Some remarks on correct rounding of functions

Jean-Michel Muller

CNRS - Laboratoire LIP

http://perso.ens-lyon.fr/jean-michel.muller/

Correct rounding of the elementary functions

- base 2, precision p;
- FP number x and integer m (with m somehow larger than p) $\rightarrow$ one can compute an approximation y to $f(x)$ whose error on the significand is $\leq 2^{-m}$.
- can be done with a possible wider format, or using double-word or triple-word arithmetic at critical places, etc.
- is already done in accurate libraries;
- deducing a correct rounding of $f(x)$ from y: may not be possible if $f(x)$ is too close to a breakpoint: a point where the rounding function changes;
- in the following: RN (round-to-nearest, ties to even), but all rounding functions are concerned.

Example with the sine function in binary32 arithmetic

x	$\sin(x)$ (in binary)
$x_1 = 12178540 \times 2^{-23}$	$0.11111110001100001000001010011001000 1 \cdots$
$x_2 = 9898372 \times 2^{-23}$	$0.11101100101100100110110110000000 11111110 \cdots$
$x_3 = 12523099 \times 2^{-23}$	$0.11111110011100100011101011111111111111 0110 \cdots$

- if s is any approximation to $\sin(x_1)$ with error $\leq 2^{-24-4}$, then $\mathrm{RN}(s) = \mathrm{RN}(\sin(x_1))$. An approximation with error 2^{-24-3} may not suffice;
- if s is any approximation to $\sin(x_2)$ with error $\leq 2^{-24-9}$, then $\mathrm{RN}(s) = \mathrm{RN}(\sin(x_1))$. An approximation with error 2^{-24-8} may not suffice;
- to obtain $\mathrm{RN}(\sin(x_3))$ with certainty, we need an approximation with error $< 2^{-24-16}$.

What can be said in general ?

Approximation with error $\leq 2^{-p-k+1}$ on the significand

$\frac{f(x)}{2^{e_{f(x)}}} = \boxed{1.xxxx\cdots xxxxxxx}$

p bits, followed by:

	$k=2$	$k=3$	$k=4$
	00	000	0000
	01	001	0001
	10	010	0010
	11	011	0011
		100	0100
		101	0101
		110	0110
		111	0111
			1000
			1001
			1010
			1011
			1100
			1101
			1110
			1111
Probability of failure:	1/2	1/4	1/8

k bits $\rightarrow$ probability of failure 2^{1-k}

Rule of thumb

If we approximate the significand of $f(x)$ with error $\leq 2^{-p-k+1}$ (roughly speaking, if we compute a $p+k$-bit approximation to $f(x)$), the probability of not being able to deduce $\mathrm{RN}(f(x))$ is around 2^{1-k}.

- exceptions to that rule: if x is tiny, not all bit strings are possible in $\sin(x)$, $\exp(x)$, etc. just after the first p bits. For instance,

$$\sin(1.xxxx\cdots x1 \times 2^{-p}) = 1.xxxx\cdots x01111111111\cdots \times 2^{-p}.$$

- in practice this is not a problem, just choose polynomial approximations where the lowest order term is exactly x for sin or sinh or $\log(1+x)$, 1 for $\exp(x)$, etc. They will automatically deliver correct rounding when x is tiny enough;
- the rule is essential for designing efficient algorithms;

