Berkeley Elementary Functions Test Suite

Zhishun Alex Liu
under direction of
Professor William Kahan

December 30, 1988

Abstract

A suite of programs is presented to test how accurately the elementary transcendental functions exp, log, sin, cos and atan have been implemented in a computer's run-time library. The suite is written in the language C and designed to run on any computer with binary floating-point arithmetic rounded in a reasonable way. The suite makes no appeal to extra-precise arithmetic; the tests use only whatever arithmetic capabilities are present in the environment where the transcendental functions are to be used. Despite this limitation, the tests run fast and deliver indication of accuracy to within a small fraction of an ULP (**U**nit in the **L**ast **P**lace) of the functions under test. This account includes the proofs of the test suite's claims to accuracy.

Part I

User's Guide to Berkeley Elementary Functions Test Suite

1 Introduction

Subprograms for elementary functions like EXP and COS are often the basic building blocks of a wide variety of applications and as such ought to be fast and accurate. How can we discover how fast and accurate they are?

The speed of, say COS(·) can be tested easily by seeing how long it takes to compute

$$\texttt{COS(T(1))}, \texttt{COS(T(2))}, \ldots, \texttt{COS(T(N))}$$

for some previously generated array of random test arguments T(1), T(2), ... , T(N) with a sufficiently large N, perhaps N = 10000. Accuracy can be tested too by comparing each COS(T(i)) with a more accurate DCOS(DBLE(T(i))), where DCOS is the double-precision analog of the single precision program COS. But how shall the accuracy of DCOS be tested in an environment that lacks any support for floating-point arithmetic more precise than double-precision? How can we test COS in an environment that lacks DCOS? That is the kind of question we answer below. In Section 2 we shall describe the design goals of our test suite, along with a discussion of a number of existing test programs attempting to achieve similar goals.

For an environment that supplies COS but not DCOS, we have found an economical way to compute a more accurate version of COS without going to the lengths of a full-scale simulation of higher-precision arithmetic. Our approach relies upon tricky formulas to compute a good estimate for the error in COS without ever explicitly computing a more accurate version of COS. These tricks permit our tests to run much faster than if we merely simulated higher-precision arithmetic, and therefore our tests can explore the accuracy of COS more thoroughly in a given amount of time. But we must pay a price to use those tricks; they work only in a somewhat restricted domain. Here are the restrictions:

First, only the functions cos, sin, atan, log, exp, log1p and expm1 are tested; except for atan, log and exp, only over ranges of test arguments restricted to the most important parts of their domains.

Second, a C compiler *supporting the desired floating-point data type* must be available on the target machine. In addition, it must fulfill a number of requirements pertaining to floating-point code generation that are perceived as among the most basic functionalities that *any* commercially significant C compiler should provide. In Section 3 these requirements are enumerated, along with instances of C compilers failing them; we then present the few simple steps our user must follow in order to be able to successfully run our tests.

Third, the tests are valid only on computers whose floating-point arithmetic is

- binary (not decimal nor hexadecimal) with at least 24 significant bits in its mantissa, and
- rounded in a reasonable way, by *reasonable* we mean the 4 basic floating-point operations $+, -, *, /$ shall produce *correctly-rounded* results except in the face of underflow or overflow; in other words, rounding errors committed by these operations shall not exceed 1/2 ULP.

Therefore the tests are valid on machines like these:

a) DEC VAX, HP 3000, HP Precision architecture,

b)
   - IBM PC/XT/AT/RT with floating-point coprocessors,
   - Sun-3 and Sun-4,
   - ELXSI 6400,
   - Apple II and Macintosh series using its SANE arithmetic,
   - ... other machines conforming to IEEE standard 754-1985 for binary floating-point arithmetic using chips like Intel's 8087, Motorola's 68881, Weitek's 1164/65 etc.

But the tests are *not* intended to run on

c) CRAYs, CDC Cybers, UNIVAC 11xx, ... (strange rounding),

d) IBM/370 and clones (hexadecimal arithmetic),

e) HP 80 series, and calculators (decimal arithmetic).

In Section 4 we shall describe results obtained by running our tests on various machines in categories (a) and (b), following a description of the output format our test suite actually produces.

Once the algorithms are designed and analyzed, implementation is always the most straightforward next step. Then comes the question of *testing* the test programs implemented. In Section 5 we shall describe the methods we use to *test* the tests.

Roughly speaking, for each approximator we employ a table-lookup technique with a table of precomputed accurate values whenever the technique seems feasible. The larger the table, the more accurate the resulting approximator will be. However, large tables occupy a significant amount of memory at run-time, and when there is a need to transmit our source code over primitive communication links, files containing *seemingly* random digits tend to require more sophisticated data transmission protocols to ensure data integrity. Taking these factors into account, we have decided to limit the size of the tables. Consequently we demand an accuracy no better than 1/16 ULPs in the design of all our algorithms. Actually, all of our approximators presently implemented can do better than 1/16 ULPs. In Section 6 the idea behind the overall design of these approximators will be illustrated in detail using `ATAN` as an example, situations under which our test suite will be most indispensable are presented, then follows an outline of the algorithms we use.

Section 7 through 11 shall be dedicated to describing in detail the algorithms used for accurately measuring the errors committed by expm1(x), exp(x), sin(x) & cos(x), log(x) and atan(x) respectively over the intervals covered. log1p(x) uses the same algorithm as the one devised for log(x), we may thus safely skip over it. Section 12 through 16 shall deliver precise accuracy statements with respect to to the above algorithms, complete with detailed proofs of the respective accuracy claims.

2 What does Our Test Suite Do?

2.1 Design Goals of Our Test Suite

to be filled in

2.2 Other Existing Useful Test Programs

2.2.1 Brent's MP Package [1]

Brent's MP package is capable of evaluating many elementary functions to any desired precision, using *only* integer arithmetic. To evaluate $\text{atan}(x)$ to 20 significant decimal digits, relative to the 4.3BSD implementation of $\text{atan}(x)$, MP $\text{atan}(x)$ is about 1000 times slower. On a VAX 11/750 with FPA, a run with $2,500,000$ random arguments to measure the accuracy of $\text{atan}(x)$ using Brent's MP package would take more than 200 CPU hours to complete. Extensive random argument tests using Brent's MP package become impractical. In contrast, those 2.5 million tests take 50 CPU minutes using our methods.

2.2.2 Cody and Waite's ELEFUNT Test Suite [2]

The ELEFUNT test suite as developed by Cody and Waite is written in FORTRAN and covers the usual assortment of algebraic, trigonometric and transcendental functions. The method they use involves measuring the error in some carefully selected mathematical identities over certain intervals. For intervals in the neighborhood of 0, a suitably truncated Taylor series is used to approximate the function under test and a random argument test is performed. The test suite provides a good indication of the numerical reliability of the functions under test. However, it does not provide a *direct* measurement of the numerical errors incurred.

2.2.3 IMSL's Elementary Functions Test [3]

to be filled in

2.2.4 Peter Tang's Test Programs in Ada [5]

to be filled in

2.2.5 W. Kahan's Floating-point Arithmetic Diagnostic Program "PARANOIA"

to be filled in

2.2.6 K. C. Ng's Exceptional/Boundary Cases Test Vector

to be filled in

3 How do You Use Our Test Suite?

3.1 What must Your C Compiler Provide?

A C compiler *supporting the desired floating-point data type* must be available on your machine which

1. generates *correct* code for all expressions involving the desired floating-point data type,
2. converts *exactly* such modest-sized integers as $2^{21} - 1$ to the desired floating-point data type,
3. allows the declaration of an *array* of numbers of the desired floating-point data type,
4. runs and passes Kahan's PARANOIA, a floating-point arithmetic diagnostic program, with no indication of anomalous rounding,
5. always performs *destructive store*; that is to say, if variables are assigned to higher precision floating-point registers, and intermediate floating-point operations are performed in that higher precision, then upon encountering each source code *assignment* statement, the result must be rounded back to the precision of the floating-point variable to which a value is assigned,
6. inhibits *rewrite* of such floating-point expressions as $(a-b)-c$ gratuitously into $a-(b+c)$.

3.2 Do All C Compilers Qualify?

Requirements 1, 2, 3 and 4 are among the most basic functionalities that *any* commercially significant C compiler should provide. However, the initial release of Borland International's Turbo C 1.0 failed requirements 1 *and* 2: the compile-time floating-point division reversed its divisor and dividend and the floating-point constant 11.0 didn't get converted at compile-time *exactly* to 11. A version of the Zilog Z8000 C compiler supported IEEE-Extended only in the form of "`register double`" while disallowing declarations of *arrays* of type "`register double`", thus failing requirement 3. Borland International's Turbo C 2.0 failed requirement 4 if software floating-point emulator is in effect: none of the 4 $+, -, *, /$ floating-point operations delivered correctly-rounded results. Requirement 5 can generally be met by compiling our code with appropriate compile-time flags. For instance, the "`-ffloat-store`" command-line flag in GNU C will force the compiler not to assign floating-point variables to floating-point registers on Sun-3s equipped with an MC68881 floating-point coprocessor, thereby alleviating the problem of unwanted "excess" precision. Requirement 6 is satisfied by most C compilers except a few super-intelligent optimizing compilers such as the MIPS C compiler. One may have to inhibit the optimization phase when compiling our code using these otherwise superb compilers.

3.3 Step-by-Step Guide to Using Our Test Suite

to be filled in

4 What will You Get? What did We Actually Get?

4.1 Output Format Explained

to be filled in

4.2 Environments under Which Our Test Suite is Known to Run

Our code has run successfully on a diverse variety of machines with a number of supported data types as indicated below:

- a VAX 8800 running VMS 4.4, D_floating and G_floating;
- a VAX 11/785 running 4.3BSD, D_floating and H_floating;
- a MIPS M1000 box with an R2010 FPU running UMIPS-BSD, IEEE 754 Single and IEEE 754 Double;
- a Sun 3/280 with an MC68881 running SunOS4.0, IEEE 754 Double;
- a Sun 3/140 with an FPA utilizing the WTL-1164/65 chip set running SunOS3.5, IEEE 754 Single and IEEE 754 Double;
- a Sun 4/280 with an FPU utilizing the WTL-1164/65 chip set running SunOS4.0, IEEE 754 Single and IEEE 754 Double;
- an IBM/PC with an Intel 8087 running PC-DOS 3.3 with Turbo C 2.0, IEEE 754 Extended;
- an Intel 80960KB processor with an integrated FPU running a version of UN*X, IEEE 754 Extended.

It is interesting to note that, due to the floating-point intensive nature of our code, earlier versions of our test suite uncovered a number of C compiler bugs, on-chip elementary function implementation glitches, and in one case even a hardware scheduling glitch.

5 Verification of Our Test Suite

5.1 Calibration

to be filled in

5.2 Cross-examination

We have implemented all the algorithms described in this paper and have thoroughly tested both the D_floating and the G_floating versions of our code on a VAX 8800 running VMS with G&H floating-point hardware and compared results delivered by our approximators with the H_floating version of the corresponding elementary functions in the VAX/VMS math library. The errors observed were reasonably less than our proved error bounds.

VAX D_floating format has a 56-bit mantissa and an 8-bit exponent, and is the default double-precision floating-point format commonly used on a VAX.

VAX G_floating format has a 53-bit mantissa and an 11-bit exponent, and is an alternate double-precision floating-point format which has been available in earlier models of VAXen only in microcoded form. It needs special WCS hardware which costs extra and doesn't come standard with the VAX. However, except for a different exponent bias and the lack of $\pm\infty$ and NaN, the VAX G_floating format is almost identical *in format* to the IEEE 754 "Double". Although the arithmetic performed on a VAX differs noticeably from IEEE 754 arithmetic, for our purposes testing the VAX G_floating version of our code should give us a fairly good idea as to how accurately our code will perform on IEEE 754 conforming machines.

Here is a brief summary of the test results with an input data of 64 and 2500 (meaning 64 subregions per test region and 2500 random arguments per subregion). All numbers are expressed in terms of ULPs of the corresponding H_floating results rounded to double. The column marked "bounds proved" summarizes the error bounds we were able to prove in a reasonably rigorous manner. NME means negative maximum error and PME means positive maximum error. The columns marked "(D)" and "(G)" are the D_floating and G_floating result respectively. $\mathcal{B}$ is $63 \cdot \log 2$ for D_floating and $969 \cdot \log 2$ for G_floating.

Name	NME/PME observed(D)	NME/PME observed(G)	Bounds proved	Intervals covered
sin	$-.0482/+.0482$	$-.0480/+.0486$	.0600	$[0, \pi/2)$
cos	$-.0479/+.0476$	$-.0475/+.0479$	.0611	$[0, \pi/2)$
atan	$-.0462/+.0460$	$-.0461/+.0463$	.0480	$[-2^{16}, 2^{16}]$
exp	$-.0112/+.0109$	$-.0111/+.0110$	.0280	$[-\mathcal{B}, \mathcal{B}]$
expm1	$-.0444/+.0432$	$-.0444/+.0431$	.0520	$[-1, 1]$
log	$-.0392/+.0378$	$-.0406/+.0375$	.0520	$[2^{-16.5}, 2^{16.5}]$
log1p	$-.0402/+.0391$	$-.0419/+.0394$	.0520	$[1/\sqrt{2}-1, \sqrt{2}-1]$

6 Mathematical Basis & Practical Importance of Our Test Suite

6.1 The Idea Behind: an Illustration

To better understand the general strategy our test suite uses, let's take atan as an example and assume that double-precision is the most precise floating-point data type understood by the C compiler we use and the subprogram under test is `DATAN`. For a double-precision argument `DX`, how can we evaluate atan(`DX`) to a few bits more accurate than any implementation of `DATAN` could possibly be? How can we even *represent* the generated test value in double-precision storage format? The answer is simple: we generate and store atan(`DX`) in several pieces.

For `DX` small enough in magnitude so that after lining up the binary points, atan(`DX`) − `DX` is shifted to the right relative to atan(`DX`) by at least 6 bits, we make use of the formula

$$\mathrm{atan}(x) = x - \frac{x}{\mathcal{R}\left(\dfrac{3}{x^2}\right)}$$

where $\mathcal{R}(u)$ is a continued fraction in $u := 3/x^2$ developed by W. Kahan (cf. [4] and Section XXX). One nice property of the continued fraction is that all constants involved have closed-form expressions and are all *rational* numbers, therefore they can be generated at run-time during the set-up phase once and for all. We can thus generate and store our test value in 2 pieces, namely

$$\begin{aligned}\mathtt{A1} &= \mathtt{DX},\\ \mathtt{A2} &= -\mathtt{DX}/\hat{\mathcal{R}}(3/(\mathtt{DX} * \mathtt{DX}))\end{aligned}$$

where both `A1` and `A2` are double-precision numbers and `A1` + `A2` approximates atan(`DX`) to at least 4 more bits than double-precision if rounding error in evaluating `A2` doesn't contaminate more than the last 2 bits of `A2`. The absolute rounding error of `DATAN(DX)` can be computed by 2 successive subtractions thus:

$$\text{absolute error} = (\mathtt{DATAN(DX)} - \mathtt{A1}) - \mathtt{A2}.$$

As long as `DATAN` is implemented not too badly, the first subtraction will be *exact* while the rounding error committed by the second subtraction can be safely disregarded. Let

$$\mathtt{A} := \mathtt{A1} + \mathtt{A2}\ \textit{rounded},$$

the relative rounding error in `DATAN(DX)` is thus

$$\text{relative error} = \text{absolute error}/\mathit{ulp}\,(\mathtt{A})$$

where

$$\begin{aligned}\mathit{ulp}\,(\mathtt{A}) &:= \mathrm{scalb}(1, \mathrm{logb}(\mathtt{A}) + 1 - t)\\ t &:= \text{number of significant bits in } \mathtt{A}.\end{aligned}$$

We will discuss in a later Section on how can $ulp\,(\mathtt{A})$ be computed portably and efficiently.

For DX not so small in magnitude, we perform the following argument reduction on DX with an appropriate shift of origin x_0 to get a small enough reduced argument DZ,

$$\mathtt{DZ} = (\mathtt{DX} - x_0)/(1 + x_0 * \mathtt{DX})$$

where x_0 belongs to a set of shifts of origin so carefully chosen that the rounding error in generating the reduced argument DZ will never exceed a small fraction, say 1/32 of an ULP of atan(DX). atan(x_0) is precomputed to 200 bits and stored in our source code in a form that a pair of double-precision numbers

$$\begin{aligned} \mathtt{A3} &:= \mathrm{atan}(x_0)\ \textit{rounded}, \\ \mathtt{A4} &:= (\mathrm{atan}(x_0) - \mathtt{A3})\ \textit{good to 16 bits} \end{aligned}$$

is easily reconstructed so that $\mathtt{A3}+\mathtt{A4}$ approximates atan(x_0) to within 0.0001 of an ULP of atan(x_0). Since atan(DZ) can be approximated as illustrated above by 2 pieces, say A1 and A2, we have

$$\begin{aligned} \mathrm{atan}(\mathtt{DX}) &\doteq \mathrm{atan}(x_0) + \mathrm{atan}(\mathtt{DZ}) \\ &\doteq \mathtt{A3} + \mathtt{A4} + \mathtt{A1} + \mathtt{A2} \end{aligned}$$

and the absolute rounding error in DATAN(DX) can be computed by 4 successive subtractions thus:

$$\text{absolute error} = (((\mathtt{DATAN(DX)} - \mathtt{A3}) - \mathtt{A1}) - \mathtt{A2}) - \mathtt{A4}.$$

The dominant rounding error in this case turns out to be the computation of $\mathtt{A1} := \mathtt{DZ}$ and by construction it never exceeds 1/32 of an ULP of atan(DX). The rest rounding errors can be explicitly estimated and bounded, they are analyzed in detail in the Appendix.

