On the maximum relative error when computing x^n in floating-point arithmetic

Jean-Michel Muller

Joint work with S. Graillat and V. Lefèvre
INVA 2014

Thank you!

ありがとうございます

Floating-Point numbers, roundings

Precision-p binary FP number (set $\mathbb{F}_p$): either 0 or

$$x = X \cdot 2^{e_x - p + 1},$$

where X and $e_x \in \mathbb{Z}$, with $2^{p-1} \leq |X| \leq 2^p - 1$.

- unlimited exponent range $\longrightarrow$ results valid unless underflow/overflow occurs;
- X: integral significand of x;
- $2^{1-p} \cdot X$: significand of x;
- e_x: exponent of x.

Floating-Point numbers, roundings

- In general, the sum, product, quotient, etc., of two FP numbers is not an FP number: it must be rounded;
- correct rounding: *Rounding function* $\circ$, and when $(a \top b)$ is performed, the returned value is $\circ(a \top b)$;
- default rounding function RN:
  - (i) for all FP numbers y, $|\mathrm{RN}(t) - t| \leq |y - t|$
  - (ii) if there are two FP numbers that satisfy (i), $\mathrm{RN}(t)$ is the one whose integral significand is even.

Relative error due to roundings, u and γ notations

Let $t \in \mathbb{R}, 2^e \leq t < 2^{e+1}$.

- we have $2^e \leq \mathrm{RN}(t) \leq 2^{e+1}$, and

$$|\mathrm{RN}(t) - t| \leq 2^{e-p}. \tag{1}$$

$\rightarrow$ upper bound on the relative error due to rounding t:

$$\left|\frac{\mathrm{RN}(t) - t}{t}\right| \leq 2^{-p}. \tag{2}$$

- $u = 2^{-p}$: rounding unit.

Relative error due to roundings, u and γ notations

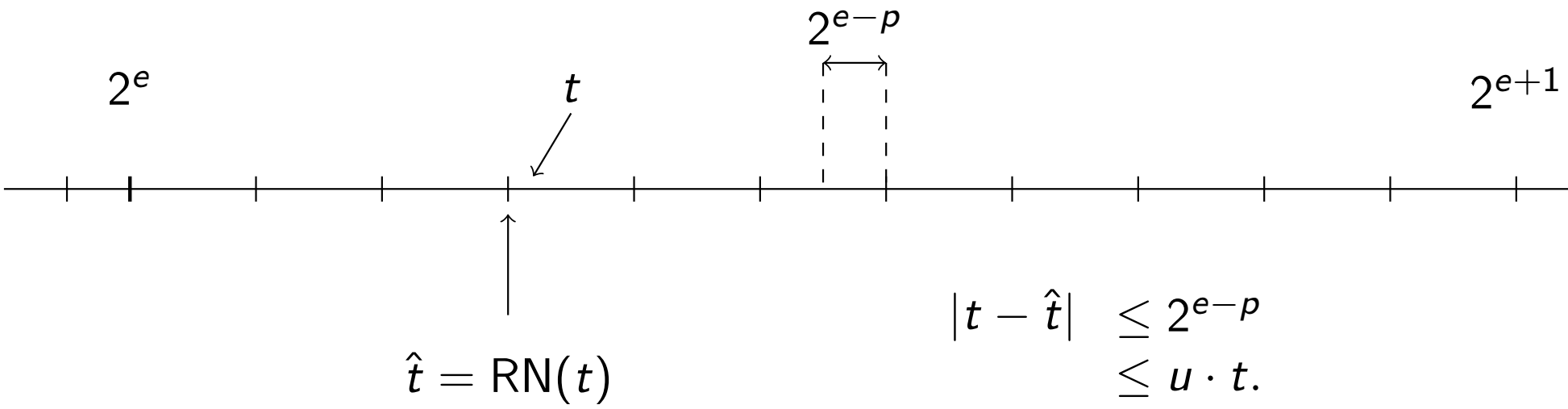

Figure 1: *In precision-p binary FP arithmetic, in the normal range, the relative error due to rounding to nearest is bounded by* $u = 2^{-p}$.

