \leq 2^{-9} \)

with \( c_0, c_1, c_2, c_3 \) exactly representable by double precision floating-point numbers.

By using \( \text{infnorm} \) function from Maple we get the following error:

\[
\texttt{<16> } \exp(r) = (1 + r + \frac{r^2}{2} + r^3).P(r).(1 + \varepsilon_{-78}) \text{ with } r \in [-2^{-9}, 2^{-9}] \bullet
\]

For efficiency reason, we will evaluate \( P(r_{hi}) \) instead of \( P(r_{hi} + r_{lo}) \). The error corresponding to this approximation is:

\[
P(r_{hi} + r_{lo}) - P(r_{hi}) = c_1 r_{lo} + c_2 (r_{lo}^2 + 2 r_{hi} r_{lo}) + c_3 (r_{lo}^3 + 3 r_{hi} r_{lo}^2 + 3 r_{hi}^2 r_{lo}) \leq \varepsilon_{-67} + \varepsilon_{-75}
\]

The property \( <16> \) becomes:
\[ \exp(r) = (1 + r + \frac{1}{2}r^2 + r^3)P(rp_{hi}) + \varepsilon_{-78} + \varepsilon_{-86} \text{ with } r \in [-2^{-9}, 2^{-9}] \]

\[ P_r = P(rp_{hi}) \] is evaluated by the following sequences of instructions:

```
static const union {int i[2]; double d;}
#define BIG_ENDIAN
_c0 = {0x3FC55555, 0x55555535}; /* 1.6666666666665769236 e-01 */
_c1 = {0x3FA55555, 0x55555538}; /* 4.16666666666643257 e-02 */
_c2 = {0x3F811111, 0x31931950}; /* 8.33333427945838573823 e-03 */
_c3 = {0x3F5E6C16C, 0x3DC3DC5E}; /* 1.3888903080471677251 e-03 */
#else
...  
#endif
#define c0 _c0.d
#define c1 _c1.d
#define c2 _c2.d
#define c3 _c3.d

double P_r;

P_r = (c_0 + rp_{hi} * (c_1 + rp_{hi} * (c_2 + (rp_{hi} * c_3))));
```

\[ \square \]

\[ |P_r| \leq 2^{-2} \]

\[ P_r = (c_0 + rp_{hi} * (c_1 + rp_{hi} * (c_2 + (rp_{hi} * c_3)))) + \varepsilon_{-55} + \varepsilon_{-65} \]

\[ P_r = P(rp_{hi}) + \varepsilon_{-55} + \varepsilon_{-65} \]

By the properties \(<17>\) and \(<18>\):

\[ \exp(r) = (1 + r + \frac{1}{2}r^2 + r^3)P_r + \varepsilon_{-78} + \varepsilon_{-81} \text{ with } r \in [-2^{-9}, 2^{-9}] \]

\[ \square \]

At the end of the polynomial evaluation scheme we have:

\[ \exp(x) = 2^k.(ex_{hi} + ex_{lo}).(1 + r + \frac{1}{2}r^2 + r^3)P_r + \varepsilon_{-78} + \varepsilon_{-81}).(1 + \varepsilon_{-69}).(1 + \varepsilon_{-109}) \]

### 8 Reconstruction

Along previous step of the algorithm we get the following results:

- \( k, r_{hi} \) and \( r_{lo} \) during the additive range reduction,
- \( ex_{hi}, ex_{lo}, rp_{hi} \) and \( rp_{lo} \) during the table range reduction,
- \( P_r \) with the polynomial range reduction.

The reconstruction step consist in merging all these results in order to get \( \exp(x) \). This step is based on the following mathematical formula:

\[ \exp(x) = 2^k.(ex_{hi} + ex_{lo}).(1 + (rp_{hi} + rp_{lo}) + \frac{1}{2}(rp_{hi} + rp_{lo})^2 + (rp_{hi} + rp_{lo})^3).P_r \]
However, some terms in this equation are too small compared to dominants terms, and should not be taken into account: We approximate:

\[ \text{Rec} = (\text{ex}_hi + \text{ex}_lo)(1 + (\text{rp}_hi + \text{rp}_lo) + \frac{1}{2}(\text{rp}_hi + \text{rp}_lo)^2 + (\text{rp}_hi + \text{rp}_lo)^3 \cdot P_r) \]

by

\[ \text{Rec}^* = \text{ex}_hi \times (1 + \text{rp}_hi + \text{rp}_lo + \frac{1}{2}(\text{rp}_hi)^2 + P_r \times (\text{rp}_hi)^3) + \text{ex}_lo \times (1 + \text{rp}_hi + \frac{1}{2}(\text{rp}_hi)^2) \]

The corresponding error is given by:

\[ \text{Rec} - \text{Rec}^* = (\text{ex}_hi + \text{ex}_lo)(\text{rp}_hi \cdot \text{rp}_lo + \frac{1}{2}(\text{rp}_lo)^2) + \text{ex}_hi \cdot \text{rp}_lo \cdot (3 \cdot (\text{rp}_hi)^2 + 3 \cdot \text{rp}_hi \cdot \text{rp}_lo + \text{rp}_lo^2) + \text{ex}_lo \cdot \text{rp}_hi \cdot (3 \cdot (\text{rp}_lo)^2 + 3 \cdot \text{rp}_hi \cdot \text{rp}_lo + \text{rp}_hi^2) + \text{ex}_lo \cdot \text{rp}_lo \]

\[ \leq 2^{-74} + 2^{-82} + 2^{-88} \]

Hence the following property:

\[ <20> \] The error done when approximating \( \text{Rec} \) by \( \text{Rec}^* \) is:

\[ \text{Rec} = \text{Rec}^* + \varepsilon_{-74} + \varepsilon_{-81} \]

The order in which are executed the instructions is chosen in order to minimize the error. These terms and the intermediate computations with their order of magnitude are given in Figure 1, page 21.

Listing 13: Reconstruction

```c
double R1, R2, R3_hi, R3_lo, R4, R5_hi, R5_lo, R6, R7, R8, R9, R10, R11, crp_hi;
R1 = rp_hi * rp_hi;
R2 = P_r * crp_hi;
Dekker(R3_hi, R3_lo, ex_hi, rp_hi);
R4 = ex_hi * rp_lo;
Dekker(R5_hi, R5_lo, ex_hi, R1);
R6 = R4 + (ex_lo * (R1 + rp_hi));
R7 = ex_hi * R2;
R7 += (R6 + R5_lo) + (R3_lo + ex_lo);
Fast2Sum(R9, R8, R7, R5_hi);
Fast2Sum(R10, tmp, R3_hi, R9);
R8 += tmp;
Fast2Sum(R11, tmp, ex_hi, R10);
R8 += tmp;
Fast2Sum(R11, R8, R11, R8);
```

\[ \diamond \text{ Proof.} \]
|rp_hi| \leq 2^{-9}, therefore:

<21> |R1| \leq 2^{-18},
R1 = (rp_hi)^2.(1 + \varepsilon_{-54}) •

By using property <21> and |rp_hi| \leq 2^{-9} we get:

<22> |crp_hi| \leq 2^{-27},
crp_hi = (rp_hi)^3.(1 + \varepsilon_{-53}) •

This operation is a division by 2, done by subtracting 1 to the exposant. This operation is valid and exact if R1 is not a denormal number, which is the case (property <1> |x| \geq 2^{-54}), and the table 2 showing that we have at most 58 bits of cancellation.