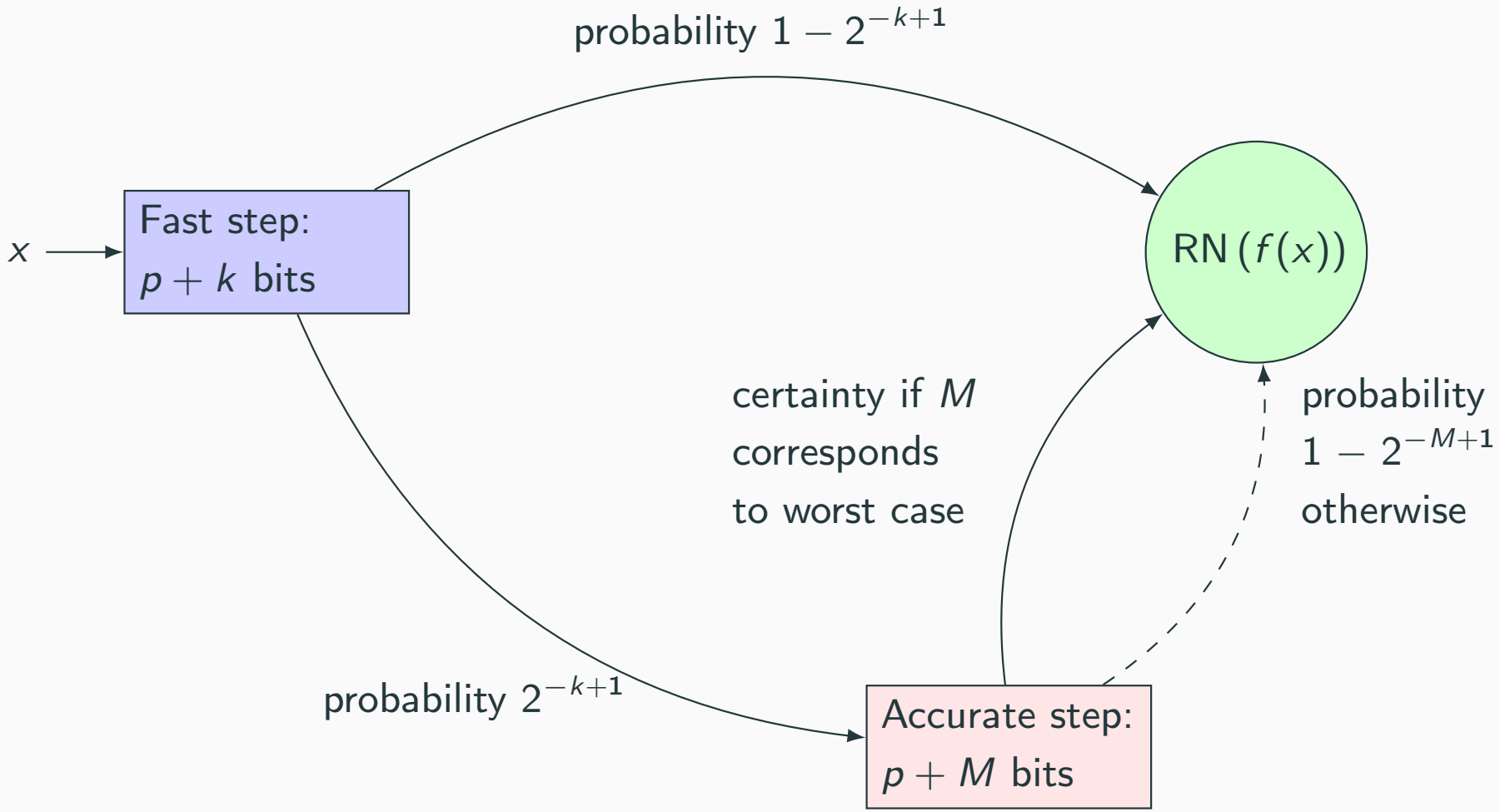
Expected vs actual worst cases

- Rule of thumb $\rightarrow$ if w is the word size, as there are $\approx 2^w$ FP numbers, with a $p+k$-bit approximation there is a total of 2^{w+1-k} failures $\rightarrow$ vanishes as soon as $k \approx w+1$.
- frequently less in practice: correlations (e.g. $\log_2$), function not defined in full range (exp because of overflow, arcsin, etc.);
- actual values (excluding tiny trivial input values):

format	$p+w+1$	exp	$\log_2$	arcsin
binary32	57	52	51	54 ($\lvert x\rvert > 2^{-23}$)
binary64	118	113	108	126 ($\lvert x\rvert > 2^{-25}$)

$\rightarrow$ the rule of thumb is not that bad.

2-step process (Ziv's strategy) – typically, $k \approx 10$



Some references

- R. C. Agarwal et al., New scalar and vector elementary functions for the IBM System/370, in IBM Journal of Res. and Dev., vol. 30, no. 2, 1986,
- L. Fousse, G. Hanrot, V. Lefèvre, P. Pélissier, and P. Zimmermann. 2007. MPFR: A Multiple-Precision Binary Floating- Point Library with Correct Rounding. ACM Trans. Math. Software 33, 2 (2007)
- B. Gladman, V. Innocente, J. Mather, and P. Zimmermann. 2024. Accuracy of Mathematical Functions in Single, Double, Extended Double and Quadruple Precision. (2024), https://hal.inria.fr/hal-03141101.
- T. Hubrecht, C.-P. Jeannerod, and P. Zimmermann. 2023. Towards a correctly-rounded and fast power function in binary64 arithmetic. In 30th IEEE Symposium on Computer Arithmetic, 2023
- J. P. Lim and S. Nagarakatte. 2021. High performance correctly rounded math libraries for 32-bit floating point representations. In PLDI'21: 42nd ACM SIGPLAN International Conference on Programming Language Design and Implementation, Virtual Event
- A. Sibidanov, P. Zimmermann, and S. Glondu. The CORE-MATH Project. In 29th IEEE Symposium on Computer Arithmetic, 2022
- C. Lauter. Arrondi Correct de Fonctions Mathématiques. Ph.D. Dissertation. École Normale Supérieure de Lyon, 2008. https://www.christoph-lauter.org/these.pdf
- N. Brisebarre, G. Hanrot, and O. Robert, Exponential sums and correctly-rounded functions, IEEE Transactions on Computers, Volume 66, No 12, 2017
- A. Ziv. Fast evaluation of elementary mathematical functions with correctly rounded last bit. ACM Trans. Math. Software 17, 3, 1991
- N. Brisebarre, G. Hanrot, J.-M. Muller, and P. Zimmermann, Correctly-rounded evaluation of a function: why, how, and at what cost? https://hal.science/hal-04474530.