Relative error due to roundings, u and γ notations

Floating-point multiplication $a \times b$:

- exact result $z = ab$;
- computed result $\hat{z} = \mathrm{RN}(z)$;

$$(1-u) \cdot z \leq \hat{z} \leq (1+u) \cdot z; \tag{3}$$

$\rightarrow$ when we approximate $\pi_n = a_1 \cdot a_2 \cdot \cdots \cdot \cdots \cdot a_n$ by

$$\hat{\pi}_n = \mathrm{RN}(\cdots \mathrm{RN}(\mathrm{RN}(a_1 \cdot a_2) \cdot a_3) \cdot \cdots) \cdot a_n),$$

we have

Theorem 1

$$(1-u)^{n-1}\pi_n \leq \hat{\pi}_n \leq (1+u)^{n-1}\pi_n. \tag{4}$$

Relative error due to roundings, u and γ notations

$\longrightarrow$ relative error on the product $a_1 \cdot a_2 \cdots \cdots \cdot a_n$ bounded by

$$\psi_{n-1} = (1+u)^{n-1} - 1.$$

- if we define (Higham)

$$\gamma_k = \frac{ku}{1-ku},$$

then, as long as $ku < 1$ (always holds in practical cases),

$$k \cdot u \leq \psi_k \leq \gamma_k.$$

$\longrightarrow$ classical bound: γ_{n-1}.

- For "reasonable" n, ψ_{n-1} is very slightly better than γ_{n-1}, yet γ_{n-1} is easier to manipulate;
- in single and double precision we never observed a relative error $\geq (n-1) \cdot u$.

Special case: $n \leq 4$

The bound on the relative error due to rounding can be slightly improved (using a remark by Jeannerod and Rump):

if $2^e \leq t < 2^{e+1}$, then $|t - \mathrm{RN}(t)| \leq 2^{e-p} = u \cdot 2^e$, and

- if $t \geq 2^e \cdot (1+u)$, then $|t - \mathrm{RN}(t)|/t \leq u/(1+u)$;
- if $t = 2^e \cdot (1 + \tau \cdot u)$ with $\tau \in [0, 1)$, then $|t - \mathrm{RN}(t)|/t = \tau \cdot u/(1 + \tau \cdot u) < u/(1+u)$,

$\longrightarrow$ the maximum relative error due to rounding is bounded by $u/(1+u)$ (attained $\longrightarrow$ no further improvement);

$\longrightarrow$ we can replace (4) by

$$\left(1 - \frac{u}{1+u}\right)^{n-1} \pi_n \leq \hat{\pi}_n \leq \left(1 + \frac{u}{1+u}\right)^{n-1} \pi_n. \qquad (5)$$

Special case: $n \leq 4$

Property 1

If $1 \leq k \leq 3$ *then*

$$\left(1 + \frac{u}{1+u}\right)^k < 1 + k \cdot u.$$

- $k = 2$:

$$\left(1 + \frac{u}{1+u}\right)^2 - (1 + 2u) = -\frac{u^2 \cdot (1 + 2u)}{(1+u)^2} < 0;$$

- $k = 3$:

$$\left(1 + \frac{u}{1+u}\right)^3 - (1 + 3u) = -\frac{u^3 \cdot (2 + 3u)}{(1+u)^3} < 0.$$

$k = n - 1 \rightarrow$ for $n \leq 4$, the relative error of the iterative product of n FP numbers is bounded by $(n-1) \cdot u$.

The particular case of computing powers

- "General" case of an iterated product: no proof for $n \geq 5$ that $(n-1) \cdot u$ is a valid bound (when starting the study we conjectured this is the case);

$\rightarrow$ focus on x^n, where $x \in \mathbb{F}_p$ and $n \in \mathbb{N}$;

- we assume the "naive" algorithm is used:

$y \leftarrow x$
for $k = 2$ to n **do**
  $y \leftarrow \mathrm{RN}(x \cdot y)$
end for
return y

- notation: $\hat{x}_j$ = value of y after the iteration corresponding to $k = j$ in the **for** loop.

Main result

We wish to prove

Theorem 2

Assume $p \geq 5$ (holds in all practical cases). If

$$n \leq \sqrt{2^{1/3} - 1} \cdot 2^{p/2},$$

then

$$|\hat{x}_n - x^n| \leq (n-1) \cdot u \cdot x^n.$$

- we can assume $1 \leq x < 2$;
- two cases: x close to 1, and x far from 1.