<23> |R1| \leq 2^{-19},
R1 = \frac{1}{2}(rp_hi)^2.(1 + \varepsilon_{-54}) •

By using properties <18> and <22>:

<24> |R2| \leq 2^{-29},
R2 = P_r \times (crp_hi).(1 + \varepsilon_{-54}),
R2 = P_r \times (rp_hi)^3.(1 + \varepsilon_{-52}) •

By using Dekker algorithm (programme 4, page 6) and properties <14> and <15> we have:

<25> |R3_hi| \leq 2^{-10} and |R3_lo| \leq 2^{-64},
R3_hi + R3_lo = ex_hi \times rp_hi exactly •

By using properties <14> and <15>:

<26> |R4| \leq 2^{-64},
R4 = ex_hi \times rp_lo.(1 + \varepsilon_{-54}) •

By using Dekker algorithm and properties <15> and <23> we have:

<27> |R5_hi| \leq 2^{-20} and |R5_lo| \leq 2^{-74},
(R5_hi + R5_lo) = ex_hi \times R1 exactly,
(R5_hi + R5_lo) = (ex_hi + \frac{1}{4}(rp_hi)^2).(1 + \varepsilon_{-54}) •

By using properties <15> and <23>:

|R1 + rp_hi| \leq 2^{-8},
R1 \oplus rp_hi = R1 + rp_hi + \varepsilon_{-62},
|ex_lo \times (R1 + rp_hi)| \leq 2^{-63},
ex_lo \otimes (R1 \oplus rp_hi) = ex_lo \times (\frac{1}{7}(rp_hi)^2 + rp_hi) + \varepsilon_{-116}.

which combined with property <26> gives us:

<28> |R6| \leq 2^{-62},
R6 = R4 + (ex_lo \times (\frac{1}{7}(rp_hi)^2 + rp_hi) + \varepsilon_{-116}) + \varepsilon_{-116}
R6 = (ex_hi \times rp_lo) + (ex_lo \times (\frac{1}{2}(rp_hi)^2 + rp_hi)) + \varepsilon_{-115} + \varepsilon_{-118} •
line 126  By using properties <15> and <24> we get:

\[ |R7| \leq 2^{-30}, \]
\[ R7 = (ex_hi \times R2) \times (1 + \epsilon_{-54}), \]
\[ R7 = (ex_hi \times P_r \times (rp_hi)^3) \times (1 + \epsilon_{-52} + \epsilon_{-53}) \]

line 127  By using properties <27> and <28> we get:

\[ |R6 + R5_lo| \leq 2^{-61}, \]
\[ R6 \oplus R5_lo = R6 + R5_lo + \epsilon_{-115} \text{ ou} \]
\[ R6 \oplus R5_lo = (ex_hi \times rp_lo) + (ex_lo \times (\frac{1}{2}(rp_hi)^2 + rp_hi)) + R5_lo + \epsilon_{-113} \]

line 128  By using properties <15> and <25> we get:

\[ |R3_lo + ex_lo| \leq 2^{-54}, \]
\[ R3_lo \oplus ex_lo = R3_lo + ex_lo + \epsilon_{-108} \]

By using properties <30> and <31> we get:

\[ |R6 + R5_lo + R3_lo + ex_lo| \leq 2^{-53} \text{ et} \]
\[ (R6 \oplus R5_lo) \oplus (R3_lo \oplus ex_lo) = (R6 \oplus R5_lo) + (R3_lo \oplus ex_lo) + \epsilon_{-107}. \]

which combined with property <29> gives us:

\[ |R7| \leq 2^{-29}, \]
\[ R7 = (ex_hi \times P_r \times (rp_hi)^3) \times (1 + \epsilon_{-52} + \epsilon_{-53}) + \]
\[ ((ex_hi \times rp_lo) + (ex_lo \times (\frac{1}{2}(rp_hi)^2 + rp_hi)) + R5_lo + \epsilon_{-113}) + \]
\[ (R3_lo + ex_lo + \epsilon_{-108}) \]
\[ = ex_hi \times (rp_lo + P_r \times (rp_hi)^3) + \]
\[ ex_lo \times (1 + rp_hi + \frac{1}{2}(rp_hi)^2) + \]
\[ R5_lo + R3_lo + \epsilon_{-80} \]

line 129  Fast2Sum algorithm guarantees:

\[ |R9| \leq 2^{-19} \text{ and } |R8| \leq 2^{-73}, \]
\[ R7 + R5_hi = R9 + R8 \text{ exactly} \]

line 131  Fast2Sum algorithm guarantee:

\[ |R10| \leq 2^{-9} \text{ and } |tmp| \leq 2^{-63}, \]
\[ R3_hi + R9 = R10 + tmp \text{ exactly} \]

line 132  By using properties <33> and <34>:

\[ |R8 + tmp| \leq 2^{-62}, \]
\[ R8 = R8 + tmp + \epsilon_{-116} \]

line 134  Fast2Sum algorithm guarantees:

\[ |R11| \leq 2^0 \text{ and } |tmp| \leq 2^{-54}, \]
\[ R11 + tmp = ex_hi + R10 \text{ exactly} \]

line 135  By using properties <35> and <36>:

\[ |R8 + tmp| \leq 2^{-53}, \]
\[ R8 = R8 + tmp + \epsilon_{-107} \]

line 137  Fast2Sum algorithm guarantees:

\[ |R11| \leq 2^1 \text{ and } |R8| \leq 2^{-53}, \]
\[ R11 + R8 = R11 + R8 \text{ exactly} \]
Therefore we have:

\[
R_{11} + R_{8} = R_{11} + \text{tmp} + R_{8} + \varepsilon_{-107} \quad <37>
\]

\[
= \text{ex}_h i + R_{10} + R_{8} + \varepsilon_{-107} \quad <36>
\]

\[
= \text{ex}_h i + R_{10} + \text{tmp} + R_{8} + \varepsilon_{-107} + \varepsilon_{-116} \quad <35>
\]

\[
= \text{ex}_h i + R_{3\_hi} + R_{9} + R_{8} + \varepsilon_{-106} \quad <34>
\]

\[
= \text{ex}_h i + R_{3\_hi} + R_{7} + R_{5\_hi} + \varepsilon_{-106} \quad <33>
\]

\[
= \text{ex}_h i + R_{3\_hi} + R_{5\_hi} + \varepsilon_{-116}
\]