6.2 Practical Importance of Our Test Suite

As illustrated above, our test suite makes use of existing floating-point arithmetic without recourse to arithmetic more precise than that in which the program under test is embedded. Thus it will be most indispensable in measuring the accuracy of run-time math libraries utilizing the most precise floating-point data type available under a specific hardware configuration. For instance,

- on such chips as NSC32081, WTL-1164/65, the most precise floating-point data type is the IEEE 754 53-bit mantissa Double;
- on such chips/boards/machines as i80x87, MC68881, WE32106, Z8070, ELXSI 6400, HP 9000 series, Apple Macintosh, the most precise floating-point data type is the IEEE 754 64-bit mantissa Extended;

- on a VAX without H_floating microcode/hardware/emulation, the most precise floating-point data type is either the 56-bit mantissa D_floating or the 53-bit mantissa G_floating;
- on a VAX with H_floating microcode/hardware/emulation, the most precise floating-point data type is the 113-bit mantissa H_floating.

6.3 Outline of The Algorithms

to be filled in

Part II
Detailed Algorithms and Proofs

7 EXPM1 — Algorithm

7.1 Continued fraction expansion over the primary interval $[-1/8, \sqrt[5]{2}/8)$

For $x \in [-1/8, \sqrt[5]{2}/8)$ we make use of the following continued fraction expansion of $\tanh(x/2)$:

$$\tanh\left(\frac{x}{2}\right) := \frac{x}{2} + \frac{x/2}{cf(-3/(x/2)^2)}$$

where

$$cf(z) := z + A_1 + \cfrac{B_1}{z + A_2 + \cfrac{B_2}{z + A_3 + \cfrac{B_3}{z + \ddots}}}$$

with

$$A_n := \frac{-6}{(4n-3)(4n+1)}, \quad B_n := \frac{-9}{(4n-1)(4n+1)^2(4n+3)}, \quad n > 0.$$

Let

$$\sigma := \tanh\left(\frac{x}{2}\right) - \frac{x}{2} = \frac{x/2}{cf(-3/(x/2)^2)}.$$

We have

$$\begin{aligned} E(x) := e^x - 1 &= \frac{2\tanh(x/2)}{1 - \tanh(x/2)} \\ &= x + \frac{x^2}{2} + \mathcal{R}(x) \end{aligned}$$

with

$$\mathcal{R}(x) := \frac{\left(\frac{x^3}{4} + 2\sigma\right) + \left(x + \frac{x^2}{2}\right)\sigma}{1 - \left(\frac{x}{2} + \sigma\right)}.$$

7.2 Table-lookups over non-primary intervals

7.2.1 Criteria for selecting breakpoints and centers

For $x \in [L, -1/8) \cup [\sqrt[5]{2}/8, R)$, we select a sequence of $N+1$ breakpoints $\{b_k\}_{0 \le k \le N}$ and a sequence of N centers $\{c_k\}_{1 \le k \le N}$ so that the following 2 conditions are satisfied:

- $L := b_0 < c_1 < b_1 < \cdots < c_k < b_k < \cdots < c_N < b_N =: R$,
- If $x \in [b_{k-1}, b_k)$, then

$$\begin{aligned} 2 \cdot ulp\left((x-c_k)^2\right) &\leq ulp\left((x-c_k)E(c_k)\right) \\ &\leq \frac{1}{16}\min\{ulp\left(\exp(x)\right), ulp\left(E(x)\right)\}, \end{aligned}$$

  and

$$|x - c_k| < \frac{1}{8}.$$

Table 1 presents a selection of $\{b_k\}$ and $\{c_k\}$ satisfying the above conditions with the added property that $2^{10} \cdot b_k$ and $2^{10} \cdot c_k$ are all integers.

Table 1: $(N = 13)$

k	$2^{10} \cdot c_k$	$2^{10} \cdot b_k$	
0		-1062	$=: 2^{10} \cdot L$
1	-1011	-961	
2	-907	-853	
3	-794	-735	
4	-669	-603	
5	-523	-443	
6	-326	-268	
7	-178	-128	
8	0	147	
9	215	342	
10	407	534	
11	612	690	
12	749	867	
13	950	1033	$=: 2^{10} \cdot R$

7.2.2 Evaluation of $E(x)$

For $x \in [b_{k-1}, b_k)$, write $\xi := x - c_k$, then

$$E(x) := E(\xi + c_k) = E(c_k) + E(\xi) + E(c_k)E(\xi).$$

Since $|\xi| < 1/8$, ξ lies in the primary interval $[-1/8, \sqrt[5]{2}/8)$, by Section 7.2.1, $E(\xi)$ can be evaluated by

$$E(\xi) = \xi + \frac{\xi^2}{2} + \mathcal{R}(\xi)$$

where

$$\mathcal{R}(\xi) := \frac{\left(\frac{\xi^3}{4} + 2\sigma\right) + \left(\xi + \frac{\xi^2}{2}\right)\sigma}{1 - \left(\frac{\xi}{2} + \sigma\right)}$$

with

$$\sigma := \tanh\left(\frac{\xi}{2}\right) - \frac{\xi}{2} = \frac{\xi/2}{cf(-3/(\xi/2)^2)}.$$

The accurate values of $E(c_k)$ were pre-calculated to 200-bit precision using symbolic mathematics. In order that the accurate values of $E(c_k)$ be easily re-constructed, we store each one of them as an array of 12 consecutive long integers: its 200-bit mantissa stored as 10 20-bit array elements; the sign bit and the binary exponent stored in the remaining 2 slots. Table 2 lists the pre-calculated values of $E(c_k)$ in standard normalized form with hexadecimal mantissa. (cf. Table 1 for the values of $\{c_k\}$)

Table 2: (N=13)

k	$E(c_k)$
1	$-2^{-1}\cdot$ `1.413D4 0950B 7B4C3 E34EF C7DA7 DBF87 2C994 512B8 0D1F7 59F82`
2	$-2^{-1}\cdot$ `1.2CD8D CA033 0AC5A EDCE7 35895 A4579 310F8 43E63 8A503 B3A69`
3	$-2^{-1}\cdot$ `1.14363 7AA69 D7147 64C22 CA981 88C8E B1C00 86744 6B561 58F54`
4	$-2^{-2}\cdot$ `1.EB327 78E8D 2B51E 7D5A3 CC724 75CDA EFC47 E6661 0D794 A1115`
5	$-2^{-2}\cdot$ `1.998C7 9DF3C F97FA 29BF1 DF7D0 481C1 BC1B3 0343A ED5AD 368CE`
6	$-2^{-2}\cdot$ `1.1733D 40CE8 48436 7029C 16F1E 4828B D1EB9 41B6F F3565 8B6F9`
7	$-2^{-3}\cdot$ `1.46C6B 159F3 46316 85128 1F1A1 37717 BAD10 8825F BE89C DF672`
8	$2^{0}\cdot$ `0.00000 00000 00000 00000 00000 00000 00000 00000 00000 00000`
9	$2^{-3}\cdot$ `1.DE795 66421 DF785 06EAE 67666 00E05 0F97D D027F FC973 0E5B3`
10	$2^{-2}\cdot$ `1.F3C13 1CDB9 90E6B CAFF6 520C2 AB2CF 74628 646EA 2D158 0A401`
11	$2^{-1}\cdot$ `1.A2BDA 7ECFC F7660 768CB C27B0 815EF 73298 F4DAC E2EA6 557A3`
12	$2^{0}\cdot$ `1.13FD2 D2BA8 5BDD9 7F284 F8128 BD785 ABFF5 49626 4C8B1 8FF71`
13	$2^{0}\cdot$ `1.875DB 20DE2 3988D 218A0 96305 544CA B9EFA EE1F6 6611C BD1E5`

To evaluate $E(x)$, we write

$$E(c_k) \doteq \hat{E}(c_k) + \check{E}(c_k)$$

where

$$\hat{E}(c_k) := E(c_k) \quad \textit{rounded},$$
$$\check{E}(c_k) := (E(c_k) - \hat{E}(c_k)) \quad \textit{rounded}.$$

December 30, 1988

Then

$$
\begin{aligned}
E(x) &= E(c_k) + E(\xi) + E(\xi)E(c_k) \\
&= (\hat{E}(c_k) + \check{E}(c_k)) + \left(\xi + \frac{\xi^2}{2} + \mathcal{R}(\xi)\right) + \\
&\quad \left(\xi + \frac{\xi^2}{2} + \mathcal{R}(\xi)\right) (\hat{E}(c_k) + \check{E}(c_k)) \\
&= \hat{E}(c_k) + \xi + \xi \cdot \hat{E}(c_k) + \frac{\xi^2}{2} + \\
&\quad \mathcal{R}(\xi) + \frac{\xi^2}{2} \cdot \hat{E}(c_k) + \mathcal{R}(\xi) \cdot \hat{E}(c_k) + \\
&\quad \check{E}(c_k) + \xi \cdot \check{E}(c_k) + \frac{\xi^2}{2} \cdot \check{E}(c_k) + \mathbf{o}(\xi).
\end{aligned}
$$

Note that the tiny quantity $\mathbf{o}(\xi) := \mathcal{R}(\xi) \cdot \check{E}(c_k)$ is ignored.

8 EXP — Algorithm

8.1 Preliminaries

For easier presentation, the symbol $\mathcal{B}$ will be used throughout this section to represent the following quantity:

$$\mathcal{B} := \left(2^m - 2^{\lfloor \log_2 t \rfloor + 1} - 1\right) \cdot \log 2 \tag{1}$$

where

$$\begin{aligned} m &:= \text{the width in bits of the exponent field} - 1, \\ t &:= \text{the number of significant bits the target floating-point data type has.} \end{aligned}$$

We may reasonably assume that

$$t - 2m \geq 10. \tag{2}$$

For $n \neq 0$, define

$$\begin{aligned} \hat{\log} 2 &:= \log 2 \textit{ rounded up} \text{ to } t - m \text{ bits if } n > 0, \\ &:= \log 2 \textit{ rounded down} \text{ to } t - m \text{ bits if } n < 0 \end{aligned}$$

and

$$\check{\log} 2 := (\log 2 - \hat{\log} 2) \textit{ rounded}.$$

Notice that with the above setup, for all $n \neq 0$, we have

$$-n \cdot (\log 2 - \hat{\log} 2) > 0, \tag{3}$$
$$-n \,\check{\log} 2 \geq 0. \tag{4}$$

8.2 $x \in (-\log 2, \log 2)$

$$\begin{aligned} \exp(x) &= 1 + (e^x - 1) \\ &= 1 + E(x) \end{aligned}$$

and $E(x)$ is approximated by the algorithm described in Section 7.

8.3 $x \in [-\mathcal{B}, \mathcal{B}] \setminus [-\log 2/2, \log 2/2]$

Let

$$\begin{aligned} n &:= x / \hat{\log} 2 \textit{ rounded}, \\ \xi &:= (x - n \cdot \hat{\log} 2) \textit{ exactly}. \end{aligned}$$

Since

$$\begin{aligned} x - n \cdot \log 2 &= \xi - n \cdot (\log 2 - \hat{\log} 2) \\ &\doteq \xi - n \cdot \check{\log} 2, \end{aligned}$$

we have

$$\begin{aligned} 2^{-n} \cdot \exp(x) &= \exp(\xi) \cdot \exp(-n \cdot (\log 2 - \hat{\log} 2)) \\ &\doteq \exp(\xi) \cdot \exp(-n \cdot \check{\log} 2) \\ &= \exp(\xi) + \exp(\xi) \cdot (e^{-n \cdot \check{\log} 2} - 1) \\ &= 1 + E(\xi) + E(-n \cdot \check{\log} 2) + E(\xi) \cdot E(-n \cdot \check{\log} 2). \end{aligned}$$

Observe that since

$$|\xi| \leq \frac{\hat{\log} 2}{2},$$

both $E(\xi)$ and $E(-n \cdot \check{\log} 2)$ are approximated by the algorithm described in Section 7.

9 COSINE & SINE — Algorithms

9.1 Preliminaries

9.1.1 $\mathcal{R}_c(x)$ & $\mathcal{R}_s(x)$

Write

$$\tan\left(\frac{x}{2}\right) := \frac{x}{2} + \cfrac{\frac{x}{2}}{\frac{3}{(x/2)^2} - \frac{6}{5} - p(x)}.$$

Compute the quantity $-p(x)$ using the following continued fraction:

$$-p(x) := \cfrac{B_1}{\frac{3}{(x/2)^2} + A_2 + \cfrac{B_2}{\frac{3}{(x/2)^2} + A_3 + \cfrac{B_3}{\frac{3}{(x/2)^2} + \ddots}}}$$

where

$$A_n := \frac{-6}{(4n-3)(4n+1)}, \quad B_n := \frac{-9}{(4n-1)(4n+1)^2(4n+3)}, \quad n > 0.$$

Compute

$$U := 3 + \left(\frac{x}{2}\right)^2 \cdot \left(-\frac{1}{5} - p(x)\right),$$
$$V := 3 + \left(\frac{x}{2}\right)^2 \cdot \left(-\frac{6}{5} - p(x)\right).$$

Let

$$\mathcal{R}_c(x) := \cos(x) - 1 + \frac{x^2}{2},$$
$$\mathcal{R}_s(x) := \sin(x) - x.$$

Express $\mathcal{R}_c(x)$ and $\mathcal{R}_s(x)$ in terms of U and V, we have

$$\mathcal{R}_c(x) = \left(\frac{x^2}{2}\right) \cdot \left(\frac{U^2 - (U+V)}{U^2 + V^2/(x/2)^2}\right),$$
$$\mathcal{R}_s(x) = (-x) \cdot \left(\frac{U^2 - V}{U^2 + V^2/(x/2)^2}\right).$$

9.1.2 $\mathcal{F}(\delta,\theta;x)$

Given θ,δ where δ can be either 0 or 1, write

$$\xi := x-\theta, \quad C := \cos(\delta\cdot\frac{\pi}{2}+\theta), \quad S := \sin(\delta\cdot\frac{\pi}{2}+\theta).$$

Let

$$\begin{array}{ll} \hat{\theta} := \theta \textit{ rounded}, & \check{\theta} := (\theta-\hat{\theta}) \textit{ rounded}; \\ \tilde{\xi} := x-\hat{\theta}, & \check{\xi} := -\check{\theta}; \\ \hat{\xi} := (\tilde{\xi}+\check{\xi}) \textit{ rounded}, & \check{\xi} := ((\tilde{\xi}-\hat{\xi})+\check{\xi}) \textit{ rounded}; \\ \hat{S} := S \textit{ rounded}, & \check{S} := (S-\hat{S}) \textit{ rounded}. \end{array}$$

Since

$$\begin{aligned} \sin(\delta\cdot\frac{\pi}{2}+x) &= \sin(\delta\cdot\frac{\pi}{2}+\theta)\cos(\xi)+\cos(\delta\cdot\frac{\pi}{2}+\theta)\sin(\xi) \\ &= S\cdot\cos(\xi)+C\cdot\sin(\xi) \\ &= S+C\cdot\xi-S\cdot\frac{\xi^2}{2}+C\cdot\mathcal{R}_s(\xi)+S\cdot\mathcal{R}_c(\xi) \\ &\doteq (\hat{S}+\check{S})+C\cdot(\tilde{\xi}+\check{\xi})-(\hat{S}+\check{S})\cdot\left(\frac{\hat{\xi}^2}{2}+\hat{\xi}\cdot\check{\xi}\right)+ \\ &\quad C\cdot\mathcal{R}_s(\hat{\xi})+(\hat{S}+\check{S})\cdot\mathcal{R}_c(\hat{\xi}), \end{aligned}$$

we may thus define

$$\begin{aligned} \mathcal{F}(\delta,\theta;x) := {} & \hat{S}+C\cdot\tilde{\xi}-\hat{S}\cdot\frac{\hat{\xi}^2}{2}+C\cdot\mathcal{R}_s(\hat{\xi})+\hat{S}\cdot\mathcal{R}_c(\hat{\xi})+ \\ & \left(\check{S}+C\cdot\check{\xi}\right)-\hat{S}\cdot\hat{\xi}\cdot\check{\xi}-\check{S}\cdot\left(\frac{\hat{\xi}^2}{2}-\mathcal{R}_c(\hat{\xi})\right). \end{aligned}$$

9.2 COSINE

For easy reference, write

$$\begin{aligned} \left[0,\frac{\pi}{2}\right) &:= I_c \cup II_c \cup III_c \\ &:= \left[0,\frac{5}{16}\right)\cup\left[\frac{5}{16},\frac{3}{4}\right)\cup\left[\frac{3}{4},\frac{\pi}{2}\right). \end{aligned}$$

9.2.1 $x \in I_c$ ...

$$\cos(x) := 1-\frac{x^2}{2}+\mathcal{R}_c(x).$$

9.2.2 $x \in II_c$ ...

$$\cos(x) := \mathcal{F}\left(1, \frac{\pi}{6}; x\right).$$

9.2.3 $x \in III_c$ ...

