Lahore February 15, 2022

University of the Punjab, Department of Mathematics
National Mathematical Society of Pakistan
PU-NMS International Schools Series
for Students and Faculty
Mathematics for Computer Science.

# Towards exact rounding of the elementary functions, an application of Diophantine approximation to scientific computing and validated numerics.

*Michel Waldschmidt*

Professeur Émérite, Sorbonne Université,
Institut de Mathématiques de Jussieu, Paris
http://www.imj-prg.fr/~michel.waldschmidt/

# Abstract

The first launch of Ariane 5 in June 1996 gave rise to what has probably been the most expensive computer mistake in the world. It was due to an arithmetic overflow.

Computers are going to play an increasing role in our life. It would be too dangerous to use them as black boxes.

In theoretical computer science, validated scientific computing (arithmetic. reliability, accuracy, and speed) includes the question of exact rounding of the elementary functions, where results of Diophantine approximation are needed.

https://en.wikipedia.org/wiki/Ariane_5

Ariane 5 is a European heavy-lift launch vehicle that is part of the Ariane rocket family, an expendable launch system designed by the Centre national d'études spatiales (CNES). It is used to deliver payloads into geostationary transfer orbit (GTO) or low Earth orbit (LEO).

# Ariane 5

Ariane 5 rockets are manufactured under the authority of the European Space Agency (ESA) and the French spatial agency Centre National d'Etudes Spatiales. Airbus Defence and Space is the prime contractor for the vehicles, leading a consortium of other European contractors.

# Ariane 5 explosion (15")

https://www.youtube.com/watch?v=kYUrqdUyEpI

# Ariane 5 rocket launch explosion (4'26")

https://www.youtube.com/watch?v=PK_yguLapgA

The rocket was using software from Ariane 4 and due to it's 5 times faster acceleration there was an arithmetic overflow.

https://en.wikipedia.org/wiki/Ariane_5

Ariane 5's first test flight (Ariane 5 Flight 501) on 4 June 1996 failed, with the rocket self-destructing 37 seconds after launch because of a malfunction in the control software. A data conversion from 64-bit floating point value to 16-bit signed integer value to be stored in a variable representing horizontal bias caused a processor trap (operand error) because the floating point value was too large to be represented by a 16-bit signed integer. The software was originally written for the Ariane 4 where efficiency considerations (the computer running the software had an 80 % maximum workload requirement) led to four variables being protected with a handler while three others, including the horizontal bias variable, were left unprotected because it was thought that they were "physically limited or that there was a large margin of safety".

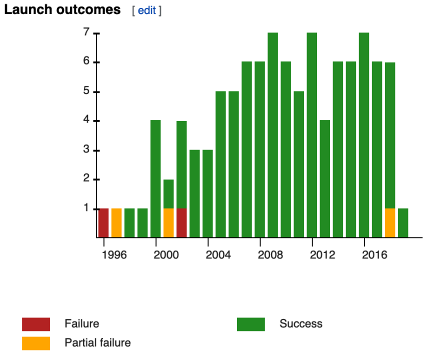
https://en.wikipedia.org/wiki/Ariane_5

The software, written in Ada, was included in the Ariane 5 through the reuse of an entire Ariane 4 subsystem despite the fact that the particular software containing the bug, which was just a part of the subsystem, was not required by the Ariane 5 because it has a different preparation sequence than the Ariane 4.

- Wired.com : "History's Worst Software Bugs" (Retrieved 3 September 2009)
- "Ariane 5 Flight 501 Failure, Report by the Inquiry Board". http://esamultimedia.esa.int/docs/esa-x-1819eng.pdf

# Launch statistics https://en.wikipedia.org/wiki/Ariane_5

Ariane 5 rockets have accumulated 103 launches since 1996, 98 of which were successful, yielding a 95.1% success rate. Between April 2003 and December 2017, Ariane 5 flew 82 consecutive missions without failure, but the rocket suffered a partial failure in January 2018.



# NASA's Mars Climate Orbiter Disaster

On September 23, 1999, NASA lost a $ 125 million Mars orbiter because a Lockheed Martin engineering team used English units of measurement while the agency's team used the more conventional metric system for a key spacecraft operation

http://edition.cnn.com/TECH/space/9909/30/mars.metric.02/

https://www.simscale.com/blog/2017/12/nasa-mars-climate-orbiter-metric/

# Avoiding accidents

Accidents need need to be avoided as much as possible.