Using property <20> we get:

\[
R_{11} + R_{8} = (\text{ex}_h i + \text{ex}_l o). (1 + r + \frac{1}{2}r^2 + r^3. P_r)
\]

\[
+ \varepsilon_{-73} + \varepsilon_{-78}
\]

By construction of values \((\text{ex}_h i + \text{ex}_l o)\), we have \((\text{ex}_h i + \text{ex}_l o) > \exp(-1) > 2^{-2}\), therefore

\[
|R_{11} + R_{8}| > 2^{-2}
\]

and

\[
\exp(x) = 2^k. (R_{11} + R_{8} + \varepsilon_{-73} + \varepsilon_{-78}). (1 + \varepsilon_{-69}). (1 + \varepsilon_{-109})
\]
Figure 1: The reconstruction step
9 Test if correct rounding is possible

We now have to round correctly the result and multiply it by \(2^k\), with \(k\) an integer. This multiplication is exact, therefore we have:

\[
\exp(x) = o \left(2^k \cdot (R11 + R8 + \epsilon_{-73} + \epsilon_{-78}) \cdot (1 + \epsilon_{-69}) \cdot (1 + \epsilon_{-109}) \right)
\]

\[
= 2^k \cdot o \left((R11 + R8 + \epsilon_{-73} + \epsilon_{-78}) \cdot (1 + \epsilon_{-69}) \cdot (1 + \epsilon_{-109}) \right)
\]

if the result is not a denormal number.

However, if the final result belongs to denormal numbers, then the precision of the result is less than the 53 bits of a normal number. Let us take an example. Let \(a = 1.75\) be a floating-point number exactly representable in double precision format (we have \(o(a) = a\)). Let us multiply this number by \(2^{-1074}\). The exact result is \(1.75 \times 2^{-1074}\) which is different from the result rounded to nearest in double precision \(o(1.75 \times 2^{-1074}) = 2^{-1073}\).

Therefore, in the case when the result is a denormal number, we have to use a special procedure.

9.1 Rounding to nearest

Listing 14: Test if rounding to nearest is possible

```c
static const union { int i[2]; double d; }
#ifdef BIG_ENDIAN
	_two1000 = {0x7E700000, 0x00000000}, /* 1.07150860718626732095e301 */
	_twom1000 = {0x01700000, 0x00000000} , /* 9.33263618503218878990e-302 */
	_errno = {0x3FF00080, 0x00000000} , /* 1.00012207031250000000e0 */
	_twom75 = {0x3B400000, 0x00000000} ; /* 2.64697796196855958e-23 */
#else
...
#endif
#define two1000 _two1000.d
#define twom1000 _twom1000.d
#define errn _errno.d
#define twom75 _twom75.d

int errd = 71303168; /* 68 * 2^20 */

double R11_new, R11_err, R13, st2mem;

union { int i[2]; long long int l; double d; } R12;

/* Résult = (R11 + R8) */
if (R11 == (R11 + R8 * errn))
if (k > -1020)
if (k < 1020)
  HI(R11) += (k<<20);
return R11;

/* we are close to +Inf */
HI(R11) += ((k-1000)<<20);
return R11*two1000;

/* We consider denormal number */
HI(R11_new) = HI(R11)+((k+1000)<<20);
LO(R11_new) = LO(R11);
R12.d = R11_new * twom1000;

HI(st2mem) = R12.i[HI_ENDIAN];
LO(st2mem) = R12.i[LO_ENDIAN];

R11_err -= st2mem * two1000;
HI(R13) = HI(R11_err) & 0x7fffffff;
LO(R13) = LO(R11_err);

if (R13 == two_m75)
  exp_R11 = HI(R11) & 0x7ff00000) - errd;
if ((HI(R8) & 0x7ff00000) < exp_R11)
  /* Difficult rounding ! */
  sn_exp(x);
```
CR-LIBM: The evaluation of the exponential

/* The error term is exactly 1/2 ulp */
if ((HI(R11_err) > 0) && (HI(R8) > 0)) R12.1 +=1;
else
if ((HI(R11_err) < 0) && (HI(R8) < 0)) R12.1 -=1;
}
return R12.d;
}

/* Challenging case */

\[
|R_{11} + R_8| 	imes 2^k - \exp(x) | \leq 2^{-68}
\]

where \( x \in [A, B] \), that leads to \( \text{err}_n = 1 + 2^{-68} \times 2^{55} = 1 + 2^{-13} \). This test is true if we are able to round correctly the result, else we need to call multiprecision procedure.

We are able to round, now we need to perform the multiplication \( R_{11} \times 2^k \) exactly.
We do this multiplication by using integer addition on the exponent of \( R_{11} \). For this operation to be valid and exact, we must be sure not to create a denormal or infinity.
This is the reason why we perform a test on the value of \( k \).

\[
(R_{11} + R_8) > 2^{-2} \text{ then } 2^k.(R_{11} + R_8) \text{ will not lead to a denormal number if } k > -1020.
\]

\[
(R_{11} + R_8) < 2^3 \text{ then } 2^k.(R_{11} + R_8) \text{ will not lead to an overflow if } k < 1020. \text{ Then we have } R_{11} = R_{11} \oplus R_8
\]

In the case when we may return an overflow as result, we make the value of \( k \) smaller, by subtracting 1000 to it. This will prevent the apparition of exception cases during the addition of \( k \) to the exponent. The result is then multiplied by the floating-point number \( twom_{1000} = 2^{-1000} \). This multiplication is exact but in case of underflow, in which case exceptions will be properly raised.

The result may be a denormal number. We need to use a specific test to check whether if we are able to round properly the result.

\[
R_{11} = R_{11} \times 2^{k+1000}
\]

In rounding to nearest we get:

\[
R_{12} = R_{11} \oplus 2^{-1000} = R_{11} \times 2^{-1000} + \varepsilon_{-1075}
\]

The error term \( \varepsilon_{-1075} \) comes from the possible truncation when the result is a denormal number.

Processors do not handle denormal number in ‘hardware’, there are treated when they are stored in memory. It means that number within register can’t be denormal. Therefore, these lines prevent the compiler from performing “dangerous” optimizations, meanwhile preventing to have extra precision by forcing \( R_{12} \) to transit through memory. These lines could be removed for exemple by using gcc and the flag \(-ffloat-store\) that will have the same effect. However this flag forces each floating-point instruction to transit through memory, and has as consequence to severely degrade the performance of the resulting program. The solution to keep good performance is to manually force a data to transit through memory, in order to have an IEEE compliant behavior for denormal numbers.

Proof.