Let

$$\frac{\hat{\pi}}{2} := \frac{\pi}{2} \textit{ chopped}, \qquad \frac{\check{\pi}}{2} := \left(\frac{\pi}{2} - \frac{\hat{\pi}}{2}\right) \textit{ rounded}.$$

Write

$$\tilde{\xi} := \frac{\hat{\pi}}{2} - x,$$

then

$$\begin{aligned}
\cos(x) &= \sin\left(\frac{\pi}{2} - x\right) \\
&\doteq \sin\left(\left(\frac{\hat{\pi}}{2} - x\right) + \frac{\check{\pi}}{2}\right) \\
&\doteq \sin(\tilde{\xi}) + \left(\frac{\check{\pi}}{2}\right) \cdot \left(1 - \frac{\tilde{\xi}^2}{2} \cdot \left(1 - \frac{\tilde{\xi}^2}{12}\right)\right).
\end{aligned}$$

9.3 SINE

Write

$$\begin{aligned}
\left[0, \frac{\pi}{2}\right) &:= I_s \cup II_s \cup III_s \cup IV_s \\
&:= \left[0, \frac{7}{16}\right) \cup \left[\frac{7}{16}, \frac{9}{16}\right) \cup \left[\frac{9}{16}, \frac{7}{8}\right) \cup \left[\frac{7}{8}, \frac{\pi}{2}\right).
\end{aligned}$$

9.3.1 $x \in I_s$ ...

Let

$$\begin{aligned}
Q_k(x) := 1 - \frac{x^2}{6 \cdot 7} \cdot \Bigg(1 - \frac{x^2}{8 \cdot 9} \cdot \Bigg(1 - \cdots \\
- \frac{x^2}{(2k-2) \cdot (2k-1)} \cdot \left(1 - \frac{x^2}{2k \cdot (2k+1)}\underbrace{\Bigg) \cdots \Bigg)}_{k-3}
\end{aligned}$$

and

$$P_k(x) := x - \frac{x^3}{6} + \frac{x^5}{120} \cdot Q_k(x), \quad k \geq 4.$$

Then

$$\sin(x) \doteq P_N(x)$$

where $N \geq 4$ is so chosen that

$$|\sin(x) - P_N(x)| \leq \frac{1}{256} ulp\,(\sin(x))\,.$$

9.3.2 $x \in II_s\ \ldots$

$$\sin(x) := \mathcal{F}\left(0, \cos^{-1}\left(\frac{7}{8}\right); x\right).$$

9.3.3 $x \in III_s\ \ldots$

$$\sin(x) := \mathcal{F}\left(0, \cos^{-1}\left(\frac{3}{4}\right); x\right).$$

9.3.4 $x \in IV_s\ \ldots$

Let

$$\frac{\hat{\pi}}{2} := \frac{\pi}{2}\ \textit{chopped}, \qquad \frac{\check{\pi}}{2} := \left(\frac{\pi}{2} - \frac{\hat{\pi}}{2}\right)\ \textit{rounded}.$$

Write

$$\tilde{\xi} := \frac{\hat{\pi}}{2} - x,$$

then

$$\begin{aligned}
\sin(x) &= \cos\left(\frac{\pi}{2} - x\right) \\
&\doteq \cos\left(\left(\frac{\hat{\pi}}{2} - x\right) + \frac{\check{\pi}}{2}\right) \\
&\doteq \cos(\tilde{\xi}) - \left(\frac{\check{\pi}}{2}\right) \cdot \tilde{\xi} \cdot \left(1 - \frac{\tilde{\xi}^2}{6}\right).
\end{aligned}$$

10 LOG — Algorithm

10.1 Preliminaries

Let

$$\rho := \frac{x - x_0}{x + x_0}$$

and

$$u := \frac{3}{\rho^2}.$$

Write

$$\tanh^{-1}(\rho) := \rho + \frac{\rho}{\mathcal{R}(u)}$$

where

$$\mathcal{R}(u) := u - A_1 - \cfrac{B_1}{u - A_2 - \cfrac{B_2}{u - A_3 - \cfrac{B_3}{u - \ddots}}}$$

and

$$A_n := \frac{12n(2n-1) - 3}{(4n-3)(4n+1)}, \quad B_n := \frac{36\bigl(n(2n+1)\bigr)^2}{\bigl((4n+1)^2 - 4\bigr)(4n+1)^2}, \qquad n > 0.$$

Define

$$\mathcal{R}_k(u) := u - A_k - \cfrac{B_k}{u - A_{k+1} - \cfrac{B_{k+1}}{u - A_{k+2} - \cfrac{B_{k+2}}{u - \ddots}}}, \qquad k > 0.$$

Notice that

$$\begin{aligned} \frac{9}{5} &= A_1 \geq A_n \searrow A_\infty := \frac{3}{2}, \\ \frac{108}{175} &= B_1 \geq B_n \searrow B_\infty := \frac{9}{16}; \end{aligned}$$

and for u large enough, say $u \geq 675$,

$$\mathcal{R}(u) \equiv \mathcal{R}_1(u) \leq \mathcal{R}_2(u) \leq \cdots \leq \mathcal{R}_k(u) \leq \mathcal{R}_{k+1}(u) \leq \cdots \leq u.$$

10.2 Continued fraction expansion over the primary interval $[1-\frac{1}{8}, 1+\frac{1}{8})$

For $x \in [1-\frac{1}{8}, 1+\frac{1}{8})$, using the ρ and u defined in Section 10.1 with $x_0 = 1$, we have

$$\begin{aligned}
\log(x) &= 2\tanh^{-1}\Big(\frac{x-1}{x+1}\Big) \\
&= 2\tanh^{-1}(\rho) \\
&= 2\rho + \frac{2\rho}{\mathcal{R}(u)} \\
&= (x-1) - \frac{(x-1)^2}{2} + \frac{(x-1)^3}{2(x+1)} + \frac{2\rho}{\mathcal{R}(u)}.
\end{aligned}$$

10.3 Table-lookups over non-primary intervals

10.3.1 Criteria for selecting breakpoints and centers

For $x \in \Big[\dfrac{1}{\sqrt{2}}, 1-\dfrac{1}{8}\Big) \cup \Big[1+\dfrac{1}{8}, \sqrt{2}\Big)$, we select a sequence of $N+1$ breakpoints $\{b_k\}_{0\le k\le N}$ and a sequence of N centers $\{c_k\}_{1\le k\le N}$ so that the following 3 conditions are satisfied:

- $\dfrac{1}{\sqrt{2}} =: b_0 < c_1 < b_1 < \cdots < c_k < b_k < \cdots < c_N < b_N := \sqrt{2}$,
- If $x \in [b_{k-1}, b_k) \cap \big([\frac{1}{\sqrt{2}}, 1-\frac{1}{8}) \cup [1+\frac{1}{8}, \sqrt{2})\big)$,

  $$ulp\left(\frac{x-c_k}{c_k}\right) \le \frac{1}{16} ulp\,(\log(x)) \tag{5}$$

  and necessarily

  $$\left|\frac{x-c_k}{c_k}\right| < \frac{1}{32} \tag{6}$$

  since for all $x \in \Big[\dfrac{1}{\sqrt{2}}, \sqrt{2}\Big)$

  $$ulp\,(\log(x)) \le \frac{1}{4} ulp\,(1)\,.$$

- $2^9 \cdot c_k$ are all integers.

Table 3 presents a selection of $\{b_k\}$ and $\{c_k\}$ satisfying the above conditions with the added property that $2^9 \cdot b_k$ are also integers except b_0 and b_N.

Table 3: $(N = 14)$

k	$2^9 \cdot c_k$	$2^9 \cdot b_k$	
0		$2^9 \cdot b_0$	$\equiv 2^9 \cdot \frac{1}{\sqrt{2}}$
1	371	382	
2	394	400	
3	406	412	
4	418	424	
5	430	436	
6	442	448	
7	512	576	
8	580	584	
9	593	602	
10	611	620	
11	629	638	
12	648	658	
13	678	699	
14	721	$2^9 \cdot b_N$	$\equiv 2^9 \cdot \sqrt{2}$

10.3.2 Evaluation of $\log(x)$

For $x \in [b_{k-1}, b_k) \cap \left([\frac{1}{\sqrt{2}}, 1 - \frac{1}{8}) \cup [1 + \frac{1}{8}, \sqrt{2})\right)$, using the ρ and u defined in Section 10.1 with $x_0 = c_k$, we have

$$\begin{aligned}
\log(x) &= \log(x_0) + \log\left(\frac{x}{x_0}\right) \\
&= \log(x_0) + 2\tanh^{-1}\left(\frac{x - x_0}{x + x_0}\right) \\
&= \log(x_0) + 2\tanh^{-1}(\rho) \\
&= \log(x_0) + 2\rho + \frac{2\rho}{\mathcal{R}(u)} \\
&= \log(x_0) + \xi - \frac{\xi^2}{2} + \frac{\rho}{2}\xi^2 + \frac{2\rho}{\mathcal{R}(u)}
\end{aligned}$$

where

$$\xi := \frac{x - x_0}{x_0}.$$

The accurate values of $\log(c_k)$ were pre-calculated to 200-bit precision using symbolic mathematics. In order that the accurate values of $\log(c_k)$ be easily re-constructed, we store each one of them as an array of 12 consecutive long integers: its 200-bit mantissa stored as 10 20-bit array elements; the sign bit and the binary exponent stored in the remaining 2 slots. Table 4 lists the pre-calculated values of $\log(c_k)$ in standard normalized form with hexadecimal mantissa. (cf. Table 3 for the values of $\{c_k\}$)

Table 4: (N=14)

k	$\log(c_k)$
1	$-2^{-2}\cdot$ `1.49DA7 F3BCC 41ECC D36BD 2E66A 6C718 21B02 EC7A5 1B6B8 0735D`
2	$-2^{-2}\cdot$ `1.0C42D 67616 2E311 62C79 D5D11 EE41E 3B351 FF419 49216 CA302`
3	$-2^{-3}\cdot$ `1.DB13D B0D48 94035 423A9 3F2D9 71062 F5613 9580F D566F 151CC`
4	$-2^{-3}\cdot$ `1.9F6C4 07089 66413 5A196 05E67 EF382 D7C64 D5883 4B04B 5F89B`
5	$-2^{-3}\cdot$ `1.6574E BE8C1 339F1 65878 5CEF2 095F4 F00EF F4801 CFCD1 34661`
6	$-2^{-3}\cdot$ `1.2D161 0C868 139D6 CCB81 B4A0D 41109 0848D 6F582 F0E24 72971`
7	$2^{0}\cdot$ `0.00000 00000 00000 00000 00000 00000 00000 00000 00000 00000`
8	$2^{-4}\cdot$ `1.FEC91 31DBE ABAAA 2E519 9F932 4E3BF E91E2 BA812 02EC6 15272`
9	$2^{-3}\cdot$ `1.2CCA0 F5F5F 25087 37280 7703F A7911 BC279 27E20 0E1E1 557AC`
10	$2^{-3}\cdot$ `1.6A079 D0F7A AD1FC 22468 A7AB0 1D0D2 2F4E8 2AE81 8A21A 51D83`
11	$2^{-3}\cdot$ `1.A57DF 28244 DCCE4 650EC D5DB1 C724D 16788 2A217 9EF90 A1D3C`
12	$2^{-3}\cdot$ `1.E2707 6E2AF 2E5E9 EA87F FE1FE 9E155 DB94E BC401 7F6F9 57DD0`
13	$2^{-2}\cdot$ `1.1F8FF 9E48A 2F28D 80819 7CED3 E58CF 23E43 622B0 B6EE3 7D610`
14	$2^{-2}\cdot$ `1.5E87B 20C29 549F4 63DDC E3E81 D7AC0 F4ABA 8BE5B 934DB A4AE6`

To evaluate $\log(x)$, we write

$$\log(x_0) \doteq \hat{\log}(x_0) + \check{\log}(x_0)$$

where

$$\begin{aligned}\hat{\log}(x_0) &:= \log(x_0) \textit{ rounded},\\ \check{\log}(x_0) &:= (\log(x_0) - \hat{\log}(x_0)) \textit{ rounded}.\end{aligned}$$

Then

$$\begin{aligned}\log(x) &\doteq (\hat{\log}(x_0) + \check{\log}(x_0)) + \xi - \frac{\xi^2}{2} + \frac{\rho}{2}\xi^2 + \frac{2\rho}{\mathcal{R}(u)}\\ &= \hat{\log}(x_0) + \xi - \frac{\xi^2}{2} + \frac{\rho}{2}\xi^2 + \frac{2\rho}{\mathcal{R}(u)} + \check{\log}(x_0).\end{aligned}$$

5 ATAN — Algorithm

5.1 Preliminaries

Let

$$u := \frac{3}{x^2}.$$

Write

$$\operatorname{atan}(x) := x - \frac{x}{\mathcal{R}(u)}$$

where

$$\mathcal{R}(u) := u + A_1 - \cfrac{B_1}{u + A_2 - \cfrac{B_2}{u + A_3 - \cfrac{B_3}{u + \ddots}}}$$

and

$$A_n := \frac{12n(2n-1)-3}{(4n-3)(4n+1)}, \quad B_n := \frac{36\big(n(2n+1)\big)^2}{\big((4n+1)^2-4\big)(4n+1)^2}, \qquad n > 0.$$

Define

$$\mathcal{R}_k(u) := u + A_k - \cfrac{B_k}{u + A_{k+1} - \cfrac{B_{k+1}}{u + A_{k+2} - \cfrac{B_{k+2}}{u + \ddots}}}, \qquad k > 0.$$

Notice that

$$\frac{9}{5} = A_1 \geq A_n \searrow A_\infty := \frac{3}{2},$$
$$\frac{108}{175} = B_1 \geq B_n \searrow B_\infty := \frac{9}{16};$$

and for u large enough, say $u \geq 57$,

$$\mathcal{R}_k(u) \geq u, \qquad k > 0.$$

Since $\operatorname{atan}(-x) \equiv -\operatorname{atan}(x)$, we are at the liberty of presenting our algorithm and proof only for positive x's with the understanding that everything we are about to say applies readily to negative x's.

5.2 Continued fraction expansion over the primary interval $[0, 0.2294)$

For $x \in [0, 0.2294)$, using the u and $\mathcal{R}$ defined in Section 5.1, we have

$$\operatorname{atan}(x) = x - \frac{x}{\mathcal{R}(u)}.$$

5.3 Table-lookups over non-primary intervals

5.3.1 Criteria for selecting breakpoints and centers

For $x \in [0.2294, 10.125)$, select a sequence of $N+1$ breakpoints $\{b_k\}_{0 \le k \le N}$ and a sequence of N centers $\{c_k\}_{1 \le k \le N}$ so that the following 3 conditions are satisfied:

- $0.2294 =: b_0 < c_1 < b_1 < \cdots < c_k < b_k < \cdots < c_N < b_N := 10.125.$
- Let $\xi := \dfrac{x - c_k}{1 + c_k \cdot x}$. If $x \in [b_{k-1}, b_k)$ then

$$ulp\,(\xi) \le \frac{3}{64} ulp\,(\operatorname{atan}(x))\,, \tag{7}$$

$$|\xi| \le \frac{1}{16}, \tag{8}$$

$$|\xi| \cdot ulp\,(1) \le \frac{1}{16} ulp\,(\operatorname{atan}(x))\,. \tag{9}$$

- $2^8 \cdot \dfrac{c_k \cdot ulp\,(1)}{ulp\,(c_k)}$ are all integers.

Table 5 presents a selection of $\{b_k\}$ and $\{c_k\}$ satisfying the above conditions with the added property that $2^8 \cdot \dfrac{b_k \cdot ulp\,(1)}{ulp\,(b_k)}$ are also integers except b_0.