Boeing 737 MAX 8

Flight 610 Lion Air
October 29, 2018

Flight 302 Ethiopian Airlines
March 10, 2019

https://fr.wikipedia.org/wiki/Vol_610_Lion_Air
https://fr.wikipedia.org/wiki/Vol_302_Ethiopian_Airlines

# No Black Box

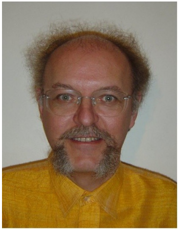
Computers will play an increasing role.
You need to understand fully all what they are doing.

IBM Releases “Black Box”
Breaker on IBM Cloud

https://www.cbronline.com/news/ai-bias-ibm

$$u_0 = 1,\, u_1 = (1-\sqrt{5})/2, \qquad u_n = u_{n-1} + u_{n-2}$$

Question : compute $u_{100}$.



Pierre Arnoux

$$\frac{1-\sqrt{5}}{2} = -0.61803398874989484820458683436 5\ldots$$

https://oeis.org/A001622

# Leonardo Pisano (Fibonacci)

Fibonacci sequence $(F_n)_{n\geq 0}$

$0,\ 1,\ 1,\ 2,\ 3,\ 5,\ 8,\ 13,\ 21,$

$34,\ 55,\ 89,\ 144,\ 233,\dots$

is defined by

$$F_0 = 0,\ F_1 = 1,$$

$$F_n = F_{n-1} + F_{n-2} \quad (n \geq 2).$$

http://oeis.org/A000045

Leonardo Pisano (Fibonacci)

(1170–1250)

# Excel file Column $A$ : $n$ Column $B$ : $u_n$

$$u_0 = 1,\ u_1 = (1-\sqrt{5})/2, \qquad u_n = u_{n-1} + u_{n-2}$$

| | A | B |
|---|---|---|
| 1 | 0 | 1 |
| 2 | 1 | `=(1-RACINE(5))/2` |

| | A | B |
|---|---|---|
| 1 | 0 | 1 |
| 2 | 1 | -0.618034 |

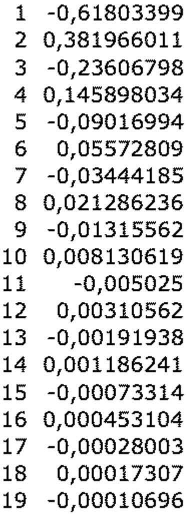
| | A | B |
|---|---|---|
| 1 | 0 | 1 |
| 2 | 1 | -0.618034 |
| 3 | =1+A2 | =B1+B2 |

| | A | B |
|---|---|---|
| 1 | 0 | 1 |
| 2 | 1 | -0.618034 |
| 3 | 2 | 0.38196601 |

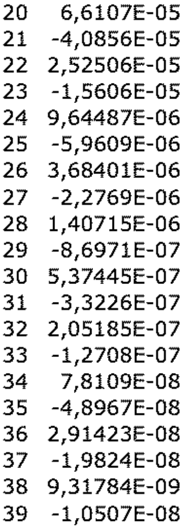
Copy $A3$ $B3$ down

`The poor man computer system`

# Excel file : $u_1$ to $u_{39}$



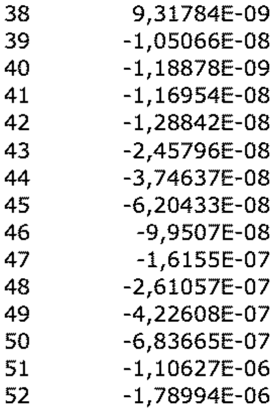
| | |
|---|---|
| 1 | -0,61803399 |
| 2 | 0,381966011 |
| 3 | -0,23606798 |
| 4 | 0,145898034 |
| 5 | -0,09016994 |
| 6 | 0,05572809 |
| 7 | -0,03444185 |
| 8 | 0,021286236 |
| 9 | -0,01315562 |
| 10 | 0,008130619 |
| 11 | -0,005025 |
| 12 | 0,00310562 |
| 13 | -0,00191938 |
| 14 | 0,001186241 |
| 15 | -0,00073314 |
| 16 | 0,000453104 |
| 17 | -0,00028003 |
| 18 | 0,00017307 |
| 19 | -0,00010696 |