This test is used to know whether we can round the result or not. More details about this trick can be found in [9].

By using property \(<39>\) we have:

\[
|R_{11} + R_8| \times 2^k - \exp(x) | \leq 2^{-68}
\]

where \( x \in [A, B] \), that leads to \( \text{err}_n = 1 + 2^{-68} \times 2^{55} = 1 + 2^{-13} \). This test is true if we are able to round correctly the result, else we need to call multiprecision procedure.

We are able to round, now we need to perform the multiplication \( R_{11} \times 2^k \) exactly.
We do this multiplication by using integer addition on the exponent of \( R_{11} \). For this operation to be valid and exact, we must be sure not to create a denormal or infinity.
This is the reason why we perform a test on the value of \( k \).

\[
(R_{11} + R_8) > 2^{-2} \text{ then } 2^k.(R_{11} + R_8) \text{ will not lead to a denormal number if } k > -1020.
\]

\[
(R_{11} + R_8) < 2^3 \text{ then } 2^k.(R_{11} + R_8) \text{ will not lead to an overflow if } k < 1020. \text{ Then we have } R_{11} = R_{11} \oplus R_8
\]

In the case when we may return an overflow as result, we make the value of \( k \) smaller, by subtracting 1000 to it. This will prevent the apparition of exception cases during the addition of \( k \) to the exponent. The result is then multiplied by the floating-point number \( twom_{1000} = 2^{-1000} \). This multiplication is exact but in case of underflow, in which case exceptions will be properly raised.

The result may be a denormal number. We need to use a specific test to check whether if we are able to round properly the result.

\[
R_{11} = R_{11} \times 2^{k+1000}
\]

In rounding to nearest we get:

\[
R_{12} = R_{11} \oplus 2^{-1000} = R_{11} \times 2^{-1000} + \varepsilon_{-1075}
\]

The error term \( \varepsilon_{-1075} \) comes from the possible truncation when the result is a denormal number.

Processors do not handle denormal number in ‘hardware’, there are treated when they are stored in memory. It means that number within register can’t be denormal. Therefore, these lines prevent the compiler from performing “dangerous” optimizations, meanwhile preventing to have extra precision by forcing \( R_{12} \) to transit through memory. These lines could be removed for exemple by using gcc and the flag \(-ffloat-store\) that will have the same effect. However this flag forces each floating-point instruction to transit through memory, and has as consequence to severely degrade the performance of the resulting program. The solution to keep good performance is to manually force a data to transit through memory, in order to have an IEEE compliant behavior for denormal numbers.
Let

\[ R_{11\_err} = R_{11} \odot R_{12} \odot 2^{1000} \]

By the Sterbenz lemma, and by the fact that a multiplication by a power of 2 is exact we have:

\[ R_{11\_err} = R_{11} - R_{12} \times 2^{1000} \]

This operation will put within \( R_{11\_err} \) the error done during the multiplication of \( R_{11} \) by \( 2^{-1000} \) (line 175).

Remove the sign information of \( R_{11\_err} \)

\[ R_{13} = |R_{11\_err}| \]

Test if \( R_{13} \) is exactly equal to the absolute error (which is also the relative error) done during the rounding process, in rounding to nearest, within denormal number, times \( 2^{1000} \):

\[ 2^{-1075} \times 2^{1000} = 2^{-75} \]

Now we want to prove that if error term is strictly less than \( 1/2ulp(R_{12}) \) then \( R_{12} \) corresponds to the correct rounding of \( R_{11} + R_{8} \).

If \( |R_{11\_err}| < \frac{1}{2}ulp(R_{12}) \) then

\[ |R_{11\_err}| \leq \frac{1}{2}ulp(R_{12}) - ulp(R_{11}) \text{ and } |R_{8}| \leq \frac{1}{2}ulp(R_{11}) \] therefore:

\[ |R_{11\_err} + R_{8}| \leq \frac{1}{2}ulp(R_{12}) - ulp(R_{11}) + \frac{1}{2}ulp(R_{11}) \]

then:

\[ |R_{11\_err} + R_{8}| < \frac{1}{2}ulp(R_{12}) \]

In that case \( R_{12} \) represents the correct rounding of \( R_{11} + R_{8} \) if \( |R_{11\_err}| < \frac{1}{2}ulp(R_{12}) \).

However, if \( |R_{11\_err}| = \frac{1}{2}ulp(R_{12}) = 2^{-75} \), during the multiplication \( R_{11} \times 2^{-1000} \), the result is rounded to odd/even due to the presence of an ambiguous value. It mean that \( R_{12} \) may not represent the rounding to nearest result of \( R_{11} + R_{8} \), we need to perform a correction:

- If \( R_{8} > 0 \), and \( R_{11\_err} = -\frac{1}{2}ulp(R_{12}) \), then the round to odd/even was done on the correct side.
- If \( R_{8} > 0 \), and \( R_{11\_err} = \frac{1}{2}ulp(R_{12}) \), then the round to odd/even wasn’t done on the correct side. We need to perform a correction by adding \( 1ulp \) to \( R_{12} \) (figure 2, case a)
- If \( R_{8} < 0 \), and \( R_{11\_err} = -\frac{1}{2}ulp(R_{12}) \), then the round to odd/even wasn’t done on the correct side. We need to perform a correction by subtracting \( 1ulp \) to \( R_{12} \) (figure 2, case b).
- If \( R_{8} < 0 \), and \( R_{11\_err} = \frac{1}{2}ulp(R_{12}) \), then the round to odd/even was done on the correct side.

When we are in presence of a consecutive sequence of 0 or 1 straddling \( R_{11} \) and \( R_{8} \), then the test done at line 161 will not detect a possible problem. This problem will arise only with denormal numbers, when \( R_{11\_err} \) is close to \( \frac{1}{2}ulp(R_{12}) \).

Therefore we have to detect if in that case (denormal, \( |R_{11\_err}| = \frac{1}{2}ulp(R_{12}) \)) we have enough precision to correctly round the result. We use a test similar to the one used to test whether we can round with rounding toward \( \pm\infty \). Indeed, problematic cases arise when \( \frac{1}{2}ulp(R_{12}) - 2^{-68}R_{11} \leq |R_{11\_err} + R_{8}| \leq \frac{1}{2}ulp(R_{12}) + 2^{-68}R_{11} \).
CR-LIBM: The evaluation of the exponential

Test if we are in presence of one of the cases described previously, and correct the result by adding or subtracting 1ulp by using the “continuity” of the representation of floating-point number.

\[ R8 > 0 \quad \text{et} \quad R11_{\text{err}} = \frac{1}{2} \text{ulp}(R12) \]

\[ R8 < 0 \quad \text{et} \quad R11_{\text{err}} = -\frac{1}{2} \text{ulp}(R12) \]

Figure 2: Description of problem with rounding to nearest of a denormal number.