5.3.2 Evaluation of $\operatorname{atan}(x)$ for $x \in [0.2294, 10.125)$

For $x \in [b_{k-1}, b_k)$, using the u and $\mathcal{R}$ defined in Section 5.1 with $x_0 = c_k$ and

$$\xi := \frac{x - x_0}{1 + x_0 \cdot x},$$

we have

$$\operatorname{atan}(x) = \operatorname{atan}(x_0) + \operatorname{atan}(\xi).$$

The accurate values of $\operatorname{atan}(c_k)$ were pre-calculated to 200-bit precision using symbolic mathematics. In order that the accurate values of $\operatorname{atan}(c_k)$ be easily re-constructed, we store each one of them as an array of 12 consecutive long integers: its 200-bit mantissa stored as 10 20-bit array elements; the sign bit

Table 5: $(N = 24)$

k	$2^9 \cdot c_k$	$2^9 \cdot b_k$	
0		$2^{10} \cdot b_0$	$\equiv 2^{10} \cdot 0.2294$
1	243	251	
2	259	267	
3	284	301	
4	318	335	
5	352	369	
6	387	405	
7	423	441	
8	460	479	
9	498	516	
10	536	556	
11	576	618	
12	662	708	
13	756	806	
14	858	912	
15	970	1032	
16	1076	1124	
17	1176	1236	
18	1304	1380	
19	1472	1552	
20	1644	1896	
21	2144	2424	
22	2880	3280	
23	4008	4736	
24	6368	$2^{10} \cdot b_N$	$\equiv 2^{10} \cdot 10.125$

and the binary exponent stored in the remaining 2 slots. Table 6 lists the precalculated values of $\mathrm{atan}(c_k)$ in standard normalized form with hexadecimal mantissa. (cf. Table 5 for the values of $\{c_k\}$)

To evaluate $\mathrm{atan}(x)$, we write

$$\mathrm{atan}(x_0) \doteq \hat{\mathrm{atan}}(x_0) + \check{\mathrm{atan}}(x_0)$$

where

$$\hat{\mathrm{atan}}(x_0) := \mathrm{atan}(x_0) \textit{ rounded},$$
$$\check{\mathrm{atan}}(x_0) := (\mathrm{atan}(x_0) - \hat{\mathrm{atan}}(x_0)) \textit{ rounded}.$$

Then

Table 6: (N=24) with $c_{N+1} := +\infty$

k	$\mathrm{atan}(c_k)$
1	$2^{-3}\cdot$ 1.DD2C6 F45DB 8B9C9 BEC03 BF27A E608D 172E2 A99D0 63F59 7EE87
2	$2^{-3}\cdot$ 1.FB5C0 55893 475A0 7ABA7 97DB3 B0407 66ED8 DAC27 BBC40 B462B
3	$2^{-2}\cdot$ 1.15097 3A9CE 546A1 A5160 18D73 61842 2D2FC 5C9B4 00B6F B8CFE
4	$2^{-2}\cdot$ 1.3454B E5720 A003B DFFFC A144B 3EC7B D9DA0 F3811 3B509 B4FCE
5	$2^{-2}\cdot$ 1.530AD 9951C D49DB 5336F EEF7E FB3D1 82425 873A6 3DE9A FA744
6	$2^{-2}\cdot$ 1.72023 E88EA 0A13D E5F2F D2AB9 D5AAE 1D797 29190 FF855 1546B
7	$2^{-2}\cdot$ 1.91234 C0BF7 1368A 81880 74F63 1BC4F 64840 976F2 1D0AA BC71A
8	$2^{-2}\cdot$ 1.B0564 20AE9 3439B 3B1AE 7272E AA9AE BBFCD 6ABC9 35ED2 469BC
9	$2^{-2}\cdot$ 1.CF839 6BC7F C8DF6 6C7D4 52684 B5192 207F1 50560 89632 9A59C
10	$2^{-2}\cdot$ 1.EDCB6 D43F8 434E0 3689C CF77B 1C4C9 B8B02 17518 6CB46 88B20
11	$2^{-1}\cdot$ 1.0657E 94DB3 0CFC5 496D4 1396C 34A2B 81E22 AB9B0 EE9BB F78AB
12	$2^{-1}\cdot$ 1.25D63 21466 46F52 AF45B 34B3F 07A5B FE468 E1936 8B554 FA6FF
13	$2^{-1}\cdot$ 1.459C6 52BAD C7F46 65A63 A384D 6F512 7A7EA DAAEB D979B E14AA
14	$2^{-1}\cdot$ 1.65147 8826E 4C87C 44D71 098F9 7A769 790D5 1DD99 14C83 32723
15	$2^{-1}\cdot$ 1.8442F B8FC6 7D2C7 C70B8 24B90 0E330 38741 08033 4F0CC C5B27
16	$2^{-1}\cdot$ 1.9ECCA 32969 5E07A 270E5 0A9E1 94A28 CDE69 19685 CF1BC 462D5
17	$2^{-1}\cdot$ 1.B5713 92769 134BE 824CA 47366 18765 9AAD8 B85EE 4CC77 E7C70
18	$2^{-1}\cdot$ 1.CF690 462A5 D2740 225CC 4CAAC 03288 81752 42401 FBAF1 AA7CE
19	$2^{-1}\cdot$ 1.ED0D9 7C904 1C8F6 CEB0E 2512D B7F4D 344F1 5D499 61BE9 ACF4C
20	$2^{0}\cdot$ 1.0383C 54504 2E95D 0EE64 52553 66A06 1698B 5E33C 97DA8 D5742
21	$2^{0}\cdot$ 1.200E5 AE0DD 61D37 DA79B B203C 2F38F A7C42 81C63 4570D 85134
22	$2^{0}\cdot$ 1.3AAB9 8641F 26AC4 796C6 7D148 72C07 2DFF7 1FBEB 3CA29 AB078
23	$2^{0}\cdot$ 1.5216A 87790 2EE45 E06F7 1ACDB 25899 83F01 D87F1 F122F EBA06
24	$2^{0}\cdot$ 1.694EB 4CD16 1D800 E8C63 F1687 2CC6B 1CB94 8380F 0918C EF2A0
25	$2^{0}\cdot$ 1.921FB 54442 D1846 9898C C5170 1B839 A2520 49C11 14CF9 8E804

$$
\begin{aligned}
\mathrm{atan}(x) &\doteq (\hat{\mathrm{atan}}(x_0) + \check{\mathrm{atan}}(x_0)) + \xi - \frac{\xi}{\mathcal{R}(u)} \\
&= \hat{\mathrm{atan}}(x_0) + \xi - \frac{\xi}{\mathcal{R}(u)} + \check{\mathrm{atan}}(x_0).
\end{aligned}
$$

5.3.3 Evaluation of $\mathrm{atan}(x)$ for $x \geq b_N := 10.125$

For $x \geq b_N$, write

$$\xi := -\frac{1}{x},$$

we have

$$\mathrm{atan}(x) = \frac{\pi}{2} + \mathrm{atan}(\xi).$$

The accurate value of $\frac{\pi}{2}$ is presented in Table 6 as $\mathrm{atan}(c_{N+1})$.
To evaluate $\mathrm{atan}(x)$, we write

$$\frac{\pi}{2} \doteq \frac{\hat{\pi}}{2} + \frac{\check{\pi}}{2}$$

where

$$\begin{aligned}
\frac{\hat{\pi}}{2} &:= \frac{\pi}{2} \textit{ rounded},\\
\frac{\check{\pi}}{2} &:= \left(\frac{\pi}{2} - \frac{\hat{\pi}}{2}\right) \textit{ rounded}.
\end{aligned}$$

Thus

$$\begin{aligned}
\mathrm{atan}(x) &\doteq \frac{\hat{\pi}}{2} + \frac{\check{\pi}}{2} + \xi - \frac{\xi}{\mathcal{R}(u)}\\
&= \frac{\hat{\pi}}{2} + \xi - \frac{\xi}{\mathcal{R}(u)} + \frac{\check{\pi}}{2}.
\end{aligned}$$

A EXPM1 — Accuracy Statement and Proof

For an arbitrary expression e, let $\varepsilon\{e\} := |fl(e) - e|$ and $\varepsilon[e] := \frac{1}{2}ulp(e)$. Here is our main result of the Section:

Theorem A.1 *Using the algorithm and the established values of L and R presented in Section 7, we have*

$$\varepsilon E(x) \leq 0.052 \cdot ulp(E(x)) \qquad \forall x \in [L, R).$$

First some preparations.

Lemma A.2

$$ulp\left(\frac{x^2}{2}\right) \leq \frac{1}{16} ulp(E(x)) \qquad \forall x \in \left(-\frac{1}{8}, \frac{\sqrt{2}}{8}\right).$$

Proof: Since

$$\begin{array}{l} \dfrac{|x|}{\sqrt{2}} \leq |E(x)| \text{ for } x \in \left[-\dfrac{\sqrt{2}}{16}, 0\right) \quad \text{and} \\ 0 \leq x \leq E(x) \text{ for all } x \geq 0, \end{array}$$

we have

$$\frac{x^2}{2} \leq \frac{1}{16}|E(x)| \text{ for } x \in \left[-\frac{\sqrt{2}}{16}, \frac{1}{8}\right).$$

For $x \in \left[\frac{1}{8}, \frac{\sqrt{2}}{8}\right)$,

$$ulp\left(\frac{x^2}{2}\right) = \frac{1}{2} \cdot \frac{ulp(x)}{ulp(1)} ulp(x) = \frac{1}{2} \cdot \frac{1}{8} ulp(x) \leq \frac{1}{16} ulp(E(x)).$$

For $x \in \left(-\frac{1}{8}, -\frac{\sqrt{2}}{16}\right)$,

$$ulp\left(\frac{x^2}{2}\right) = \frac{ulp(x)}{ulp(1)} ulp(x) = \frac{1}{16} ulp(x) = \frac{1}{16} ulp(E(x)).$$

$\mathcal{QED}$

Lemma A.3 *Let*

$$\mathcal{R}(x) := E(x) - x - \frac{x^2}{2},$$

then

$$ulp(\mathcal{R}(x)) \leq \frac{1}{16} ulp\left(\frac{x^2}{2}\right) \qquad \forall |x| < \frac{\sqrt[3]{2}}{8}.$$

Proof: Since

$$|\mathcal{R}(x)| \leq \frac{|x|^3}{4} \quad \text{for } |x| < \frac{\sqrt[3]{2}}{8},$$

for $|x| < \dfrac{1}{8}$ we have

$$ulp\,(\mathcal{R}(x)) \leq \frac{1}{4} ulp\left(|x| \cdot x^2\right) \leq \frac{1}{16} ulp\left(\frac{x^2}{2}\right).$$

For $|x| \in \left[\dfrac{1}{8}, \dfrac{\sqrt[3]{2}}{8}\right)$, since $ulp\left(x^3\right) = \dfrac{ulp\,(x)}{ulp\,(1)} ulp\left(x^2\right) = \dfrac{1}{8} ulp\left(x^2\right)$,

$$ulp\,(\mathcal{R}(x)) \leq \frac{1}{4} ulp\left(x^3\right) = \frac{1}{16} ulp\left(\frac{x^2}{2}\right).$$

QED

Lemma A.4 *Let*

$$p(x) := \frac{-x/2}{x/2 - \tanh(x/2)} + \frac{12}{x^2} + \frac{6}{5},$$

then

$$(1 - \frac{x^2}{9})\frac{x^2}{700} \leq p(x) \leq \frac{\frac{x^2}{700}}{1 - \frac{x^2}{10}} \leq (1 + \frac{x^2}{9})\frac{x^2}{700} \qquad \forall\, |x| < 1.$$

In particular, we have

$$0 \leq p(x) \leq (1 + \frac{x^2}{9})\frac{x^2}{700}\bigg|_{x = \frac{\sqrt[5]{2}}{8}} \leq \frac{1}{2^{15}} \text{ for } |x| \leq \frac{\sqrt[5]{2}}{8}.$$

Proof: Using the fact that the Taylor series expansion of $\tanh(x)$ around 0 is alternating and

$$\frac{x}{2} - \tanh\left(\frac{x}{2}\right) = \frac{x^3}{24}(1 - \frac{1}{10}x^2 + \frac{17}{1680}x^4 - \mathbf{o}(x^4)), \quad \mathbf{o}(x^4) \geq 0,$$

it is not hard to establish the following expressions for $p(x)$:

$$\begin{aligned} p(x) &= \frac{\frac{x^2}{700}(1 - \mathbf{o}(1))}{1 - \frac{x^2}{10} + \mathbf{o}(x^2)} \\ &= \frac{\frac{x^2}{700}(1 - \frac{x^2}{9} + \mathbf{o}(x^2))}{1 - \mathbf{o}(1)} \end{aligned}$$

where all occurrences of $\mathbf{o}(1)$ and $\mathbf{o}(x^2)$ are non-negative. Hence follows the Lemma. *QED*

Lemma A.5 *Let*

$$\sigma := \tanh\left(\frac{x}{2}\right) - \frac{x}{2} = \frac{x/2}{cf} = \frac{x/2}{-3/(x/2)^2 + (-1.2 - p(x))}$$

and

$$p(x) = \cfrac{3/175}{-3/(x/2)^2 - 2/15 + \cfrac{-1/693}{-3/(x/2)^2 + \ddots}},$$

then

$$\varepsilon\{\sigma\} \leq \frac{1}{2} ulp(\sigma) + \frac{3}{2} \cdot |\sigma| \cdot ulp(1) \qquad \forall x \in \left[-\frac{1}{8}, \frac{\sqrt[5]{2}}{8}\right).$$

Proof: We may safely assume that

$$\varepsilon\{p(x)\} \leq 2^{12} ulp(p(x)).$$

By Lemma A.4,

$$0 \leq p(x) \leq \frac{1}{2^{15}} \text{ for } |x| \leq \frac{\sqrt[5]{2}}{8};$$

we thus have

$$\begin{cases} \varepsilon\{p(x)\} \leq 2^{12} \cdot \dfrac{1}{2^{15}} ulp(1) = \dfrac{1}{8} ulp(1), & (\dagger) \\ \left|\dfrac{-3}{(x/2)^2}\right| + |-1.2 - p(x)| = |cf|, & (\ddagger) \\ 1.3 \geq |-1.2 - p(x)| = 1.2 + |p(x)| \geq 1.2 \text{ for } |x| \leq \dfrac{\sqrt[5]{2}}{8}. & (\star) \end{cases}$$

To estimate $\varepsilon\{cf\}$, notice that

$$\varepsilon\{cf\} \leq \varepsilon\{-1.2 - p(x)\} + \varepsilon\left\{\frac{-3}{(x/2)^2}\right\} + \varepsilon\left[\frac{-3}{(x/2)^2} + (-1.2 - p(x))\right],$$

since

$$\begin{aligned} \varepsilon\{-1.2 - p(x)\} &\leq \varepsilon[-1.2] + \varepsilon\{-p(x)\} + \varepsilon[(-1.2) + (-p(x))] \\ &\overset{(\dagger),(\star)}{\leq} \left(\frac{1}{2} + \frac{1}{8} + \frac{1}{2}\right) \cdot ulp(1) \overset{(\star)}{\leq} |-1.2 - p(x)| \cdot ulp(1) \\ \varepsilon\left\{\frac{-3}{(x/2)^2}\right\} &\leq \varepsilon\left[\frac{-3}{(x/2)^2}\right] + \left|\frac{-3}{(x/2)^2}\right| \cdot \frac{\varepsilon\left[(x/2)^2\right]}{(x/2)^2} \\ &\leq \frac{1}{2} ulp\left(\frac{-3}{(x/2)^2}\right) + \frac{1}{2} \cdot \left|\frac{-3}{(x/2)^2}\right| \cdot \frac{ulp\left((x/2)^2\right)}{(x/2)^2} \\ &\leq \frac{1}{2} \cdot \left|\frac{-3}{(x/2)^2}\right| \cdot ulp(1) + \frac{1}{2} \cdot \left|\frac{-3}{(x/2)^2}\right| \cdot ulp(1) \\ &= \left|\frac{-3}{(x/2)^2}\right| \cdot ulp(1), \end{aligned}$$

we got

$$\begin{aligned}\varepsilon\{cf\} &\le \left(|-1.2-p(x)| + \left|\frac{-3}{(x/2)^2}\right|\right)\cdot ulp\,(1) + \frac{1}{2}ulp\,(cf)\\ &\overset{(\ddagger)}{\le} |cf|\cdot ulp\,(1) + \frac{1}{2}\cdot|cf|\cdot ulp\,(1) = \frac{3}{2}\cdot|cf|\cdot ulp\,(1)\,.\end{aligned}$$

To complete the proof, note that

$$\begin{aligned}\varepsilon\{\sigma\} &\le \varepsilon\left[\frac{x/2}{cf}\right] + |\sigma|\cdot\frac{\varepsilon\{cf\}}{|cf|}\\ &\le \frac{1}{2}ulp\,(\sigma) + \frac{3}{2}\cdot|\sigma|\cdot ulp\,(1)\,.\end{aligned}$$

$\mathcal{QED}$

Lemma A.6 *Let*

$$\sigma := \tanh\left(\frac{x}{2}\right) - \frac{x}{2}, \qquad |x| \le \frac{\sqrt[5]{2}}{8},$$

then

1. $\left|x + \frac{x^2}{2}\right| \le 0.154$
2. $x\cdot\sigma < 0$ *unless* $x = 0$; $\left|\frac{x^3}{24.1}\right| \le |\sigma| \le \left|\frac{x^3}{24}\right|$
3. $\max\left\{\left|\frac{x^3}{4}+2\sigma\right|, \left|(\frac{x^3}{4}+2\sigma)+(x+\frac{x^2}{2})\sigma\right|\right\} \le \frac{1}{16}\left(\frac{x^2}{2}\right)$
4. $ulp\,(x^3) \le \frac{1}{4}ulp\left(\frac{x^2}{2}\right)$
5. $|\sigma|\cdot ulp\,(1) \le \frac{1}{48}ulp\left(\frac{x^2}{2}\right)$
6. $ulp\,(\sigma) \le \frac{1}{64}ulp\left(\frac{x^2}{2}\right)$
7. $\varepsilon\{\sigma\} \le \frac{5}{128}ulp\left(\frac{x^2}{2}\right)$.