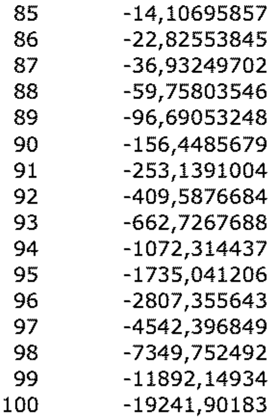
| | |
|---|---|
| 20 | 6,6107E-05 |
| 21 | -4,0856E-05 |
| 22 | 2,52506E-05 |
| 23 | -1,5606E-05 |
| 24 | 9,64487E-06 |
| 25 | -5,9609E-06 |
| 26 | 3,68401E-06 |
| 27 | -2,2769E-06 |
| 28 | 1,40715E-06 |
| 29 | -8,6971E-07 |
| 30 | 5,37445E-07 |
| 31 | -3,3226E-07 |
| 32 | 2,05185E-07 |
| 33 | -1,2708E-07 |
| 34 | 7,8109E-08 |
| 35 | -4,8967E-08 |
| 36 | 2,91423E-08 |
| 37 | -1,9824E-08 |
| 38 | 9,31784E-09 |
| 39 | -1,0507E-08 |

# Excel (continued)

$$u_{100} = -19\,241.901\,833\,167\ldots$$



| | |
|---|---|
| 38 | 9,31784E-09 |
| 39 | -1,05066E-08 |
| 40 | -1,18878E-09 |
| 41 | -1,16954E-08 |
| 42 | -1,28842E-08 |
| 43 | -2,45796E-08 |
| 44 | -3,74637E-08 |
| 45 | -6,20433E-08 |
| 46 | -9,9507E-08 |
| 47 | -1,6155E-07 |
| 48 | -2,61057E-07 |
| 49 | -4,22608E-07 |
| 50 | -6,83665E-07 |
| 51 | -1,10627E-06 |
| 52 | -1,78994E-06 |



| | |
|---|---|
| 85 | -14,10695857 |
| 86 | -22,82553845 |
| 87 | -36,93249702 |
| 88 | -59,75803546 |
| 89 | -96,69053248 |
| 90 | -156,4485679 |
| 91 | -253,1391004 |
| 92 | -409,5876684 |
| 93 | -662,7267688 |
| 94 | -1072,314437 |
| 95 | -1735,041206 |
| 96 | -2807,355643 |
| 97 | -4542,396849 |
| 98 | -7349,752492 |
| 99 | -11892,14934 |
| 100 | -19241,90183 |

# Exact value of $u_n$

Observations : The signs of $u_n$ alternate, the absolute value is decreasing.

Set $\widetilde{\Phi} = (1-\sqrt{5})/2$. Notice that $\widetilde{\Phi}$ is a root of $X^2 - X - 1$, the other root is $\Phi = (1+\sqrt{5})/2$, the golden ratio.

From $\widetilde{\Phi}^n = \widetilde{\Phi}^{n-1} + \widetilde{\Phi}^{n-2}$ with $u_0 = 1$, $u_1 = \widetilde{\Phi}$, we deduce by induction $u_n = \widetilde{\Phi}^n$.

# Exact value of $u_{39}$

Numerical values :

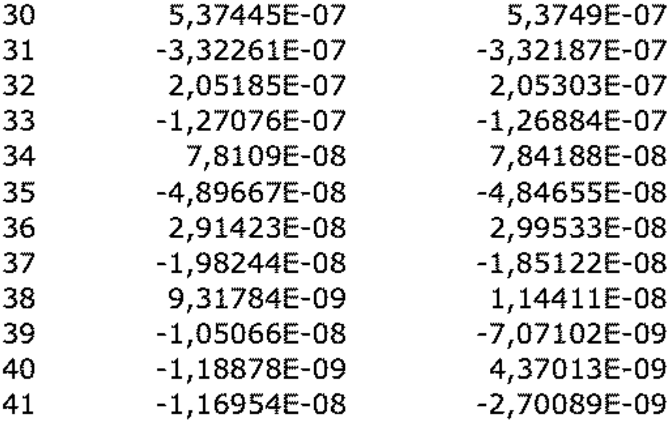
$$\widetilde{\Phi} = -0.618\,033\,988\,749\,895\ldots,$$

$$\log|\widetilde{\Phi}| = -0.481\,211\,825\,059\,603\,4\ldots$$

$$u_{39} = -\widetilde{\Phi}^{39} = -\mathrm{e}^{-18.767\,261\,177\,324{,}453\ldots} = -7.071\,019\ldots10^{-9}.$$