### 9.2 Rounding toward \( +\infty \)

Listing 15: Test if rounding toward \( +\infty \) is possible

```c
static const union {int i[2]; double d;}
#else BIG_ENDIAN
_two1000 = {0x7E700000, 0x00000000}, /* 1.07150860718626732095 e301 */
_twom1000 = {0x01700000, 0x00000000}, /* 9.33263618503218878990 e-302 */
#endif
#define two1000 _two1000.d
#define twom1000 _twom1000.d
int errd = 71303168; /* 68 * 2^20 */
union {int i[2]; long long int l; double d;} R12;
/* Result = (R11 + R8) */
exp_R11 = (HI(R11) & 0x7ff00000) - errd;
if ((HI(R8) & 0x7ff00000) > exp_R11) {
  /* We are able to round the result */
  if (k > -1020) {
    HI(R11) += (k<<20);
  } else {
    HI(R11) += ((k+1000)<<20);
    R11 += two1000;
  }
  if (HI(R8) > 0) {
    R12.d = R11;
    R12.l += 1;
    R11 = R12.d;
  }
  return R11;
} else {
  /* We are with denormal number */
  HI(R11) += ((k+1000)<<20);
  R12.d = R11 + twom1000;
  HI(st2mem) = R12.i[HIG_ENDIAN];
  LO(st2mem) = R12.i[LO_ENDIAN]
  R11 = st2mem + two1000;
```
if ((HI(R11) > 0) || ((HI(R11) == 0) && (HI(R8) > 0))) R12.l += 1;
return R12.d;
}
else {
/* Difficult case */
su_exp(x);
}

\begin{proof}
The program used to check whether correct rounding toward $+\infty$ is possible is similar to the one used with rounding to nearest.

- **Line 221**: This test is valid even if the final result is a denormal number.
- **Line 231**: We add 1ulp to the result if $R8$ is positive.
- **Line 245**: Like for rounding to nearest, the quantity $R11$ represents the rounding error that comes from the operation $R11 \times \text{twom}1000$ in line 240.
- **Line 246**: This test checks whether the error from line 240 is strictly positive or if it is equal to zero and if $R8$ is strictly positive. If we are in one of these two cases, by definition of rounding toward $+\infty$, we need to add 1ulp to the result.

\end{proof}

\section{Rounding toward $-\infty$}

Listing 16: Test if rounding toward $-\infty$ is possible

\begin{verbatim}
static const union{int i[2]; double d;}
#ifdef BIG_ENDIAN
	_two1000 = {0x7E700000, 0x00000000}, /* 1.07150860718626732095 e301 */
	_twom1000 = {0x01700000, 0x00000000}, /* 9.33263618503218878990 e-302 */
#else
#endif
#define two1000 _two1000.d
#define twom1000 _twom1000.d
int errd = 71303168; /* 68 * 2^20 */
int exp_R11;
union {int i[2]; long long int l; double d} R12;
/* Result = (R11 + R8) */
exp_R11 = (HI(R11) & 0x7ff00000) - errd;
if ((HI(R8) & 0x7ff00000) > exp_R11) {
/* We are able to round the result */
if (k > -1020) {
if (k < -1020) {
HI(R11) += (k<<20);
} else {
/* We are close to +Inf */
HI(R11) += ((k-1000)<<20);
R11 *= twom1000;
}
} else if (HI(R8) > 0) {
R12.d = R11;
R12.l += 1;
R11 = R12.d;
} return R11;
} else {
/* We are with denormal number */
HI(R11) += ((k+1000)<<20);
R12.d = R11 * twom1000;
HI(st2mem) = R12.i[HII_ENDIAN];
LO(st2mem) = R12.i[LO_ENDIAN];
}
\end{verbatim}
\begin{verbatim}
R11 = st2mem * two1000;
if ((HI(R11) < 0) || ((HI(R11) == 0) && (HI(R8) < 0))) R12.1 = 1;
return R12.d;
else {
  /* Difficult case */
su_exp(x);
}
\end{verbatim}

\begin{itemize}
  \item \textbf{Proof.}
  The program used to check whether correct rounding toward \(-\infty\) is possible is similar to the previous one.
\end{itemize}

9.4 \textbf{Rounding toward 0}

The program used to check whether correct rounding toward 0 is possible is identical to the one used for rounding to \(-\infty\) because \(\exp(x)\) is a positive function.

10 \textbf{Accurate phase}

When the previous computation failed, it means that the rounding of the result is difficult to decide. We need to use more accurate methods:

\begin{itemize}
  \item \textit{su\_exp} with rounding to nearest,
  \item \textit{su\_exp} with rounding toward \(+\infty\),
  \item \textit{sd\_exp} with rounding toward \(-\infty\),
\end{itemize}

These methods are based on SCS library \cite{5}, with 30 bits of precision per digit and 8 digits per vector. The guaranteed precision with this format is 211 bits at least. Even if there is no proof for these operators yet, the proof for correct rounding of the exponential only requires the following properties, which are easy to check and/or satisfy:

\begin{property} \textbf{(Addition)} \label{prop:addition}
Let \(a \oplus b\) represent the multiprecision operation performing an addition between \(a\) and \(b\) with at least 210 bits of precision for the result. Like for double precision floating point number, the SCS addition may lead to a cancellation. We have:
\[
  a + b = (a \oplus b).(1 + \varepsilon_{211})
\]
\end{property}

\begin{property} \textbf{(Multiplication)} \label{prop:multiplication}
Let \(a \oslash b\) represent the multiprecision operation performing a multiplication between \(a\) and \(b\) with at least 210 bits of precision for the result. This operation does not produce a cancellation.
\[
  a \times b = (a \oslash b).(1 + \varepsilon_{211})
\]
\end{property}

10.1 \textbf{Overview of the algorithm}

Here is the algorithm used for the second part of the evaluation:

1. \textbf{No special case handling}
   Special cases have been handled by the first part.

2. \textbf{Range reduction}
   We compute the reduced argument \(r\) and the integer \(k\) such that:
\[
  r = \frac{x - k \cdot \ln(2)}{512}
\]
with \(-\frac{\ln(2)}{1024} \leq r \leq -\frac{\ln(2)}{1024}\)

such that

\[\exp(x) = \exp(r)^{512} \times 2^k\]

3. Polynomial evaluation

We compute the polynomial \(P(r)\) of degree 11:

\[\exp(r) = (1 + r + P(r)).(1 + \varepsilon_{-179})\]

4. Powering the result

\[\exp(r)^{512} = \left(\left(\left(\left(\left(\left(\left(\left(\left(\left(\left(\left(\left(\exp(r)^2\right)^2\right)^2\right)^2\right)^2\right)^2\right)^2\right)^2\right)^2\right)^2\right)^2\right)^2\right)^2\]

5. Reconstruction

\[\exp(x) = \exp(r)^{512} \times 2^k.(1 + \varepsilon)\]

with \(|\varepsilon| \leq 2^{-170}\)

We have chosen this evaluation scheme, because the reconstruction step uses the squaring multiprecision operator. This operator facilitates the error computation and is very economic: its cost is 0.7 times that of a true multiprecision multiplication.