Proof: (A.6.1) holds since $|x| \le \frac{\sqrt[5]{2}}{8}$; (A.6.2) follows from the fact that the Taylor series expansion of $\tanh(x)$ being an alternating series and making use of its first 3 terms; (A.6.3) follows from (A.6.2). (A.6.4) holds trivially for $|x| < \frac{1}{8}$; for $|x| \in [\frac{1}{8}, \frac{\sqrt[5]{2}}{8})$, note that

$$ulp\,(x^3) = \frac{ulp\,(x)}{ulp\,(1)}ulp\,(x^2) = \frac{1}{8}ulp\,(x^2)\,.$$

(A.6.5) follows from (A.6.2) for $|x| < \frac{1}{8}$; for $|x| \in [\frac{1}{8}, \frac{\sqrt[5]{2}}{8})$, note that

$$|x|^3 \cdot ulp\,(1) \le 2 \cdot ulp\left(x^3\right) = 2\frac{ulp\,(x)}{ulp\,(1)} ulp\left(x^2\right) = \frac{1}{4} ulp\left(x^2\right).$$

(A.6.6) follows from (A.6.5); (A.6.7) follows from (A.6.5), (A.6.6) and Lemma A.5. $\mathcal{QED}$

Lemma A.7 *Recall from Lemma A.3 that*

$$\mathcal{R}(x) := e^x - 1 - x - \frac{x^2}{2},$$

$\mathcal{R}(x)$ *can be expressed in the form of* A/B *with*

$$A := \left(\frac{x^3}{4} + 2\sigma\right) + \left(x + \frac{x^2}{2}\right)\sigma$$
$$B := 1 - \left(\frac{x}{2} + \sigma\right)$$

where σ *was defined in Lemma A.5. Then for* $|x| \le \frac{\sqrt[5]{2}}{8}$, *we have*

1. $\varepsilon\{A\} \le 0.2172 \cdot ulp\left(\frac{x^2}{2}\right)$,
2. $\varepsilon\{B\} \le 0.532 \cdot ulp\,(1)$,
3. $\varepsilon\{\mathcal{R}(x)\} \le 0.323 \cdot ulp\left(\frac{x^2}{2}\right)$.

Proof: of (A.7.1). We have the following estimates:

$$\frac{1}{2} ulp\left(\frac{x^3}{4} + 2\sigma\right) + \frac{1}{2} ulp\,(A) \overset{(A.6.3)}{\le} \left(\frac{1}{2} + \frac{1}{2}\right) ulp\left(\frac{1}{16} \cdot \frac{x^2}{2}\right) = \frac{1}{16} ulp\left(\frac{x^2}{2}\right)$$

$$\varepsilon\left\{\frac{x^3}{4}\right\} \le \frac{1}{2}|x| \cdot ulp\left(\frac{x^2}{4}\right) + \frac{1}{2} ulp\left(\frac{x^3}{4}\right) \overset{(A.6.4)}{\le} \left(\frac{\sqrt[5]{2}}{32} + \frac{1}{32}\right) ulp\left(\frac{x^2}{2}\right)$$

$$\varepsilon\{2\sigma\} \overset{(A.6.7)}{\le} 2 \cdot \frac{5}{128} ulp\left(\frac{x^2}{2}\right) = \frac{5}{64} ulp\left(\frac{x^2}{2}\right)$$

$$\left(x + \frac{x^2}{2}\right)\varepsilon\{\sigma\} \overset{(A.6.1)}{\le} 0.154 \cdot \frac{5}{64} ulp\left(\frac{x^2}{2}\right) \le 0.00602 \cdot ulp\left(\frac{x^2}{2}\right)$$

$$\frac{1}{2} ulp\left(\left(x + \frac{x^2}{2}\right)\sigma\right) \overset{(A.6.1)}{\le} \frac{1}{2} ulp\,(0.154\sigma) \le \frac{1}{8} ulp\,(\sigma) \overset{(A.6.6)}{\le} \frac{1}{512} ulp\left(\frac{x^2}{2}\right)$$

$$\varepsilon\left\{x + \frac{x^2}{2}\right\} \cdot |\sigma| \le |\sigma| \cdot \left[\frac{1}{2} ulp\left(x + \frac{x^2}{2}\right) + \frac{1}{2} ulp\left(\frac{x^2}{2}\right)\right]$$
$$\overset{(A.6.1)}{\le} \frac{1}{16}|\sigma| \cdot ulp\,(1) + \frac{|x|^3}{48} ulp\left(\frac{x^2}{2}\right) \overset{(A.6.5)}{\le} 0.00137 \cdot ulp\left(\frac{x^2}{2}\right).$$

Hence

$$\begin{aligned}
\varepsilon\{A\} &\le \left(\varepsilon\left[\left(\frac{x^3}{4}+2\sigma\right)+\left(x+\frac{x^2}{2}\right)\sigma\right]+\varepsilon\left[\frac{x^3}{4}+2\sigma\right]\right)+ \\
&\quad \varepsilon\left\{\frac{x^3}{4}\right\}+\varepsilon\{2\sigma\}+\varepsilon\left\{(x+\frac{x^2}{2})\sigma\right\} \\
&\le \left(\frac{1}{16}+\frac{\sqrt[5]{2}+1}{32}+\frac{5}{64}+(0.00602+\frac{1}{512}+0.00137)\right)\cdot ulp\left(\frac{x^2}{2}\right) \\
&\le 0.2172\cdot ulp\left(\frac{x^2}{2}\right)
\end{aligned}$$

$\mathcal{QED}$

Proof: of (A.7.2). By using the following 3 estimates,

$$\begin{aligned}
ulp\left(1-\left(\frac{x}{2}+\sigma\right)\right) &\overset{(A.6.2)}{\le} ulp\left(1+\left|\frac{x}{2}\right|\right)=ulp\,(1)\,, \\
ulp\left(\frac{x}{2}+\sigma\right) &\overset{(A.6.2)}{\le} ulp\left(\frac{x}{2}\right)\le\frac{1}{16}ulp\,(1)\,, \\
\varepsilon\{\sigma\} &\overset{(A.6.7)}{\le} \frac{5}{128}ulp\left(\frac{x^2}{2}\right)\le\frac{5}{2^{14}}ulp\,(1)\,,
\end{aligned}$$

we obtain

$$\begin{aligned}
\varepsilon\{B\} &\le \varepsilon\left[1-\left(\frac{x}{2}+\sigma\right)\right]+\varepsilon\left[\frac{x}{2}+\sigma\right]+\varepsilon\{\sigma\} \\
&\le \left(\frac{1}{2}+\frac{1}{32}+\frac{5}{2^{14}}\right)ulp\,(1) \\
&\le 0.532\cdot ulp\,(1)\,.
\end{aligned}$$

$\mathcal{QED}$

Proof: of (A.7.3). Observe that

$$\begin{aligned}
&B\ge 1-\left|\frac{x}{2}\right|\ge 0.9282, \\
&\left|\frac{A}{B}\right|=\left|e^x-1-x-\frac{x^2}{2}\right|\le\frac{|x|^3}{5}\text{ and }ulp\left(\frac{A}{B}\right)\le\frac{1}{16}ulp\left(\frac{x^2}{2}\right) \\
&\frac{|x|^3}{5}ulp\,(1)\le\frac{1}{10}ulp\left(\frac{x^2}{2}\right),
\end{aligned}$$

therefore

$$\varepsilon\left\{\frac{A}{B}\right\} \quad\le\quad \varepsilon\left[\frac{A}{B}\right]+\frac{1}{B}\left(\varepsilon\{A\}+\left|\frac{A}{B}\right|\cdot\varepsilon\{B\}\right)$$

$$
\begin{aligned}
&\overset{(A.7.1),\,(A.7.2)}{\leq} \frac{1}{2} ulp\left(\frac{A}{B}\right) + \frac{1}{B}\left(0.2172 \cdot ulp\left(\frac{x^2}{2}\right) + 0.532 \cdot \frac{|x|^3}{5} ulp\,(1)\right) \\
&\leq \left(\frac{1}{32} + \frac{0.2172 + 0.0532}{0.9282}\right) \cdot ulp\left(\frac{x^2}{2}\right) \leq 0.323 \cdot ulp\left(\frac{x^2}{2}\right).
\end{aligned}
$$

$\mathcal{QED}$

We can now easily prove our theorem for x in the primary interval. For $x \in [-\frac{1}{8}, \frac{\sqrt[5]{2}}{8})$, our $E(x)$ approximator says

$$E(x) = x + \frac{x^2}{2} + \mathcal{R}(x),$$

hence

$$
\begin{aligned}
\varepsilon E(x) &\leq \varepsilon\left[\frac{x^2}{2}\right] + \varepsilon\{\mathcal{R}(x)\} \\
&\overset{(A.7.3)}{\leq} \frac{1}{2} ulp\left(\frac{x^2}{2}\right) + 0.323 \cdot ulp\left(\frac{x^2}{2}\right) \\
&= 0.823 \cdot ulp\left(\frac{x^2}{2}\right) \overset{(A.3)}{\leq} \frac{0.823}{16} ulp\,(E(x)) \leq 0.052 \cdot ulp\,(E(x))
\end{aligned}
$$

as claimed.

For x not in the primary interval, say $x \in [b_{k-1}, b_k)$. Let

$$\xi := x - c_k.$$

Our $E(x)$ approximator says

$$
\begin{aligned}
E(x) &= E(\xi + c_k) \\
&= E(\xi) + E(c_k) + E(\xi)E(c_k) \\
&= \xi + \hat{E}(c_k) + \check{E}(c_k) + \\
&\quad \xi \cdot \hat{E}(c_k) + \frac{\xi^2}{2} + \mathcal{R}(\xi) + \frac{\xi^2}{2} \cdot \hat{E}(c_k) + \\
&\quad \mathcal{R}(\xi) \cdot \hat{E}(c_k) + \xi \cdot \check{E}(c_k) + \frac{\xi^2}{2} \cdot \check{E}(c_k) + \mathbf{o}(\xi).
\end{aligned}
$$

The ignored quantity $\mathbf{o}(\xi) := \mathcal{R}(\xi) \cdot \check{E}(c_k)$.

Lemma A.8 *For reference, here are the inequalities to be used to complete the proof of Theorem A.1:*

1. $|\xi| < \frac{1}{8}, \quad ulp\,(\xi) \leq \frac{1}{16} ulp\,(1),$
2. $ulp\left(\xi \cdot \hat{E}(c_k)\right) \leq \frac{1}{16} ulp\,(E(x)) \quad \forall k,$

3. $ulp\left(\frac{\xi^2}{2}\right) \le \frac{1}{4} ulp\left(\xi \cdot \hat{E}(c_k)\right), \quad \forall k,$

4. $ulp\left(\xi^2\right) \le \sqrt{2} \cdot |\xi| \cdot ulp\,(\xi),$

5. $\frac{1}{2}|u| \cdot ulp\,(v) \le ulp\,(u \cdot v) \le 2 \cdot |u| \cdot ulp\,(v),$

6. $|\mathcal{R}(\xi)| \le \frac{|\xi|}{320}; \quad ulp\,(\mathcal{R}(\xi)) \le \frac{1}{256} ulp\,(\xi),$

7. $|\check{E}(c_k)| \le \frac{1}{2} ulp\left(\hat{E}(c_k)\right) \le 2^{-24} \cdot |\hat{E}(c_k)| \quad \forall k.$

Proof: of Theorem A.1. To complete our proof for $x \in [b_{k-1}, b_k)$ outside of the primary interval, note that

$$\begin{aligned}
\varepsilon E(x) \quad &\le \quad \varepsilon\left[\xi \cdot \hat{E}(c_k)\right] + \varepsilon\left[\frac{\xi^2}{2}\right] + \varepsilon\left\{\mathcal{R}(\xi)\right\} + \varepsilon\left[\frac{\xi^2}{2}\right] \cdot |\hat{E}(c_k)| + \\
&\quad\quad \varepsilon\left[\mathcal{R}(\xi)\right] \cdot |\hat{E}(c_k)| + \varepsilon\left[\frac{\xi^2}{2} \cdot \hat{E}(c_k)\right] + \varepsilon\left[\mathcal{R}(\xi) \cdot \hat{E}(c_k)\right] + \\
&\quad\quad \varepsilon\left[\xi \cdot \check{E}(c_k)\right] + \varepsilon\left[\frac{\xi^2}{2}\right] \cdot |\check{E}(c_k)| + \varepsilon\left[\frac{\xi^2}{2} \cdot \check{E}(c_k)\right] + |\mathbf{o}(\xi)| \\
&\le \quad \left(\frac{1}{2} + \frac{1}{8} + \frac{0.323}{4} + \frac{\sqrt{2}}{32} + \right. \\
&\quad\quad \frac{0.323 \cdot \sqrt{2}}{16} + \frac{1}{32} + \frac{1}{512} + \\
&\quad\quad \left. \frac{1}{2^{25}} + \frac{1}{2^{28.5}} + \frac{1}{2^{29}} + \frac{1}{320}\right) \cdot ulp\left(\xi \cdot \hat{E}(c_k)\right) \\
&\le \quad 0.8149 \cdot ulp\left(\xi \cdot \hat{E}(c_k)\right) \\
&\overset{(A.8.2)}{\le} \quad \frac{0.8149}{16} \cdot ulp\,(E(x)) \\
&\le \quad 0.052 \cdot ulp\,(E(x)).
\end{aligned}$$

$\mathcal{QED}$

B EXP — Accuracy Statement and Proof

Let $\mu := ulp\,(1)$. Here is our main result of the Section:

Theorem B.1 *Using the algorithm and notation established in Section 8, we have*

$$\varepsilon\exp(x) \le 0.028 \cdot ulp\,(E(x)) \qquad \forall x \in [-\mathcal{B}, \mathcal{B}].$$

First some preparations.