PARI GP : https://pari.math.u-bordeaux.fr/

# Comparing the excel values with the exact values



| | excel value | exact value |
|---|---|---|
| 30 | 5,37445E-07 | 5,3749E-07 |
| 31 | -3,32261E-07 | -3,32187E-07 |
| 32 | 2,05185E-07 | 2,05303E-07 |
| 33 | -1,27076E-07 | -1,26884E-07 |
| 34 | 7,8109E-08 | 7,84188E-08 |
| 35 | -4,89667E-08 | -4,84655E-08 |
| 36 | 2,91423E-08 | 2,99533E-08 |
| 37 | -1,98244E-08 | -1,85122E-08 |
| 38 | 9,31784E-09 | 1,14411E-08 |
| 39 | -1,05066E-08 | -7,07102E-09 |
| 40 | -1,18878E-09 | 4,37013E-09 |
| 41 | -1,16954E-08 | -2,70089E-09 |

# Exact value of $u_{100}$

The answer to initial question is

$$u_{100} = \widetilde{\Phi}^{100}$$

$\widetilde{\Phi} = -0.618\,033\,988\,749\,895\ldots$, $\log|\widetilde{\Phi}| = -0.481\,211\,825\,059\,603\,4\ldots$

$\widetilde{\Phi}^{100} = \mathrm{e}^{-48.121\,182\,505\,960\,34\ldots} = 1.262\,513\,338\,064\ldots 10^{-21}$.

# The linear recurrence sequence $u_n = u_{n-1} + u_{n-2}$

From the two solutions $\Phi^n$ and $\widetilde{\Phi}^n$ one deduces that any solution is of the form $u_n = a\Phi^n + b\widetilde{\Phi}^n$.

Since $|\Phi| > 1$, the term $\Phi^n$ tends to $\infty$.

Since $|\widetilde{\Phi}| < 1$, the term $b\widetilde{\Phi}^n$ tends to $0$.

If $a \neq 0$, then $|u_n|$ tends to infinity like $a\Phi^n$.

If $a = 0$, then $u_n = b\widetilde{\Phi}^n$ tends to $0$.

If two consecutive terms are of the same sign, then all the next ones have the same sign and $|u_n|$ tends to infinity.

# Two computers may give different answers

One of the objectives of the *Aric* project (Arithmetic and Computing)

http://www.ens-lyon.fr/LIP/AriC/

is to build correctly rounded mathematical function programs.

The IEEE 754-2008 standard

https://en.wikipedia.org/wiki/IEEE_754

specifies the behavior of floating-point arithmetic. This standard defines rounding rules : properties to be satisfied when rounding numbers during arithmetic and conversions.

Institute of Electrical and Electronics Engineers (IEEE).

# Decimal expansion of real numbers

A real number has a decimal expansion

$a_k 10^k + a_{k-1} 10^{k-1} + \cdots + a_1 10 + a_0 + b_1 10^{-1} + b_2 10^{-2} + \cdots$

where the digits $a_i$ and $b_j$ belong to $\{0, 1, \ldots 9\}$.
Any sequence of digits defines a real number, but some numbers have two decimal expansions, namely the rational numbers with denominator a power of $10$.
From the relation

$$1 + a + a^2 + a^3 + \cdots + a^m + \cdots = \frac{1}{1-a}$$

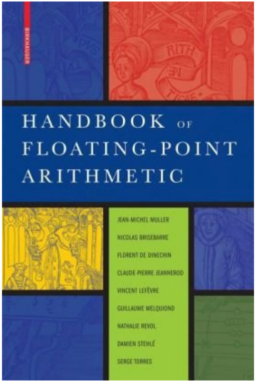
which is valid for $-1 < a < 1$ we deduce