We will notice that there exists an alternative to the squaring solution. We can tabulate values \(2^{\pi N}\) for \(N = 1, 2, \ldots, 511\) and use the formula \(\exp(x) = \exp(r) \times 2^N \times 2^{\pi N}\) with \(k = M + N/512\). However we prefer the squaring method that do not request the storage of SCS numbers and the associated quantity of memory.

10.2 Function calls

With Lefèvre worst cases on table makers dilemma we get the following theorem:

**Theorem 6** (Correct rounding for the exponential) Let \(y\) be the exact value of the exponential of a floating-point number in double precision \(x\). Let \(y^*\) be an approximation of \(y\) such that the distance between \(y\) and \(y^*\) be bounded by \(\varepsilon\). Then if \(\varepsilon \leq 2^{-157}\), for each of the four rounding mode, rounding \(y^*\) is equivalent to rounding \(y\);

To round the multiprecision result in SCS format depending on the rounding mode, we use the following procedure \((\text{scs\_get\_d}, \text{scs\_get\_d\_pinf}, \text{scs\_get\_d\_minf})\).

10.2.1 Rounding to nearest

Listing 17: Compute the rounding to nearest of the exponential in multiprecision

```c
double sn_exp(double x) {
    scs_t res_scs;
    scs_db_number res;
    exp_SC(res_scs, x);
    scs_get_d(&res.d, res_scs); res.d = x;
    return res.d;
}
```
10.2.2 Rounding toward $+\infty$

Listing 18: Compute the rounding toward $+\infty$ of the exponential in multiprecision

```c
double su_exp(double x){
    scs_t res_scs;
    scs_db_number res;
    exp_SC(res_scs, x);
    scs_get_d_pinf(&res.d, res_scs);
    return res.d;
}
```

10.2.3 Rounding toward $-\infty$

Listing 19: Compute the rounding toward $-\infty$ of the exponential in multiprecision

```c
double sd_exp(double x){
    scs_t res_scs;
    scs_db_number res;
    exp_SC(res_scs, x);
    scs_get_d_minf(&res.d, res_scs);
    return res.d;
}
```

10.3 Software

The function $\text{exp\_SC}$ approximate the exponential of $x$ with 170 bits of precision and put the result in $\text{res\_scs}$.

Listing 20: Compute the exponential in multiprecision

```c
void exp_SC(scs_ptr res_scs, double x){
    scs_t sc1, red;
    scs_db_number db;
    int i, k;

    // 1) Range reduction */
    db.d = x/512 (= 2^9) */
    db.i[HI_ENDIAN] -= (9 << 20);
    scs_set_d(sc1, db.d);

    DOUBLE2INT(k, (db.d * iln2_o512.d));

    // 2) Polynomial evaluation */
    scs_mul_ui(red, (unsigned int) k);
    scs_mul_ui(red_low, (unsigned int) k);

    if (k>0){
        scs_mul_ui(red, (unsigned int)(-k));
        scs_mul_ui(red_low, (unsigned int)(-k));
        red->sign *= -1;
        red_low->sign *= -1;
    }
    scs_sub(red, sc1, red);
    scs_sub(red, red, red_low);

    scs_mul(res_scs, constant_poly_ptr[0], red);
    for(i=1; i<11; i++){
        scs_add(res_scs, constant_poly_ptr[i], res_scs);
        scs_mul(res_scs, red, res_scs);
    }
```
```c
scs_add(res_scs, SCS_ONE, res_scs);
scs_mul(res_scs, red, res_scs);
scs_add(res_scs, SCS_ONE, res_scs);

/* 3) Powering the result exp(r)^512 */
for (i = 0; i < 9; i++) {
    scs_square(res_scs, res_scs);
}

/* 4) Multiplication by 2^k */
res_scs->index += (int)(k/30);
if ((k%30) > 0)
    scs_mul_ui(res_scs, (unsigned int)(1LL<<(k%30)));
else if ((k%30) < 0) {
    res_scs->index --;
    scs_mul_ui(res_scs, (unsigned int)(1LL<<(30+(k%30))));
}
}
```
By the properties <40> and <42> we have:

\[ red = \frac{x}{512} \times \left( k \times \frac{\ln 2}{512} \right) \left( 1 + \varepsilon_{-411} \right) \]

In addition we have seen in the quick phase that at most 58 bits could be cancelled during this subtraction.

\[ |red| \leq \frac{\ln 2}{1024} \leq 2^{-10}, \]
\[ red = \frac{x}{512} - k \times \frac{\ln 2}{512} \varepsilon_{-210} \]
We now perform the polynomial evaluation where the coefficient have the following properties.

\[
\begin{align*}
|\text{constant}_\text{poly}_\text{ptr}[0]| &= c_0 \leq 2^{-28}, \\
|\text{constant}_\text{poly}_\text{ptr}[1]| &= c_1 \leq 2^{-25}, \\
|\text{constant}_\text{poly}_\text{ptr}[2]| &= c_2 \leq 2^{-21}, \\
|\text{constant}_\text{poly}_\text{ptr}[3]| &= c_3 \leq 2^{-18}, \\
|\text{constant}_\text{poly}_\text{ptr}[4]| &= c_4 \leq 2^{-15}, \\
|\text{constant}_\text{poly}_\text{ptr}[5]| &= c_5 \leq 2^{-12}, \\
|\text{constant}_\text{poly}_\text{ptr}[6]| &= c_6 \leq 2^{-9}, \\
|\text{constant}_\text{poly}_\text{ptr}[7]| &= c_7 \leq 2^{-6}, \\
|\text{constant}_\text{poly}_\text{ptr}[8]| &= c_8 \leq 2^{-4}, \\
|\text{constant}_\text{poly}_\text{ptr}[9]| &= c_9 \leq 2^{-2}, \\
|\text{constant}_\text{poly}_\text{ptr}[10]| &= c_{10} \leq 2^{-1}.
\end{align*}
\]

We have:

1. \(P_0 = c_1 \oplus (\text{red} \boxtimes c_0)\) therefore \(|P_0| \leq 2^{-24}\) et \(P_0 = (c_1 + (\text{red} \times c_0))(1 + \varepsilon - 210 + \varepsilon - 244)\)

2. \(P_1 = c_2 \oplus (\text{red} \boxtimes P_0)\) therefore \(|P_1| \leq 2^{-20}\) et \(P_1 = (c_2 + (\text{red} \times P_0))(1 + \varepsilon - 210 + \varepsilon - 240)\)