Lemma B.2 *For $x \in (-\log 2, \log 2)$, we have*

$$ulp\,(E(x)) \le \frac{1}{2} ulp\,(\exp(x))\,.$$

Proof: For $x \in (-\log 2, 0)$, since $|E(x)| < \frac{1}{2}$, we have

$$ulp\,(E(x)) \le \frac{1}{4}\mu = \frac{1}{2} ulp\,(\exp(x))\,.$$

For $x \in (0, \log 2)$, since $|E(x)| < 1$, we have

$$ulp\,(E(x)) \le \frac{1}{2}\mu = \frac{1}{2} ulp\,(\exp(x))\,.$$

$\mathcal{QED}$

With the formula stated in Section 8.2, we have

Lemma B.3

$$\varepsilon\exp(x) \le 0.052 \cdot ulp\,(E(x)) \le 0.026 \cdot ulp\,(exp(x))\,.$$

Proof: The first inequality follows from Section A while the second from Lemma B.2. $\mathcal{QED}$

Lemma B.4 *Assume $x \in [-\mathcal{B}, \mathcal{B}] \setminus [-\log 2, \log 2]$. Recall that*

$$\xi := x - n \cdot \hat{\log} 2, \qquad |\xi| \le \frac{\hat{\log} 2}{2}.$$

Let

$$\delta := -n \cdot \check{\log} 2.$$

We have

1. $\frac{1}{2}\mu \le ulp\,(\exp(\xi))$,
2. $ulp\,(E(\xi)) \le \frac{1}{2} ulp\,(\exp(\xi))$, $|E(\xi)| \le \sqrt{2} - 1$,
3. $0 \le \delta \le \frac{1}{2^{10}}$,

4. $\exp(\xi) = 2^{-n} \cdot \exp(x) \cdot \exp(-\delta) \leq 2^{-n} \cdot \exp(x)$
5. $1 \leq E(\delta) \leq (1 + \frac{1}{2^{10}}) \cdot \delta \leq (1 + \frac{1}{2^{10}}) \cdot \frac{1}{2^{10}}$,
6. $\varepsilon\{E(\delta)\} \leq 0.552 \cdot ulp\,(E(\delta)) \leq \frac{0.552}{2^{10}} \cdot \mu \leq \frac{0.552}{2^9} \cdot ulp\,(\exp(\xi))$,
7. $|E(\xi)| \cdot \varepsilon\{E(\delta)\} \leq (\sqrt{2} - 1) \cdot \frac{0.552}{2^9} \cdot ulp\,(\exp(\xi))$,
8. $\varepsilon\{E(\xi)\} \cdot E(\delta) \leq (1 + \frac{1}{2^{10}}) \cdot \frac{0.552}{2^{10}} ulp\,(E(\xi)) \leq (1 + \frac{1}{2^{10}}) \cdot \frac{1}{4} \cdot \frac{0.552}{2^9} ulp\,(\exp(\xi))$
9. $\varepsilon E(\xi) \leq 0.052 \cdot ulp\,(E(\xi)) \leq 0.026 \cdot ulp\,(\exp(\xi))$,
10. *Finally, we have*

$$\begin{aligned} \varepsilon 2^{-n} \exp(x) &\leq \varepsilon E(\xi) + \varepsilon\{E(\delta)\} + \varepsilon\{E(\xi) \cdot E(\delta)\} \\ &\leq \left(0.026 + \left(\sqrt{2} + \frac{1}{4} \cdot \left(1 + \frac{1}{2^{10}}\right)\right) \cdot \frac{0.552}{2^9}\right) \cdot ulp\,(\exp(\xi)) \\ &\leq 0.028 \cdot ulp\,\left(2^{-n} \exp(x)\right). \end{aligned}$$

Proof: (B.4.1) and (B.4.2) are trivial. To prove (B.4.3), recall that by construction,

$$\begin{aligned} t - 2m &\geq 10, \\ n &< 2^m, \\ |\,\check{\log} 2| &\leq 2^m \cdot ulp\,(\hat{\log} 2) = 2^{m-1}\mu. \end{aligned}$$

Hence

$$0 \leq \delta := -n \cdot \check{\log} 2 < 2^m \cdot 2^{m-1}\mu = 2^{2m-t} \leq \frac{1}{2^{10}}$$

as desired. (B.4.4) follows from (B.4.3) and by way of construction. (B.4.5) follows from (B.4.4) and the inequality

$$e^t - 1 \leq t + t^2 \qquad \text{for } t \in [0, 1].$$

(B.4.6) follows from (B.4.2), (B.4.5) and Section A. (B.4.7) follows from (B.4.2) and (B.4.6). (B.4.8) follows from (B.4.2), (B.4.5) and (B.4.6). (B.4.9) follows from (B.4.2) and Section A. (B.4.10) follows from (B.4.4), (B.4.7), (B.4.8) and (B.4.9). $\mathcal{QED}$

With Lemma B.3 and Lemma B.4, the proof of Theorem B.1 is now complete.

C COSINE & SINE — Accuracy Statements and Proofs

Let $\mu := ulp\,(1)$. Here is our main result of the Section:

Theorem C.1 *For* $x \in \left[0, \frac{\pi}{2}\right)$,

$$\varepsilon \cos(x) \leq 0.0611 \cdot ulp\,(\cos(x))\,,$$
$$\varepsilon \sin(x) \leq 0.0600 \cdot ulp\,(\sin(x))\,.$$

First some preparations.

Lemma C.2 *For* m, $n > 0$, *let*

$$x_0 := \sqrt[n+1]{m \cdot 2^{\left\lfloor \frac{\tau^{n+1}}{m} \right\rfloor}}$$

where τ *is the number closest to but smaller than* 2 *in the floating-point arithmetic we are interested in, then*

$$ulp\left(\frac{x^{n+1}}{m}\right) \leq x_0 \cdot \frac{|x|^n}{m} ulp\,(x)\,.$$

In particular,

$$ulp\left(x^{n+1}\right) \leq \sqrt[n+1]{2^n} \cdot |x|^n ulp\,(x)\,.$$

Proof: It suffices to prove the inequality for $x \in [1, \tau]$. Let $x_0 \in [1, \tau]$ be the largest such x that $\frac{x^{n+1}}{m}$ is an integral power of 2. It is easy to see that

$$x_0 = \sqrt[n+1]{m \cdot 2^{\left\lfloor \frac{\tau^{n+1}}{m} \right\rfloor}}.$$

For $x \in [x_0, \tau]$, by choice of x_0,

$$\begin{aligned} ulp\left(\frac{x^{n+1}}{m}\right) = ulp\left(\frac{x_0^{n+1}}{m}\right) &= x_0 \cdot \frac{x_0^n}{m}\mu \\ &\leq x_0 \cdot \frac{x^n}{m}\mu = x_0 \cdot \frac{x^n}{m} ulp\,(x)\,; \end{aligned}$$

for $x \in [1, x_0)$,

$$ulp\left(\frac{x^{n+1}}{m}\right) \leq \frac{x^n}{m}\mu \leq x_0 \cdot \frac{x^n}{m}\mu = x_0 \cdot \frac{x^n}{m} ulp\,(x)\,.$$

$\mathcal{QED}$

Lemma C.3 *Let*

$$\mathcal{T}_0(x) \equiv 1, \quad \mathcal{T}_{n+1}(x) := 1 - \frac{x^2}{\alpha_{n+1}}\mathcal{T}_n(x), \quad \Sigma_n(x) := \mathcal{T}_n(x\sqrt{-1}).$$

If for all $n > 0$, α_n are positive integers and

$$\mathcal{T}_n \geq 0,$$

and if both $\mathcal{T}_n$ and Σ_n converges for all $|x| < 1$ as $n \to \infty$, then

$$\varepsilon\left\{\mathcal{T}_n(x)\right\} \leq \mu \cdot \left[\Sigma_n(x_k) - \frac{3}{4}\right] \qquad \forall\, |x| < x_k := \frac{1}{2^k}.$$

In particular, with the Q_n defined in Section 9.3.1, we have

$$\varepsilon\left\{Q_n(x)\right\} \leq \mu \cdot \left[\frac{120}{x_k^5}\left(\sinh(x_k) - x_k - \frac{x_k^3}{6}\right) - \frac{3}{4}\right] \qquad \forall\, |x| < x_k.$$

Proof: First notice that

$$\varepsilon\left\{\frac{x^2}{\alpha}\right\} \leq \frac{1}{2}ulp\left(\frac{x^2}{\alpha}\right) + \frac{\frac{1}{2}ulp\left(x^2\right)}{\alpha} \leq \frac{1}{2}\cdot\frac{x^2}{\alpha}\mu + \frac{1}{4}\cdot\frac{x_k^2}{\alpha}\mu \leq \frac{3}{4}\mu\frac{x_k^2}{\alpha}.$$

We now apply induction on n. For $n = 0$,

$$\varepsilon\left\{\mathcal{T}_0(x)\right\} = 0 \leq \frac{1}{4}\mu = \mu \cdot \left[\Sigma_0(x_k) - \frac{3}{4}\right].$$

Assume that the proposition holds true for n,

$$\begin{aligned}
\varepsilon\left\{\mathcal{T}_{n+1}(x)\right\} &\leq \frac{1}{4}\mu + \varepsilon\left\{\frac{x^2}{\alpha_{n+1}}\right\}\mathcal{T}_n(x) + \frac{x^2}{\alpha_{n+1}}\varepsilon\left\{\mathcal{T}_n(x)\right\} \\
&\leq \frac{1}{4}\mu + \frac{3}{4}\mu\frac{x_k^2}{\alpha_{n+1}} \cdot 1 + \frac{x_k^2}{\alpha_{n+1}} \cdot \mu\left[\Sigma_n(x_k) - \frac{3}{4}\right] \\
&= \frac{1}{4}\mu + \mu \cdot \frac{x_k^2}{\alpha_{n+1}}\Sigma_n(x_k) \\
&= \mu\left[\Sigma_{n+1}(x_k) - \frac{3}{4}\right]
\end{aligned}$$

as desired. $\mathcal{QED}$

Lemma C.4 *Let P_n, Q_n be defined as in Section 9.3.1*

1. $ulp\,(x) \leq x\mu \leq 2 \cdot ulp\,(\sin(x)) \qquad \forall\, x \in (0,1)$

2. For $x > 0$ we have

$$\varepsilon\left\{\frac{x^3}{6}\right\} \leq \frac{ulp\left(x^2\right)x + ulp\left(x^3\right)}{12} + \frac{1}{2}ulp\left(\frac{x^3}{6}\right),$$

$$\varepsilon\left\{\frac{x^5}{120}\right\} \leq \frac{ulp\left(x^2\right)x^3 + ulp\left(x^3\right)x^2 + ulp\left(x^4\right)x + ulp\left(x^5\right)}{240} + \frac{1}{2}ulp\left(\frac{x^5}{120}\right).$$

3. For $x \in [0, \arcsin(\frac{1}{4}))$,

$$\varepsilon\left\{\frac{x^3}{6}\right\} \leq \frac{\sqrt{2}+\sqrt[3]{4}+\sqrt[3]{6}}{6}x^2 ulp\left(\sin(x)\right) \leq 0.052 \cdot ulp\left(\sin(x)\right),$$

$$\varepsilon\left\{\frac{x^5}{120}\right\} \leq \frac{\sqrt{2}+\sqrt[3]{4}+\sqrt[4]{8}+\sqrt[5]{16}+\sqrt[5]{32}}{120}x^4 ulp\left(\sin(x)\right) \leq 0.000157 \cdot ulp\left(\sin(x)\right).$$

4. For $x \in [\arcsin(\frac{1}{4}), \frac{7}{16}]$, $\quad ulp\left(\sin(x)\right) \equiv \frac{1}{4}\mu$ and

$$\varepsilon\left\{\frac{x^3}{6}\right\} \leq \frac{7}{128}ulp\left(\sin(x)\right),$$

$$\varepsilon\left\{\frac{x^5}{120}\right\} \leq \frac{145}{131072}ulp\left(\sin(x)\right).$$

5. For $x \in [0, \frac{7}{16})$,

$$\varepsilon\left\{Q_n\right\} \leq 0.256\mu,$$

$$\frac{x^5}{120}\varepsilon\left\{Q_n\right\} \leq \left.\frac{0.256}{120}x^4\right|_{x=\frac{7}{16}} \cdot x\mu \leq 0.000157 \cdot ulp\left(\sin(x)\right).$$

6. For $x \in I_s := [0, \frac{7}{16})$,

$$\varepsilon\sin(x) \leq |P_n(x) - \sin(x)| + \varepsilon\left\{\frac{x^3}{6}\right\} + \varepsilon\left\{\frac{x^5}{120}\right\}|Q_n(x)| + \frac{x^5}{120}\varepsilon\left\{Q_n(x)\right\} \leq 0.05986 \cdot ulp\left(\sin(x)\right).$$

Proof: (C.4.1) and (C.4.2) are trivial.
(C.4.3) follows from (C.4.1), (C.4.2) and Lemma C.2.
(C.4.4) follows from (C.4.3) and the fact that $ulp\left(\sin(x)\right)$ stays at $\frac{1}{4}\mu$ throughout $[\arcsin(\frac{1}{4}), \frac{7}{16})$.
(C.4.5) follows from (C.4.1) and Lemma C.3.
(C.4.6) follows from (C.4.4), (C.4.5) and the inequalities $|Q_n(x)| < 1$ and $|P_n(x) - \sin(x)| \leq \frac{1}{256}ulp\left(\sin(x)\right)$. $\mathcal{QED}$

Lemma C.5 *Let U, V, $\mathcal{R}_c(x)$, $\mathcal{R}_s(x)$ be defined as in Section 2.1.1. Write*

$$\begin{aligned} A &:= U^2 - V, \\ A' &:= U^2 - (U+V), \\ B &:= U^2 + V^2 \,/ (x/2)^2 \,. \end{aligned}$$

Assume

$$\varepsilon\{p(x)\} \le 100 \cdot ulp\,(p(x))\,,$$

then for $|x| < \frac{5}{16}$ we have

1. $\dfrac{x^2}{700} \le p(x) \le \dfrac{x^2}{700}\left(1 + \dfrac{x^2}{9}\right)$
2. $\varepsilon\{p(x)\} \le \dfrac{1}{80}\mu, \quad 2.97 \le V \le U \le 3$
3. $\dfrac{A}{B} \le \dfrac{x^2}{6}, \quad \dfrac{A'}{B} \le \dfrac{x^2}{12}$
4. $|\mathcal{R}_s(x)| \le \dfrac{|x|^3}{6}, \quad \mathcal{R}_c(x) \le \dfrac{x^4}{24}$
5. $\varepsilon\{\mathcal{R}_s(x)\} \le |x|^3\mu, \quad \varepsilon\{\mathcal{R}_c(x)\} \le \dfrac{3}{8}x^4\mu.$

Proof: (C.5.1), (C.5.2), (C.5.3) and (C.5.4) are all straightforward. We now proceed to prove (C.5.5). Write

$$u := \frac{13}{40} + p(x), \qquad v := \frac{169}{80} + p(x).$$

We have

$$\begin{aligned} \varepsilon\{U\} \le\ & \frac{ulp\,(U)}{2} + \frac{1}{2}ulp\Big(\left(\frac{x}{2}\right)^2(\frac{1}{5} + p(x))\Big) + \frac{1}{2}ulp\Big(\left(\frac{x}{2}\right)^2\Big)(\frac{1}{5} + p(x)) + \\ & \left(\frac{x}{2}\right)^2\left(\frac{1}{2}ulp\left(\frac{1}{5} + p(x)\right) + \varepsilon\left\{\frac{1}{5}\right\} + \varepsilon\{p(x)\}\right) \\ \le\ & \mu + 2\cdot\left(\frac{1}{10} + \frac{1}{2}p(x)\right)\left(\frac{x}{2}\right)^2\mu + \left(\frac{x}{2}\right)^2\left(\frac{1}{2}\cdot\frac{1}{8} + \frac{1}{20} + \frac{1}{80}\right)\mu \\ =\ & \left(1 + u\left(\frac{x}{2}\right)^2\right)\mu, \\ \varepsilon\{V\} \le\ & \frac{ulp\,(V)}{2} + \frac{1}{2}ulp\Big(\left(\frac{x}{2}\right)^2(\frac{6}{5} + p(x))\Big) + \frac{1}{2}ulp\Big(\left(\frac{x}{2}\right)^2\Big)(\frac{6}{5} + p(x)) + \\ & \left(\frac{x}{2}\right)^2\left(\frac{1}{2}ulp\left(\frac{6}{5} + p(x)\right) + \varepsilon\left\{\frac{6}{5}\right\} + \varepsilon\{p(x)\}\right) \end{aligned}$$

$$
\begin{aligned}
&\le \mu + 2\cdot\left(\frac{3}{5}+\frac{1}{2}p(x)\right)\left(\frac{x}{2}\right)^2\mu + \left(\frac{x}{2}\right)^2\left(\frac{1}{2}\cdot 1+\frac{2}{5}+\frac{1}{80}\right)\mu \\
&= \left(1+v\left(\frac{x}{2}\right)^2\right)\mu, \\
\varepsilon\left\{U^2\right\} &\le \frac{1}{2}ulp\left(U^2\right)+2U\varepsilon\{U\} \le 4\mu + 2\cdot 3\cdot\left(1+u\left(\frac{x}{2}\right)^2\right)\mu \\
&= \left(10+6u\left(\frac{x}{2}\right)^2\right)\mu, \\
\varepsilon\{A\} &\le \frac{1}{2}ulp\left(U^2-V\right)+\varepsilon\left\{U^2\right\}+\varepsilon\{V\} \\
&\le 2\mu + \left(10+6u\left(\frac{x}{2}\right)^2\right)\mu + \left(1+v\left(\frac{x}{2}\right)^2\right)\mu \\
&= \left(13+(6u+v)\left(\frac{x}{2}\right)^2\right)\mu, \\
\varepsilon\{A'\} &\le 0+\varepsilon\left\{U^2\right\}+\frac{1}{2}\varepsilon\{U+V\}+\varepsilon\{U\}+\varepsilon\{V\} \\
&\le \left(14+(7u+v)\left(\frac{x}{2}\right)^2\right)\mu, \\
\varepsilon\left\{\frac{V^2}{(x/2)^2}\right\} &\le \frac{1}{2}ulp\left(\frac{V^2}{(x/2)^2}\right)+\frac{V^2}{(x/2)^2}\cdot\frac{\frac{1}{2}ulp\left((x/2)^2\right)}{(x/2)^2}+\frac{\varepsilon\left\{V^2\right\}}{(x/2)^2} \\
&\le \frac{V^2}{(x/2)^2}\mu + \frac{\mu}{(x/2)^2}\left(10+6v\left(\frac{x}{2}\right)^2\right), \\
\frac{\varepsilon\{B\}}{B} &\le \frac{ulp\,(B)}{2B}+\frac{\varepsilon\left\{U^2\right\}}{V^2/(x/2)^2}+\frac{\varepsilon\left\{V^2/(x/2)^2\right\}}{V^2/(x/2)^2} \\
&\le \frac{3}{2}\mu+\frac{\mu}{V^2}\left(10+(6v+10)\left(\frac{x}{2}\right)^2+6u\left(\frac{x}{2}\right)^4\right), \\
\frac{\varepsilon\{A\}}{B} &\le \frac{\mu}{V^2}\left(13\left(\frac{x}{2}\right)^2+(6u+v)\left(\frac{x}{2}\right)^4\right), \\
\varepsilon\left\{\frac{A}{B}\right\} &\le \frac{1}{2}ulp\left(\frac{A}{B}\right)+\frac{\varepsilon\{A\}}{B}+\frac{A}{B}\cdot\frac{\varepsilon\{B\}}{B} \\
&\le \frac{x^2}{3}\mu+\frac{x^2}{12}\mu\cdot\frac{1}{V^2}\left(59+(18u+15v+20)\left(\frac{x}{2}\right)^2+12u\left(\frac{x}{2}\right)^4\right) \\
&\le \frac{11}{12}x^2\mu, \\
\varepsilon\{\mathcal{R}_s\} &\le \frac{1}{2}ulp\,(\mathcal{R}_s)+|x|\cdot\varepsilon\left\{\frac{A}{B}\right\} \le |x|^3\mu; \\
\frac{\varepsilon\{A'\}}{B} &\le \frac{\mu}{V^2}\left(14+(7u+v)\left(\frac{x}{2}\right)^2\right)\left(\frac{x}{2}\right)^2,
\end{aligned}
$$