$$1 + \frac{1}{10} + \frac{1}{100} + \frac{1}{1000} + \cdots \frac{1}{10^m} + \cdots = \frac{1}{1 - \dfrac{1}{10}} = \frac{10}{9},$$

hence

$$0.999\,999\,999\,\cdots = 1.$$

# Handbook of floating-point arithmetic



Jean-Michel Muller, Nicolas Brisebarre, Florent de Dinechin, Claude-Pierre Jeannerod, Vincent Lefèvre, Guillaume Melquiond, Nathalie Revol, Damien Stehlé, Serge Torres. *Handbook of floating-point arithmetic.*
Birkhäuser Basel, 2010.

Y. V. Nesterenko and M. Waldschmidt. On the approximation of the values of exponential function and logarithm by algebraic numbers (in Russian). Mat. Zapiski, 2 :23–42, 1996. Available in English at
http://www.math.jussieu.fr/~miw/articles/ps/Nesterenko.ps

# Connection with Diophantine approximation

Many functions considered in the IEEE 754-2008 standard are transcendental, including the exponentials, logarithms, trigonometric functions, and inverse trigonometric functions.

*The Table Maker's Dilemma.*
Accurate rounding of transcendental mathematical functions is difficult because the number of extra digits that need to be calculated to resolve whether to round up or down cannot be known in advance.
https://en.wikipedia.org/wiki/Rounding

# The Table Maker's Dilemma for the exponential function

Let $\alpha$ be a precision-$p$ floating-point number in $[1, 2]$. The exact value $\exp(\alpha)$ belongs to the interval $[\mathrm{e}, \mathrm{e}^2)$. We now use the theorem of Nesterenko and Waldschmidt with $E = \mathrm{e} = 2.7182818\ldots$ and $\theta = \alpha'$, where $\alpha'$ is any precision-$p$ floating-point number in $[1, 6)$. We obtain the following :

$$|\mathrm{e}^{\alpha'} - \alpha| \geq 2^{-688p^2 - 992p\log(p+1) - 67514p - 71824\log(p+1) - 1283614}.$$

Reference : *Handbook of floating-point arithmetic*, § 12.4. Solving the Table Maker's Dilemma for Arbitrary Functions, p. 431.

# $|e^b - a|$ for $a$ and $b$ rational integers

Kurt Mahler
(1903 – 1988)

Maurice Mignotte

Franck Wielonsky

http://www-history.mcs.st-and.ac.uk/Biographies/Mahler.html
https://www.i2m.univ-amu.fr/perso/franck.wielonsky/

# $|\mathrm{e}^b - a|$ for $a$ and $b$ rational integers

K. Mahler noticed that an integer power of $\mathrm{e}$ is never an integer, since $\mathrm{e}$ is transcendental. Hence when $a$ and $b$ are rational integers, we have $\mathrm{e}^b \neq a$.

Mahler obtained a lower bound for $|\mathrm{e}^b - a|$ in 1953 and 1967. His estimates were improved by M. Mignotte (1974), and later by F. Wielonsky (1997). The sharpest known estimate is

$$|\mathrm{e}^b - a| > b^{-20b}.$$

# $|\mathrm{e}^b - a|$ for $a$ and $b$ rational integers

Mahler asked whether there exists an absolute constant $c > 0$ such that, for $a$ and $b$ positive integers,

$$|\mathrm{e}^b - a| > a^{-c}?$$

This is not yet solved. He also noticed that the inequality

$$|b - \log a| < \frac{1}{a}$$

has infinitely many solutions in positive integers $a$ and $b$. Indeed, if $a$ denotes the integral part of $\mathrm{e}^b$, then we have

$$0 < \mathrm{e}^b - a < 1, \qquad 0 < a(b - \log a) < \mathrm{e}^b - a < \mathrm{e}^b(b - \log a),$$

hence

$$0 < b - \log a < \frac{\mathrm{e}^b - a}{a} < \frac{1}{a}\cdot$$

# Mahler's conjecture

Mahler's conjecture arises by considering the numbers $\log a - b_a$ for $a = 1, \ldots, A$, where $b_a$ is the nearest integer to $\log a$, for growing values of $A$, and assuming that these numbers are more or less evenly distributed in the interval $(-1/2, 1/2)$.