3. \(P_2 = c_3 \oplus (\text{red} \boxtimes P_1)\) therefore \(|P_2| \leq 2^{-17}\) et \(P_2 = (c_3 + (\text{red} \times P_1))(1 + \varepsilon - 210 + \varepsilon - 236)\)

4. \(P_3 = c_4 \oplus (\text{red} \boxtimes P_2)\) therefore \(|P_3| \leq 2^{-14}\) et \(P_3 = (c_4 + (\text{red} \times P_2))(1 + \varepsilon - 210 + \varepsilon - 233)\)

5. \(P_4 = c_5 \oplus (\text{red} \boxtimes P_3)\) therefore \(|P_4| \leq 2^{-11}\) et \(P_4 = (c_5 + (\text{red} \times P_3))(1 + \varepsilon - 210 + \varepsilon - 230)\)

6. \(P_5 = c_6 \oplus (\text{red} \boxtimes P_4)\) therefore \(|P_5| \leq 2^{-8}\) et \(P_5 = (c_6 + (\text{red} \times P_4))(1 + \varepsilon - 210 + \varepsilon - 228)\)

7. \(P_6 = c_7 \oplus (\text{red} \boxtimes P_5)\) therefore \(|P_6| \leq 2^{-5}\) et \(P_6 = (c_7 + (\text{red} \times P_5))(1 + \varepsilon - 210 + \varepsilon - 224)\)

8. \(P_7 = c_8 \oplus (\text{red} \boxtimes P_6)\) therefore \(|P_7| \leq 2^{-3}\) et \(P_7 = (c_8 + (\text{red} \times P_6))(1 + \varepsilon - 210 + \varepsilon - 221)\)

9. \(P_8 = c_9 \oplus (\text{red} \boxtimes P_7)\) therefore \(|P_8| \leq 2^{-1}\) et \(P_8 = (c_9 + (\text{red} \times P_7))(1 + \varepsilon - 210 + \varepsilon - 219)\)

10. \(P_9 = c_{10} \oplus (\text{red} \boxtimes P_8)\) therefore \(|P_9| \leq 1 + 2^{-10}\) et \(P_9 = (c_{10} + (\text{red} \times P_8))(1 + \varepsilon - 210 + \varepsilon - 217)\)

11. \(P_{10} = 1 \oplus (\text{red} \boxtimes P_9)\) therefore \(|P_{10}| \leq 1 + 2^{-9}\) et \(P_{10} = (1 + (\text{red} \times P_9))(1 + \varepsilon - 210 + \varepsilon - 216)\)

12. \(P_{11} = 1 \oplus (\text{red} \boxtimes P_{10})\) therefore \(|P_{11}| \leq 1 + 2^{-8}\) et \(P_{11} = (1 + (\text{red} \times P_{10}))(1 + \varepsilon - 210 + \varepsilon - 216)\)

Therefore

\[
\text{res}_{\text{scs}} = P_{11}(1 + \varepsilon - 209)
\]

We build the polynomial such that

\[
\exp(r) = (1 + r + c_{10}, r^2 + \cdots + c_0, r^{12})(1 + \varepsilon - 179)
\]

Therefore

\[
|\text{res}_{\text{scs}}| \leq 1 + 2^{-8} \text{ with }
\exp(x) = 2^{k} (\exp(r))^{512} = 2^{k} (\text{res}_{\text{scs}}(1 + \varepsilon - 209), (1 + \varepsilon - 179))^{512}
\]
We perform a squaring of the result 9 times, which corresponds to raising the result to the power 512. At each iteration we perform a rounding error equal to $\varepsilon_{-207}$. Finally

$$|\text{res}_{\text{scs}}| \leq 2^{3} \text{ and } \exp(x) = 2^{k}.\exp(r)^{512}.(1 + \varepsilon_{-170})$$

With these lines we perform the multiplication of $\text{res}_{\text{scs}}$ by $2^{k}$. This multiplication is done by a shift on the index of $k/30$, where 30 correspond to the number of bits used within a multiprecision number. This shift is exact. Then a multiplication of $\text{res}_{\text{scs}}$ by $2$ to the power the rest of the euclidian division of $k$ by 30 is done. At the end of these instructions we have:

$$\exp(x) = (\text{res}_{\text{scs}}).(1 + \varepsilon_{-170})$$

We are approximating the exponential with a relative error less than $2^{(-170)}$. This result combines with property 1, gives us the proof of correct rounding for the four rounding modes.

\[\square\]

## 11 Analysis of the exponential

### 11.1 Test conditions

Table 9 lists the combinations of processor, OS and default libm used for our tests.

<table>
<thead>
<tr>
<th>Processor</th>
<th>OS</th>
<th>compiler</th>
<th>default libm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pentium III</td>
<td>Debian GNU/Linux</td>
<td>gcc-2.95</td>
<td>glibc, derived from fdlibm</td>
</tr>
<tr>
<td>UltraSPARC III</td>
<td>SunOS 5.8</td>
<td>gcc-2.95</td>
<td>Sun optimized</td>
</tr>
<tr>
<td>Xeon (Pentium 4)</td>
<td>Debian GNU/Linux</td>
<td>gcc-2.95</td>
<td>glibc, derived from fdlibm</td>
</tr>
<tr>
<td>PowerPC G4</td>
<td>MacOS 10.2</td>
<td>gcc-2.95</td>
<td>Apple specific</td>
</tr>
<tr>
<td>Itanium</td>
<td>Debian GNU/Linux</td>
<td>gcc-2.95, gcc-3.2</td>
<td>Intel optimized</td>
</tr>
</tbody>
</table>

Table 9: The systems tested

The following presents tests performed under such conditions as to suppress most of the impact of the memory hierarchy: A small loops performs 10 identical calls to the function, and the minimum timing is reported, ensuring that both code and data have been loaded in the cache, and that interruptions by the operating system do not alter the timings.

These timings are taken on random values between $-745$ and $+744$, which is the practical range for the exponential. We also report the timing for the worse case for the correct rounding in rounding to nearest mode of the exponential, which is $x = 7.5417527749959590085206221e - 10$.

Our library was tuned to take into account the adequation of the evaluation scheme to the memory hierarchies of current processors (our program for the exponential evaluation uses 2.8Kbytes of table for the four rounding modes, whereas the one from fdlibm use 13Kbytes). However we do not have tested the impact over performance of this concern and is part of our futur works.

### 11.2 Results

Tables 10 gives a summary of the timings of the various libraries. The timings are normalized to the average time of the default libm on each system, which should be taken with care as the libms used by default on the tested systems are all different. Table 11 gives absolute times.
Table 10: Accuracy and timings for the exponential function from various libraries. Timings are normalized to the average time for the default \texttt{libm}. For each processor, the first line gives the average time, and the second line gives the worst-case time.

