## Nothing's perfect

Assume IEEE-754 single-precision arithmetic. The machine number which is closest to $\arctan(2^{30})$ is

$$\frac{13176795}{8388608} = 1.57079637050628662109375 > \frac{\pi}{2}.$$

If the arctangent function is implemented in round-to-nearest mode, we get an arctangent larger than $\pi/2$. A consequence of this is that in our system,

$$\tan\left(\arctan\left(2^{30}\right)\right) = -2.2877\cdots \times 10^{7}.$$

## Monotonicity

Assume that the function $f$ to be computed is increasing in the considered domain, and that its approximation $F$ is such that an *absolute error* bounded by $\epsilon$ is made. If

$$\epsilon \leq \frac{1}{2}\mathsf{ulp}(x) \times \min_{t \in [x, x+\mathsf{ulp}(x)]} |f'(t)|, \tag{1}$$

then the obtained implementation of $f$ will be monotonic.

**Theorem 1 (Ferguson and Brightman)** *Let $f(x)$ be a monotonic function defined on the interval $[a, b]$. Let $F(x)$ be an approximation of $f(x)$ whose associated relative error is less than or equal to $\epsilon$, $\epsilon < 1$. If for every pair $m < m^+$ of consecutive machine numbers in $[a, b]$*

$$\epsilon < \frac{|f(m^+) - f(m)|}{|f(m^+)| + |f(m)|},$$

*then $F(x)$ exhibits on the set of machine numbers in $[a, b]$ the same monotonic behavior exhibited by $f(x)$ on $[a, b]$.*

## Exact Rounding

Assume that from any real number $x$ and any integer $m$ (with $m > n$), we are able to compute an approximation of $f(x)$ with an error on its mantissa $y$ smaller than or equal to $2^{-m}$.

The problem is to get an $n$-bit floating-point exactly rounded result from the mantissa $y$ of an approximation of $f(x)$, with error $\pm 2^{-m}$. This is not possible if $y$ has the form:

– in rounding to the nearest mode,

$$\overbrace{\underbrace{1.xxxxx\cdots xxx}_{n \text{ bits}}1000000\cdots000000}^{m \text{ bits}}xxx\cdots$$

or

$$\overbrace{\underbrace{1.xxxxx\cdots xxx}_{n \text{ bits}}0111111\cdots111111}^{m \text{ bits}}xxx\cdots;$$

– in rounding towards $+\infty$, or towards $-\infty$ modes,

$$\overbrace{\underbrace{1.xxxxx\cdots xxx}_{n \text{ bits}}0000000\cdots000000}^{m \text{ bits}}xxx\cdots$$

or

$$\overbrace{\underbrace{1.xxxxx\cdots xxx}_{n \text{ bits}}1111111\cdots111111}^{m \text{ bits}}xxx\cdots.$$

## Examples

– assume a radix-2 floating-point arithmetic with 6-bit mantissa,

$$\sin(11.1010) = 0.0\boxed{111011}01111110\cdots,$$

– In IEEE-754 double-precision floating-point arithmetic, the exponential of the "machine number"

$$x = \boxed{0.1101011001100111101111100100011010110100111011110000}$$

is

$$\boxed{10.010011111000010111001001011110000011110111001110000}01111$$
$$1111111111111111111111111111111111111111111111111101\cdots.$$

This example: worst case for exponentials of double precision numbers between $1/2$ and $1$.

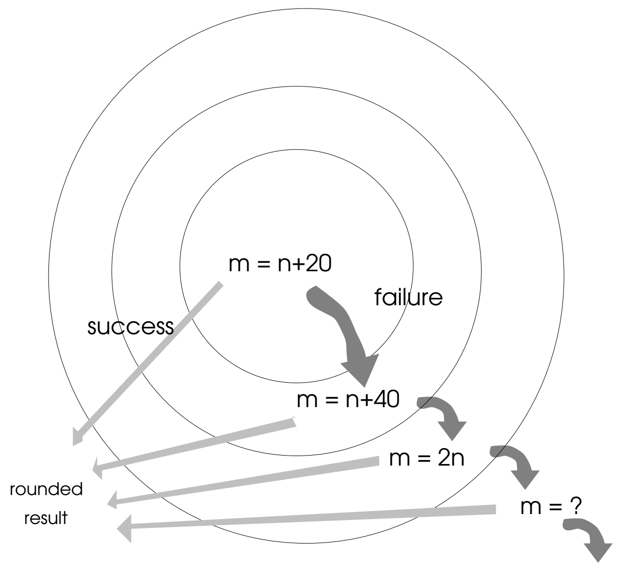


FIG. 1 - Ziv's multilevel strategy.

Let $p/q$ be a rational number, with $q > 0$ and $\gcd(p, q) = 1$. Define $H(p/q) = \max\{|p|, q\}$.

**Theorem 2 (Y. Nesterenko and M. Waldschmidt (1995))** *Let $\alpha, \beta$ be rational numbers. Let $A$, $B$, and $E$ be positive real numbers with $E \geq e$ satisfying*

$$A \geq \max\left(H(\alpha), e\right), \quad B \geq H(\beta).$$

*Then*

$$\begin{aligned} &|e^\beta - \alpha| \geq \\ &\exp\Big(-211 \times (\ln B + \ln\ln A + 2\ln(E|\beta|_+) + 10) \\ &\times (\ln A + 2E|\beta| + 6\ln E) \times (3.3\ln(2) + \ln E) \times (\ln E)^{-2}\Big), \end{aligned}$$

*where $|\beta|_+ = \max(1, |\beta|)$.*

This theorem with $\alpha$ and $\beta$ being machine numbers: lower bound on the distance between a machine number and the exponential of a machine number.

**Theorem 3 (Brent, 1976)** *If* $\pi$ *and* $\ln 2$ *are precomputed, the elementary function* $f(x)$ *can be evaluated to precision* $n$ *in* $(K + o(1))\, M(n) \log_2 n$ *operations, where*

$$K = \begin{cases} 13 & \textit{if} \quad f(x) = \ln(x) \textit{ or } \exp(x) \\ 34 & \textit{if} \quad f(x) = \arctan(x) \textit{ or } \sin(x) \textit{ or } \cos x \end{cases}$$