Mahler's conjecture is equivalent to the existence of a constant $c > 0$ such that, for $a$ and $b$ positive integers,

$$|\mathrm{e}^b - a| > \mathrm{e}^{-cb}.$$

# Stronger conjecture

I suggest that the numbers $\mathrm{e}^b - a_b$ for $b = 1, \ldots, B$, for growing values of $B$, are evenly distributed in the interval $(-1/2, 1/2)$, where $a_b$ is the nearest integer to $\mathrm{e}^b$. This amounts to suggest the stronger conjecture that there exists a constant $c > 0$ for which

$$|\mathrm{e}^b - a| > b^{-c}.$$

This conjecture is equivalent to the existence of a constant $c > 0$ for which

$$|\mathrm{e}^b - a| > \frac{1}{a(\log a)^c}.$$

# $|\mathrm{e}^b - a|$ for $a$ and $b$ rational numbers

Define $H(p/q) = \max\{|p|, q\}$.

Then for $a$ and $b$ in $\mathbb{Q}$ with $b \neq 0$, the estimate is

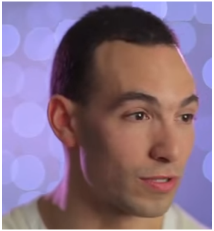
$$|\mathrm{e}^b - a| \geq \exp\{-1,3 \cdot 10^5 (\log A)(\log B)\}$$

where $A = \max\{H(a),\ A_0\}$, $B = \max\{H(b),\ 2\}$.

Yu. V. Nesterenko & M. Waldschmidt – *On the approximation of the values of exponential function and logarithm by algebraic numbers*. (In russian) Mat. Zapiski, **2** *Diophantine approximations, Proceedings of papers dedicated to the memory of Prof. N. I. Feldman*, ed. Yu. V. Nesterenko, Centre for applied research under Mech.-Math. Faculty of MSU, Moscow (1996), 23–42.
http://fr.arXiv.org/abs/math/0002047

# $|\mathrm{e}^b - a|$ for $a$ and $b$ rational numbers

A refinement of our estimate has been obtained in
SAMY KHÉMIRA & PAUL VOUTIER.
*Diophantine approximation and Hermite-Padé approximants of type I of exponential functions.*
Ann. Sci. Math. Québec 35 (2011), no. 1, 85–116.

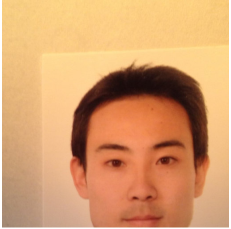


Samy Khemira

Paul Voutier

https://www.youtube.com/watch?v=1WnoyYPu65g
Parlons Passion : Samy donne des cours aux enfants hospitalisés

# $|e^b - a|$ for $a$ and $b$ rational numbers



Makoto Kawashima

Makoto Kawashima,
Linear independence of values of logarithms revisited,
April 3, 2019
https://arxiv.org/abs/1904.01737

New lower bound for linear form in

$$1, \log(1+\alpha), \ldots, \log^{m-1}(1+\alpha)$$

with algebraic integer coefficients in both complex and $p$–adic case. Refinement of the result of Nesterenko-Waldschmidt on the lower bound of linear form in certain values of power of logarithms.

# Further applications of Diophantine Approximation

Hua Loo Keng & Wang Yuan – *Application of number theory to numerical analysis*, Springer Verlag (1981).

Hua Loo Keng

(1910 – 1985)

Wang Yuan

Further applications of Diophantine Approximation include equidistribution modulo 1, discrepancy, numerical integration, interpolation, approximate solutions to integral and differential equations.

http://www-history.mcs.st-and.ac.uk/Biographies/Hua.html

http://www-history.mcs.st-and.ac.uk/PictDisplay/Wang_Yuan.html

Lahore *February 15, 2022*

University of the Punjab, Department of Mathematics
National Mathematical Society of Pakistan
PU-NMS International Schools Series
for Students and Faculty
Mathematics for Computer Science.

# Towards exact rounding of the elementary functions, an application of Diophantine approximation to scientific computing and validated numerics.

*Michel Waldschmidt*

Professeur Émérite, Sorbonne Université,
Institut de Mathématiques de Jussieu, Paris
http://www.imj-prg.fr/~michel.waldschmidt/