$$
\begin{aligned}
\varepsilon\left\{\frac{A'}{B}\right\} &\leq \frac{1}{2}ulp\left(\frac{A'}{B}\right) + \frac{\varepsilon\{A'\}}{B} + \frac{A'}{B}\cdot\frac{\varepsilon\{B\}}{B} \\
&\leq \frac{x^2}{6}\mu + \frac{x^2}{12}\mu\cdot\frac{1}{V^2}\left(52 + (21u+9v+10)\left(\frac{x}{2}\right)^2 + 6u\left(\frac{x}{2}\right)^4\right) \\
&\leq \frac{2}{3}x^2\mu, \\
\varepsilon\{\mathcal{R}_c\} &\leq \frac{1}{2}ulp(\mathcal{R}_c) + \frac{1}{2}ulp\left(\frac{x^2}{2}\right)\cdot\frac{A'}{B} + \frac{x^2}{2}\cdot\varepsilon\left\{\frac{A'}{B}\right\} \leq \frac{3}{8}x^4\mu.
\end{aligned}
$$

$\mathcal{QED}$

Lemma C.6 *Making use of Lemma C.5, it is now straightforward to compute the error bounds for regions II_s, III_s and II_c. The results are summarized in the following two tables. Table 7 presents in some detail as to how each individual error bound is computed. Table 8 gives the final error bounds as the sum of columns 3, 4 and 5 and are registered in the last column.*

$\mathcal{F}$	S	C	$\lvert\xi\rvert_{\max}$	$\frac{1}{2}ulp(\hat{S}\cdot\frac{\hat{\xi}^2}{2})+$ $\hat{S}\cdot\frac{1}{2}ulp(\frac{\xi^2}{2})$	$\frac{1}{2}ulp(\hat{C}\cdot\mathcal{R}_s)+$ $\hat{C}\cdot\varepsilon\{\mathcal{R}_s\}$	$\frac{1}{2}ulp(\hat{S}\cdot\mathcal{R}_c)+$ $\hat{S}\cdot\varepsilon\{\mathcal{R}_c\}$
sin	$\frac{\sqrt{15}}{8}$	$\frac{7}{8}$	$\cos^{-1}(\frac{7}{8})-\frac{7}{16}$	$\frac{1}{2}\cdot\frac{\mu}{1024}+$ $\frac{\sqrt{15}}{8}\cdot\frac{\mu}{1024}$	$2^{-16}\mu+$ $\frac{7}{8}\lvert\xi\rvert^3_{\max}\mu$	$2^{-23}\mu+$ $\frac{\sqrt{15}}{8}\cdot\frac{3}{8}\lvert\xi\rvert^4_{\max}\mu$
sin	$\frac{\sqrt{7}}{4}$	$\frac{3}{4}$	$\cos^{-1}(\frac{3}{4})-\frac{9}{16}$	$\frac{1}{2}\cdot\frac{\mu}{128}+$ $\frac{\sqrt{7}}{4}\cdot\frac{\mu}{256}$	$2^{-12}\mu+$ $\frac{3}{4}\lvert\xi\rvert^3_{\max}\mu$	$2^{-17}\mu+$ $\frac{\sqrt{7}}{4}\cdot\frac{3}{8}\lvert\xi\rvert^4_{\max}\mu$
cos	$\frac{\sqrt{3}}{2}$	$\frac{1}{2}$	$\frac{3}{4}-\frac{\pi}{6}$	$\frac{1}{2}\cdot\frac{\mu}{64}+$ $\frac{\sqrt{3}}{2}\cdot\frac{\mu}{128}$	$2^{-12}\mu+$ $\frac{1}{2}\lvert\xi\rvert^3_{\max}\mu$	$2^{-15}\mu+$ $\frac{\sqrt{3}}{2}\cdot\frac{3}{8}\lvert\xi\rvert^4_{\max}\mu$

Table 7: II_s, III_s, II_c (Part 1)

$\mathcal{F}$	*region*	$\varepsilon\left\{\hat{S}\cdot\frac{\hat{\xi}^2}{2}\right\}$	$\varepsilon\left\{\hat{C}\cdot\mathcal{R}_s\right\}$	$\varepsilon\left\{\hat{S}\cdot\mathcal{R}_c\right\}$	$ulp(\mathcal{F})$	$\frac{\varepsilon\mathcal{F}}{ulp(\mathcal{F})}$
sin	II_s	$.00097\mu$	$.00029\mu$	$.00001\mu$	$.25\mu$	$\frac{.000127}{.25} = .00508$
sin	III_s	$.00649\mu$	$.00333\mu$	$.00018\mu$	$.50\mu$	$\frac{.01000}{.50} = .02000$
cos	II_c	$.01458\mu$	$.00605\mu$	$.00089\mu$	$.50\mu$	$\frac{.02152}{.50} = .04304$

Table 8: II_s, III_s, II_c (Part 2)

Lemma C.7 *For $x \in I_c$, we have*

$$\varepsilon\cos(x) \leq 0.03942\cdot ulp(\cos(x)).$$

Proof: Note that

$$ulp\,(\cos(x)) \equiv \frac{1}{2}\mu \qquad \forall\, x \in I_c := \Big[0, \frac{5}{16}\Big).$$

Making use of (C.5.5), we have

$$\begin{aligned}
\varepsilon \cos(x) &\le \varepsilon \left\{\frac{x^2}{2}\right\} + \varepsilon \left\{\mathcal{R}_c\right\} \\
&\le \frac{1}{2} ulp \left(\frac{x^2}{2}\right) + \frac{3}{8} x^4 \mu \\
&\le \left(\frac{1}{64} + \frac{3}{8}\Big(\frac{5}{16}\Big)^4\right)\mu \\
&\le \left(\frac{1}{32} + \frac{3}{4}\Big(\frac{5}{16}\Big)^4\right) ulp\,(\cos(x))
\end{aligned}$$

as desired. $\mathcal{QED}$

Lemma C.8 *Let*

$$\xi := \frac{\hat{\pi}}{2} - x \quad \textit{where} \quad \frac{\hat{\pi}}{2} := \frac{\pi}{2} \textit{ chopped},$$

then

1. *For* $x \in III_c$, $\xi \cdot \mu \le 2.25 \cdot ulp\,(cos(x))$.
2. *For* $x \in III_c$, $\varepsilon \cos(x) \le 0.06103 \cdot ulp\,(\cos(x))$.
3. *For* $x \in IV_s$, $\varepsilon \sin(x) \le 0.04576 \cdot ulp\,(\sin(x))$.

Proof: Note that $x \in III_c \iff \xi \in (0, \frac{\hat{\pi}}{2} - \frac{3}{4}]$. Since $\sin(\xi)/\xi$ is a decreasing function of ξ over $(0, \frac{\hat{\pi}}{2} - \frac{3}{4}]$, we have

$$\frac{\sin(\xi)}{\xi} \ge \frac{\sin\big(\frac{\hat{\pi}}{2} - \frac{3}{4}\big)}{\frac{\hat{\pi}}{2} - \frac{4}{3}} \ge \frac{8}{9}, \quad \textit{or} \quad \xi \le 1.125 \sin(\xi).$$

Since

$$\frac{\check{\pi}}{2} := \frac{\pi}{2} - \frac{\hat{\pi}}{2} \ge 0,$$

$$\begin{aligned}
\xi \cdot \mu \le 1.125 \cdot \sin(\xi)\mu &\le 2.25 \cdot ulp\,(\sin(\xi)) \\
&= 2.25 \cdot ulp \left(\cos\Big(x + \frac{\check{\pi}}{2}\Big)\right) \le 2.25 \cdot ulp\,(\cos(x))
\end{aligned}$$

as desired. This proves (C.8.1).
Now, ignoring second and higher order terms, for $x \in III_c$,

$$\begin{aligned}
\varepsilon\cos(x) &\le \varepsilon\{\sin(\xi)\} + \frac{\xi^6}{6!}\cdot\frac{\check{\pi}}{2} \\
&\le 0.05986 \cdot ulp\,(\sin(\xi)) + \frac{\xi^5}{720}\xi\cdot\mu \\
&\le 0.05986 \cdot ulp\left(\sin\left(\xi + \frac{\check{\pi}}{2}\right)\right) + \frac{2.25}{720}\xi^5 ulp\,(\cos(x)) \\
&= \left(0.05986 + \frac{2.25}{720}\left(\frac{\hat{\pi}}{2} - \frac{3}{4}\right)^5\right) ulp\,(\cos(x)) \\
&\le 0.06103 \cdot ulp\,(\cos(x))
\end{aligned}$$

as desired. This proves (C.8.2).
For $x \in IV_s$, since $ulp\,(\sin(x)) \equiv \frac{1}{2}\mu = ulp\,(\cos(\xi))$,

$$\begin{aligned}
\varepsilon\sin(x) &\le \varepsilon\{\cos(\xi)\} + \frac{\xi^5}{5!}\cdot\frac{\check{\pi}}{2} \\
&\le 0.04304 \cdot ulp\,(\cos(\xi)) + \frac{\xi^5}{120}\mu \\
&\le 0.04304 \cdot ulp\,(\sin(x)) + \frac{\xi^5}{60} ulp\,(\sin(x)) \\
&= \left(0.04304 + \frac{1}{60}\left(\frac{\hat{\pi}}{2} - \frac{7}{8}\right)^5\right) ulp\,(\sin(x)) \\
&\le 0.04576 \cdot ulp\,(\sin(x))
\end{aligned}$$

as desired. This proves (C.8.3). $\mathcal{QED}$

In short, (C.4.6) covers I_s, Lemma C.6 covers II_s, III_s and II_c, Lemma C.7 covers I_c, (C.8.2) covers III_c, (C.8.3) covers IV_s, thus the proof of Theorem C.1 is now complete.

D LOG — Accuracy Statement and Proof

Let $\mu := ulp\,(1)$. Here is our main result of the Section:

Theorem D.1 *For* $x \in \left[\frac{1}{\sqrt{2}}, \sqrt{2}\right)$,

$$\varepsilon \log(x) \leq 0.052 \cdot ulp\,(\log(x))\,.$$

Lemma D.2 *For* $x \in \left[1 - \frac{1}{8}, 1 + \frac{1}{8}\right)$,

$$\varepsilon \left\{ \frac{(x-1)^2}{2} \right\} \leq \frac{1}{32} ulp\,(\log(x))\,.$$

Proof: For $x - 1 \in \left[-\frac{1}{8}, 0\right)$, since

$$|x-1| \leq |\log(x)|,$$

we have

$$ulp\left((x-1)^2\right) \leq ulp\,((x-1)\log(x)) \leq \frac{1}{8} ulp\,(\log(x))\,.$$

For $x - 1 \in \left(0, \frac{\sqrt{2}}{16}\right)$, since

$$x - 1 \leq \sqrt{2}\log(x),$$

we have

$$ulp\left((x-1)^2\right) \leq ulp\left(\sqrt{2}(x-1)\log(x)\right) \leq \frac{1}{8} ulp\,(\log(x))\,.$$

For $x - 1 \in [\frac{\sqrt{2}}{16}, \frac{1}{8})$,

$$ulp\left((x-1)^2\right) \leq \frac{1}{128}\mu = \frac{1}{8} \cdot \frac{1}{16}\mu = \frac{1}{8} ulp\,(\log(x))\,.$$

$\mathcal{QED}$

Lemma D.3 *For* $x \in \left[1 - \frac{1}{8}, 1 + \frac{1}{8}\right)$,

$$\varepsilon \left\{ \frac{(x-1)^3}{2(x+1)} \right\} \leq 0.009439 \cdot ulp\,(\log(x))\,.$$

Proof:

$$\begin{aligned} \varepsilon \left\{ \frac{(x-1)^3}{2(x+1)} \right\} &\leq \frac{1}{4} ulp\left(\frac{(x-1)^3}{x+1}\right) + \frac{ulp\left((x-1)^3\right)}{4(x+1)} + \\ &\quad \left|\frac{x-1}{x+1}\right| \frac{ulp\left((x-1)^2\right)}{4} + \left|\frac{(x-1)^3}{x+1}\right| \frac{ulp\,(x+1)}{4(x+1)} \\ &=: \epsilon_1 + \epsilon_2 + \epsilon_3 + \epsilon_4. \end{aligned}$$

By using a simple tabulation we find that, relative to $ulp\,(\log(x))$, ϵ_1, ϵ_2, ϵ_3 achieved their maximum at $\exp\big(-\big(\frac{1}{8}\big)^-\big)$ and ϵ_4 achieved its maximum at $\big(1+\frac{1}{8}\big)^-$. Since

$$\begin{aligned}
\epsilon_1\left(\exp\left(-\left(\frac{1}{8}\right)^-\right)\right) &= \frac{1}{4}\cdot\frac{1}{128}ulp\,(\log(x))\,,\\
\epsilon_2\left(\exp\left(-\left(\frac{1}{8}\right)^-\right)\right) &\le \frac{1}{4}\cdot\frac{1}{120.4798}ulp\,(\log(x))\,,\\
\epsilon_3\left(\exp\left(-\left(\frac{1}{8}\right)^-\right)\right) &\le \frac{1}{4}\cdot\frac{1}{128.1666}ulp\,(\log(x))\,,\\
\epsilon_4\left(\left(1+\frac{1}{8}\right)^-\right) &\le \frac{1}{4}\cdot\frac{1}{72.25}ulp\,(\log(x))\,,
\end{aligned}$$

the result follows. $\mathcal{QED}$

Lemma D.4 *For $x \in \left[\frac{1}{\sqrt{2}}, \sqrt{2}\right)$, using the u, ρ, $\mathcal{R}_k$ and $\mathcal{R}$ defined as in Section 10.1, we have*

1. $ulp\,(A_n) \equiv \mu, \quad ulp\,(B_n) \equiv \frac{1}{2}\mu.$

2. $u > \cdots > \mathcal{R}_2(u) > \mathcal{R}_1(u) \equiv \mathcal{R}(u) = \dfrac{2\rho}{\log(x) - 2\rho}.$

3. $\varepsilon\{\rho\} \le \frac{1}{2}ulp\,(\rho) + |\rho| \cdot \dfrac{\frac{1}{2}ulp\,(x+x_0)}{x+x_0} \le \frac{1}{2}(ulp\,(\rho) + |\rho|\cdot\mu).$

4. $\varepsilon\{u\} \le \frac{1}{2}ulp\,(u) + u \cdot \dfrac{\varepsilon\{\rho^2\}}{\rho^2} \le \frac{1}{2}ulp\,(u) + \frac{5}{2}u\cdot\mu.$

5. *Let*

$$\begin{aligned}
\alpha &:= \frac{1}{2}ulp\,(u) + \varepsilon\{u\} + \mu + \frac{1}{2}ulp\left(\frac{B_1}{\mathcal{R}}\right) + \frac{\mu}{4\mathcal{R}},\\
\beta &:= \frac{B_1}{\mathcal{R}^2}.
\end{aligned}$$