<table>
<thead>
<tr>
<th>Processor</th>
<th>\texttt{default libm}</th>
<th>correctlib libraries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pentium III</td>
<td># errors 1/587</td>
<td>time 1.21\hspace{1em}</td>
</tr>
<tr>
<td>Itanium</td>
<td># errors 1/491</td>
<td>time 13.7\hspace{1em}</td>
</tr>
<tr>
<td>UltraSparc Ili</td>
<td># errors 1/41</td>
<td>time 12.9\hspace{1em}</td>
</tr>
<tr>
<td>PowerPC G4</td>
<td># errors 1/4739</td>
<td>time 1.32\hspace{1em}</td>
</tr>
</tbody>
</table>

Table 11: Absolute timings for the exponential

<table>
<thead>
<tr>
<th>Processor</th>
<th>libm</th>
<th>crlibm</th>
<th>MPFR</th>
<th>libultim</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pentium III (cycles)</td>
<td>average time</td>
<td>462</td>
<td>562</td>
<td>21114</td>
</tr>
<tr>
<td>Overall worst-case time</td>
<td>837</td>
<td>18413</td>
<td>43316</td>
<td>66050</td>
</tr>
<tr>
<td>Correct rounding worst case time</td>
<td>448</td>
<td>15963</td>
<td>38969</td>
<td>154215</td>
</tr>
<tr>
<td>Xeon (cycles)</td>
<td>average time</td>
<td>982</td>
<td>1178</td>
<td>24991</td>
</tr>
<tr>
<td>Overall worst-case time</td>
<td>3124</td>
<td>34236</td>
<td>138768</td>
<td>108920</td>
</tr>
<tr>
<td>Correct rounding worst case time</td>
<td>1080</td>
<td>30384</td>
<td>50436</td>
<td>2592660</td>
</tr>
<tr>
<td>Itanium (cycles)</td>
<td>average time</td>
<td>202</td>
<td>518</td>
<td>30995</td>
</tr>
<tr>
<td>Overall worst-case time</td>
<td>2767</td>
<td>9349</td>
<td>70660</td>
<td>139415</td>
</tr>
<tr>
<td>Correct rounding worst case time</td>
<td>131</td>
<td>6434</td>
<td>33388</td>
<td>2928718</td>
</tr>
<tr>
<td>UltraSparc Ili (cycles)</td>
<td>average time</td>
<td>762</td>
<td>2033</td>
<td>63570</td>
</tr>
<tr>
<td>Overall worst-case time</td>
<td>9823</td>
<td>112366</td>
<td>292129</td>
<td>157987</td>
</tr>
<tr>
<td>Correct rounding worst case time</td>
<td>292</td>
<td>91190</td>
<td>126827</td>
<td>2724912</td>
</tr>
<tr>
<td>PowerPC G4 (ms)</td>
<td>average time</td>
<td>2.27</td>
<td>2.07</td>
<td>30.7</td>
</tr>
<tr>
<td>Overall worst-case time</td>
<td>3</td>
<td>21</td>
<td>61</td>
<td>116</td>
</tr>
<tr>
<td>Correct rounding worst case time</td>
<td>2</td>
<td>20</td>
<td>59</td>
<td>3354</td>
</tr>
</tbody>
</table>

11.3 Analysis

Processor-specific libraries

Documentation[15] from Intel labs claim to provide an exponential in only 48 cycles. This performance is possible through the wide use of non portable tricks such as inverse approximation, fused multiply and add and double extended precision. However, our tests show that the environmental cost (mainly the cost of a function call) is about 80 clock cycles! Our tests have also shown that there exists a slower path that takes up to 2787 clock cycles, which is 14 times slower. This path seems to be taken very often since the average cost is 1.5 times more expensive than the smallest execution time.

The same conclusion can be done for the mathematical library used on Ultra-SPARC Ili system, where there exists a path 13 times slower than a normal execution.

These two observations show that our two-step procedure, with a much slower second step, could
be viable in the commercial world. The mathematical library used on the PowerPC G4 with gcc is the one from Apple. This library do not provide correct rounding and is 1.1 slower than the version provided with crlibm. It is, however, the most accurate of the tested libraries.

**The cost of correct rounding**

The libutlim library provides correct rounding for an average cost between 0.9 (on a Pentium III) and 1.8 (on an Itanium) times the cost of the standard library. Our exponential return a result for an average cost between 0.91 (on Power-PC) and 2.66 (on Ultra-SPARC III) times the cost of the standard library, which is reasonable. On the other hand, MPFR provide correct rounding for an average cost between 13.5 (on Power-PC) and 153 (on Itanium) compared to the libm.

The main advantage of crlibm over libutlim is the upper bound on the execution time. On our tests, this bound for crlibm is 147 times the average libm cost, whereas for libtultim this bound goes up to 14499 times the average libm cost. Our two steps strategy fully benefits from knowing bounds on correct rounding worst cases.

We notice that our second step is in average 3 times faster than the multiprecision library MPFR. It shows that our multiprecision operators from scslib, hand tuned for 200 bits of precision perfectly fulfill the performance requirement of the second step.

**Relations between the two steps of crlibm**

For some of our test, we have disconnected the call to the second phase in the evaluation scheme and counted one miss-rounding result over 2097152 (≈ 2^{21}). As we can see in the first column of table 10, it mean that we are far more precise than others library.

Our second phase is 30 times slower than the first step and is called only once over 2^{13}. The cost of the second step over the average cost is:

$$\frac{1 \times (2^{13} - 1) + 30 \times 1}{2^{13}} = 1.003540039$$

which corresponds to a 0.35% overhead. This small overhead in average means that a possible performance improvement is to reduce the precision of the first step, and by the same way the number of instructions, to increase to number of time that the second step is called. It will also make the proof simpler.

**12 CONCLUSION AND PERSPECTIVES**

We have presented a library of elementary functions correctly rounded in double precision in the four IEEE-754 rounding modes. Although only one function is complete, we have thus shown that correct rounding can be obtained with performance (both average and worst-case) comparable to libraries without this property. Improvements over previous works include

- proven correct rounding, thanks to recent theoretical results, with detailed proofs of the code published along the code itself,
- availability of the directed rounding modes,
- bounded worst-case performance acceptable for real-time applications.

Future work include, of course, tuning and completing the library. It is obvious from our performance measurements that our first step is too accurate and too slow for a balanced average time. We will take this experience into consideration when writing first steps for other functions. The IBM library seems to get a better balance although its second and later steps are much slower. Its code, unfortunately, is little documented and difficult to prove.

Writing the proofs is a very time-consuming task, which could be partially automated for one step which is common to most function: The accumulation of error terms in order to compute the final error.

We hope that this work is a step towards making correctly rounded elementary functions a standard.
References

[3] MPFR, the Multiprecision Precision Floating-Point Reliable library.