We have for all $k > 0$,

$$\varepsilon\{\mathcal{R}_k\} \le \alpha + \beta\cdot\varepsilon\{\mathcal{R}_{k+1}\}$$

and thus

$$\varepsilon\{\mathcal{R}\} \le \frac{\alpha}{1-\beta}.$$

6. With the α, β defined as in 5, we have

$$\begin{aligned}\varepsilon\left\{\frac{2\rho}{\mathcal{R}(u)}\right\} &\le \frac{1}{2}ulp\left(\frac{2\rho}{\mathcal{R}}\right) + \frac{2\varepsilon\{\rho\}}{\mathcal{R}} + \frac{2|\rho|}{\mathcal{R}^2}\cdot\varepsilon\{\mathcal{R}\} \\ &\le \frac{1}{2}ulp\left(\frac{2\rho}{\mathcal{R}}\right) + \frac{ulp(\rho)+|\rho|\cdot\mu}{\mathcal{R}} + \frac{2|\rho|}{\mathcal{R}^2}\cdot\frac{\alpha}{1-\beta} \\ &=: \tau_1+\tau_2+\tau_3.\end{aligned}$$

Proof: (D.4.1) and (D.4.2) follow directly from definition. (D.4.3) and (D.4.4) are straightforward. (D.4.5) follows from (D.4.1) through (D.4.4) and the fact that $\varepsilon\{\mathcal{R}_k\}$ always stays bounded. (D.4.6) follows from (D.4.4) and (D.4.5). $\mathcal{QED}$

Lemma D.5 *For* $x \in \left[1-\frac{1}{8}, 1+\frac{1}{8}\right)$,

1. $|x-1| \le \dfrac{1}{8}, \quad |\rho| = \left|\dfrac{x-1}{x+1}\right| \le \dfrac{1}{15}, \quad u = \dfrac{3}{\rho^2} \ge 675, \quad \mathcal{R}(u) \ge 673.$

2. $\varepsilon\left\{\dfrac{2\rho}{\mathcal{R}(u)}\right\} \le 0.011180\cdot ulp(\log(x)).$

Proof: Making use of Lemma D.4 and note that

$$\tau_3 \le \frac{2|\rho|}{\mathcal{R}^2 - B_1}\left(ulp(u) + \left(\frac{5}{2}u + 1 + \frac{1}{4096} + \frac{1}{2692}\right)\mu\right),$$

we once again resort to tabulation. Relative to $ulp(\log(x))$, we find that

$$\tau_1+\tau_2+\tau_3$$

achieves its maximum at $x = \exp\left(-\left(\frac{1}{8}\right)^-\right)$ where $ulp(\log(x)) = \frac{1}{16}\mu$. Since

$$\begin{aligned}\tau_1\left(\exp\left(-\left(\frac{1}{8}\right)^-\right)\right) &\le 0.0009765625\cdot ulp(\log(x)), \\ \tau_2\left(\exp\left(-\left(\frac{1}{8}\right)^-\right)\right) &\le 0.0019509242\cdot ulp(\log(x)), \\ \tau_3\left(\exp\left(-\left(\frac{1}{8}\right)^-\right)\right) &\le 0.0082518444\cdot ulp(\log(x)),\end{aligned}$$

the result follows. $\mathcal{QED}$

Lemma D.2, D.3 and D.5 can now be combined into:

Lemma D.6 *For* $x \in \left[1-\frac{1}{8}, 1+\frac{1}{8}\right)$, *we have*

$$\varepsilon\log(x) \le 0.052\cdot ulp(\log(x)).$$

For the rest of this Section, assume $x_0 = c_k \neq 1$, $x \in [b_{k-1}, b_k)$. Recall that

$$\xi := \frac{x - x_0}{x_0}.$$

Lemma D.7

$$\varepsilon\{\xi\} \leq \frac{1}{32} ulp\left(\log(x)\right).$$

Proof: This claim follows directly from the way x_0 was selected, cf. Section 10.3.1 $\mathcal{QED}$

Lemma D.8

$$\varepsilon\left\{\frac{\xi^2}{2}\right\} \leq \frac{3}{2^{11}} ulp\left(\log(x)\right).$$

Proof:

$$\begin{aligned}
\varepsilon\left\{\frac{\xi^2}{2}\right\} &\leq \frac{ulp\left(\xi^2\right)}{4} + |\xi| \cdot \varepsilon\{\xi\} \\
&\leq \frac{1}{4} \cdot \frac{1}{32} ulp(\xi) + \frac{1}{32} \cdot \frac{1}{32} ulp\left(\log(x)\right) \\
&\leq \left(\frac{1}{4} \cdot \frac{1}{32} \cdot \frac{1}{16} + \frac{1}{32} \cdot \frac{1}{32}\right) ulp\left(\log(x)\right).
\end{aligned}$$

$\mathcal{QED}$

Lemma D.9 *Recall that* $\rho := \dfrac{x - x_0}{x + x_0}$. *We have*

1. $|\rho| \leq |\xi| \leq \dfrac{1}{32}$.
2. $ulp(\rho) \leq ulp(\xi) \leq \dfrac{1}{16} ulp\left(\log(x)\right)$,
3. $\varepsilon\{\rho\} \leq \dfrac{ulp(\rho) + |\rho| \cdot \mu}{2} \leq \dfrac{ulp(\xi) + |\xi| \cdot \mu}{2} \leq \dfrac{3}{32} ulp\left(\log(x)\right)$.
4. $\varepsilon\left\{\dfrac{\rho}{2}\xi^2\right\} \leq \dfrac{7}{2^{16}} ulp\left(\log(x)\right)$.

Proof: (D.9.1) and (D.9.2) hold true by construction. (D.9.3) follows from (D.9.1) and (D.9.2). As for (D.9.4),

$$\begin{aligned}
\varepsilon\left\{\frac{\rho}{2}\xi^2\right\} &\leq \frac{1}{2} ulp\left(\frac{\rho}{2}\xi^2\right) + \frac{1}{2}\varepsilon\{\rho\} \cdot \xi^2 + \frac{|\rho|}{2} \cdot \varepsilon\left\{\xi^2\right\} \\
&\leq \left(\frac{1}{4} \cdot \frac{1}{2^{10}} \cdot \frac{1}{2^4} + \frac{1}{2} \cdot \frac{3}{32} \cdot \frac{1}{2^{10}} + \frac{1}{2} \cdot \frac{1}{2^5} \cdot \frac{3}{2^{10}}\right) ulp\left(\log(x)\right) \\
&\leq \frac{7}{2^{16}} ulp\left(\log(x)\right)
\end{aligned}$$

as desired. $\mathcal{QED}$

Lemma D.10 *For* $\left([\frac{1}{\sqrt{2}}, 1-\frac{1}{8}) \cup [1+\frac{1}{8}, \sqrt{2})\right)$, *We have*

1. $u \geq 3072, \quad \mathcal{R}(u) \geq 3070,$

2. $\dfrac{u}{\mathcal{R}} \leq 1 + \dfrac{A_1 + \dfrac{B_1}{\mathcal{R}}}{\mathcal{R}} \leq 1.000586385,$

3. $\varepsilon \left\{ \dfrac{2\rho}{\mathcal{R}(u)} \right\} \leq 0.00037681 \cdot ulp\left(\log(x)\right).$

Proof: (D.10.1) is trivial. (D.10.2) holds true by construction. As for (D.10.3), we make use of Lemma D.4, Lemma D.9, (D.10.1) and (D.10.2) to find

$$\begin{aligned}
\tau_1 &\leq \frac{1}{2^{11}} \cdot \frac{1}{2^4} ulp\left(\log(x)\right), \\
\tau_2 &\leq 2 \cdot \frac{3}{32} \cdot \frac{1}{3070} ulp\left(\log(x)\right), \\
\tau_3 &\leq \frac{2|\rho| \cdot \mu}{\mathcal{R}^2 - B_1}\left(3.5021 \cdot \mathcal{R} + 1 + \frac{1}{2^{14}} + \frac{1}{12280}\right) \\
&\leq \frac{2 \cdot 2}{16} \cdot 0.001141 \cdot ulp\left(\log(x)\right).
\end{aligned}$$

Summing the above 3 bounds up and (D.10.3) follows. $\mathcal{QED}$

Lemma D.7, D.8, D.9 and D.10 can therefore be combined into:

Lemma D.11 *For* $\left([\frac{1}{\sqrt{2}}, 1-\frac{1}{8}) \cup [1+\frac{1}{8}, \sqrt{2})\right)$, *we have*

$$\varepsilon \log(x) \leq 0.0332 \cdot ulp\left(\log(x)\right).$$

With Lemma D.6 and D.11, the proof of Theorem D.1 is now complete.

E ATAN — Accuracy Statement and Proof

Let $\mu := ulp\,(1)$. Here is our main result of the Section:

Theorem E.1 *For all $x \geq 0$,*

$$\varepsilon\,\text{atan}(x) \leq 0.048 \cdot ulp\,(\text{atan}(x))\,.$$

Lemma E.2 *Using the u and $\mathcal{R}$ defined in 5.1, for all $u \geq 57$, we have*

$$\varepsilon\,\{\mathcal{R}(u)\} \leq \frac{u^2}{u^2-1}\Big(\frac{1}{2}ulp\,(u+1.8) + \varepsilon\,\{u\} + 1.01\mu\Big).$$

Proof: Observe that

$$\begin{array}{ll} u \leq \mathcal{R}_k(u) \leq u+1.8, & k>0, \\ 1.8 =: A_1 \geq A_n \searrow A_\infty := 1.5, & ulp\,(A_n) \equiv \mu, \\ \dfrac{108}{175} =: B_1 \geq B_n \searrow B_\infty := \dfrac{9}{16}, & ulp\,(B_n) \equiv \dfrac{1}{2}\mu. \end{array}$$

Thus,

$$\begin{aligned}
\varepsilon\,\{\mathcal{R}_k\} &\leq \frac{1}{2}ulp\,(\mathcal{R}) + \varepsilon\,\{u\} + \varepsilon\left\{A_k - \frac{B_k}{\mathcal{R}_{k+1}}\right\} \\
&\leq \frac{1}{2}ulp\,(\mathcal{R}) + \varepsilon\,\{u\} + \frac{1}{2}ulp\left(A_k - \frac{B_k}{\mathcal{R}_{k+1}}\right) + \frac{1}{2}ulp\,(A_k) + \varepsilon\left\{\frac{B_k}{\mathcal{R}_{k+1}}\right\} \\
&\leq \frac{1}{2}ulp\,(\mathcal{R}) + \varepsilon\,\{u\} + \mu + \frac{1}{2}ulp\Big(\frac{B_k}{\mathcal{R}_{k+1}}\Big) + \frac{ulp\,(B_k)}{2\mathcal{R}_{k+1}} + \frac{B_k}{\mathcal{R}_{k+1}^2}\cdot\varepsilon\,\{\mathcal{R}_{k+1}\} \\
&\leq \frac{1}{2}ulp\,(u+1.8) + \varepsilon\,\{u\} + \left(1 + \frac{B_k+0.5}{2u}\right)\mu + \frac{1}{u^2}\cdot\varepsilon\,\{\mathcal{R}_{k+1}\} \\
&\leq \frac{1}{2}ulp\,(u+1.8) + \varepsilon\,\{u\} + 1.01\mu + \frac{1}{u^2}\cdot\varepsilon\,\{\mathcal{R}_{k+1}\}\,.
\end{aligned}$$

Since $\varepsilon\,\{\mathcal{R}_k\}$ stays bounded, the result follows from a simple limiting argument. $\mathcal{QED}$

Lemma E.3 *For $x \in [0, 0.2294)$, let $u := \dfrac{3}{x^2}$. We have*

1. $\varepsilon\,\{u\} \leq \dfrac{1}{2}\Big(ulp\,(u) + u\cdot\dfrac{ulp\,(x^2)}{x^2}\Big)$,
2. $\dfrac{1}{2}ulp\left(\dfrac{x}{\mathcal{R}}\right) \leq \dfrac{1}{128}ulp\,(\text{atan}(x))$,
3. $\dfrac{|x|}{\mathcal{R}^2}\cdot\varepsilon\,\{\mathcal{R}\} \leq 0.035\cdot ulp\,(\text{atan}(x))$,

4. $\varepsilon\,\mathrm{atan}(x) \le \varepsilon\left\{\frac{x}{\mathcal{R}}\right\} \le 0.043 \cdot \mathit{ulp}\,(\mathrm{atan}(x))$.

Proof: (E.3.1) is trivial. (E.3.2) follows from $\frac{x}{\mathcal{R}} = x - \mathrm{atan}(x)$ and a simple tabulation relative to $\mathit{ulp}\,(\mathrm{atan}(x))$. (E.3.3) follows from Lemma E.2, (E.3.1) and a simple tabulation relative to $\mathit{ulp}\,(\mathrm{atan}(x))$; in fact the expression achieves its maximum at $x = \frac{\sqrt{3}}{8}$. (E.3.4) follows from (E.3.2) and (E.3.3). $\mathcal{QED}$

Lemma E.4 *For $x \in [b_{k-1}, b_k)$, let $x_0 = c_k$, $\xi := \frac{x - x_0}{1 + x_0 \cdot x}$ and $u := \frac{3}{\xi^2}$. We have*

1. $\varepsilon\{u\} \le u \cdot \mu + \frac{2u}{|\xi|} \cdot \varepsilon\{\xi\}$,
2. $\varepsilon\{\mathcal{R}\} \le \frac{u^2}{u^2-1}(1.5u + 1.91) \cdot \mu + \frac{2u^3}{|\xi| \cdot (u^2-1)} \cdot \varepsilon\{\xi\}$,
3. $\varepsilon\left\{\frac{\xi}{\mathcal{R}}\right\} \le \left(\frac{1}{2u} + \frac{1.5u + 1.91}{u^2-1}\right) \cdot |\xi| \cdot \mu + \left(\frac{1}{u} + \frac{2u}{u^2-1}\right) \cdot \varepsilon\{\xi\}$,
4. $\varepsilon\left\{\frac{\xi}{\mathcal{R}}\right\} \le 0.0004 \cdot \mathit{ulp}\,(\mathrm{atan}(x))$,
5. $\varepsilon\,\mathrm{atan}(x) \le \varepsilon\{\xi\} + \varepsilon\left\{\frac{\xi}{\mathcal{R}}\right\} \le 0.048 \cdot \mathit{ulp}\,(\mathrm{atan}(x))$.

Proof: (E.4.1) is trivial. (E.4.2) follows from Lemma E.2 and (E.4.1). (E.4.3) follows from (E.4.2) and the fact that $u \le \mathcal{R}$. (E.4.4) follows from (E.4.3) and the inequalities

$$\begin{aligned} |\xi| \cdot \mu &\le \frac{1}{16}\mathit{ulp}\,(\mathrm{atan}(x))\,, \\ \varepsilon\{\xi\} &\le \frac{3}{64}\mathit{ulp}\,(\mathrm{atan}(x))\,, \\ u &\ge 768. \end{aligned}$$

(E.4.5) follows from (E.4.4). $\mathcal{QED}$

Lemma E.5 *For $x \ge b_N := 10.125$, let $\xi := -\frac{1}{x}$, $u := \frac{3}{\xi^2}$. We have*

1. $\mathit{ulp}\,(\mathrm{atan}(x)) \equiv \mu, \quad |\xi| \le \frac{1}{10.125}, \quad u \ge 300$,
2. $|\xi| \cdot \mu \le \frac{1}{10.125} \cdot \mathit{ulp}\,(\mathrm{atan}(x))$,

3. $\varepsilon\{\xi\} \leq \frac{1}{32} ulp(\mathrm{atan}(x))$,

4. $\varepsilon\left\{\frac{\xi}{\mathcal{R}}\right\} \leq 0.001 \cdot ulp(\mathrm{atan}(x))$,

5. $\varepsilon\,\mathrm{atan}(x) \leq \varepsilon\{\xi\} + \varepsilon\left\{\frac{\xi}{\mathcal{R}}\right\} \leq 0.033 \cdot ulp(\mathrm{atan}(x))$.

Proof: (E.5.1) is trivial. (E.5.2) and (E.5.3) follow from (E.5.1). (E.5.4) follows from Lemma E.4, (E.5.2) and (E.5.3). (E.5.5) follows from (E.5.3) and (E.5.4). $\mathcal{QED}$

With Lemma E.3, Lemma E.4 and Lemma E.5, the proof of Theorem E.1 is now complete.

References

[1] R. P. Brent, a FORTRAN Multiple-Precision Arithmetic Package, *ACM Transactions on Mathematical Software,* 4, no. 1, March 1978, pp. 57-70.

[2] W. Cody and W. Waite, *Software Manual for the Elementary Functions,* Prentice-Hall, Englewood Cliffs, N.J., 1980.

[3] IMSL Inc., Tests of Elementary Functions, *internal report,* IMSL Inc., Houston, Texas, 1983.

[4] W. Kahan, Continued Fractions for *a priori* Computation of Elementary Transcendental Functions, *unpublished manuscript,* March 1988.

[5] P-T P. Tang, Accurate and Efficient Testing of the Exponential and Logarithm Functions, *pre-print,* August 1988.