Preliminary results

First,

$$(1-u)^{n-1} \geq 1-(n-1)\cdot u$$

for all $n \geq 2$ and $u \in [0,1]$.

$\rightarrow$ the left-hand bound of

$$(1-u)^{n-1}\pi_n \leq \hat{\pi}_n \leq (1+u)^{n-1}\pi_n.$$

suffices to show that

$$1-(n-1)\cdot u \cdot x_n \leq \hat{x}_n$$

$\rightarrow$ to establish the Theorem, we only need to focus on the right-hand bound.

Preliminary results

For $t \neq 0$, define

$$\overline{t} = \frac{t}{2^{\lfloor \log_2 |t| \rfloor}}.$$

We have,

Lemma 3

Let t be a real number. If

$$2^e \leq w \cdot 2^e \leq |t| < 2^{e+1}, e \in \mathbb{Z} \tag{6}$$

(in other words, if $w \leq |\overline{t}|$) then

$$\left| \frac{\mathrm{RN}(t) - t}{t} \right| \leq \frac{u}{w}.$$

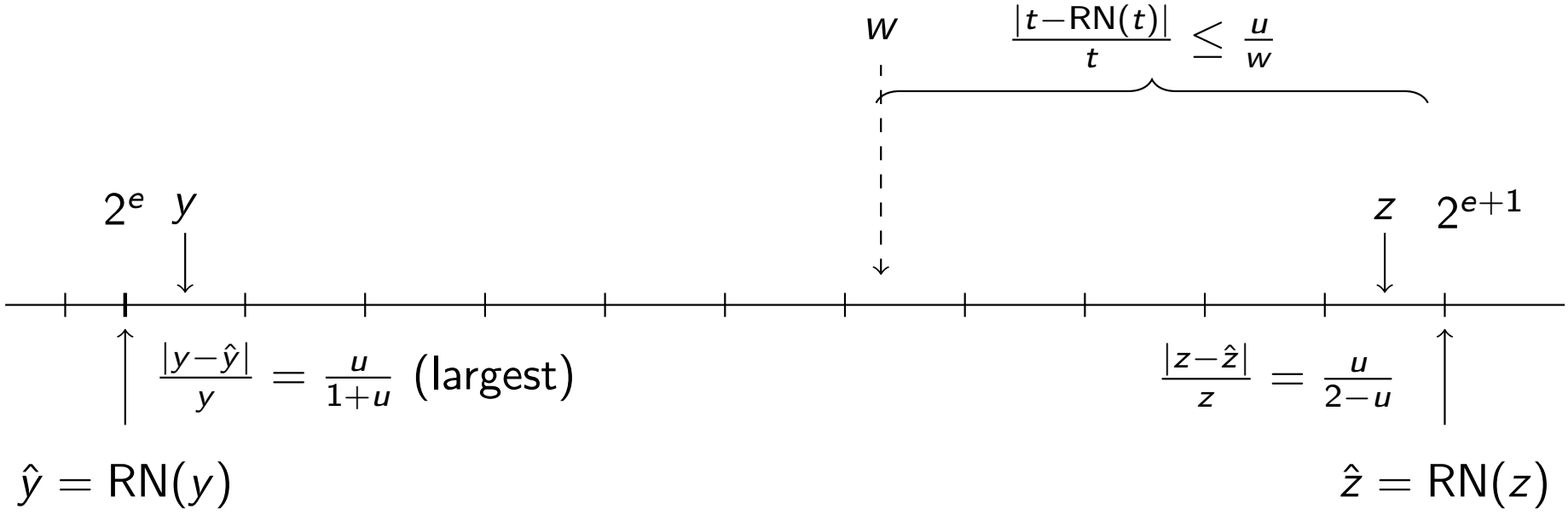

Figure 2: *The bound on the relative error due to rounding to nearest can be reduced to* $u/(1+u)$. *Furthermore, if we know that* $w \leq \overline{t} = t/2^e$, *then* $|\mathrm{RN}(t) - t|/t \leq u/w$.

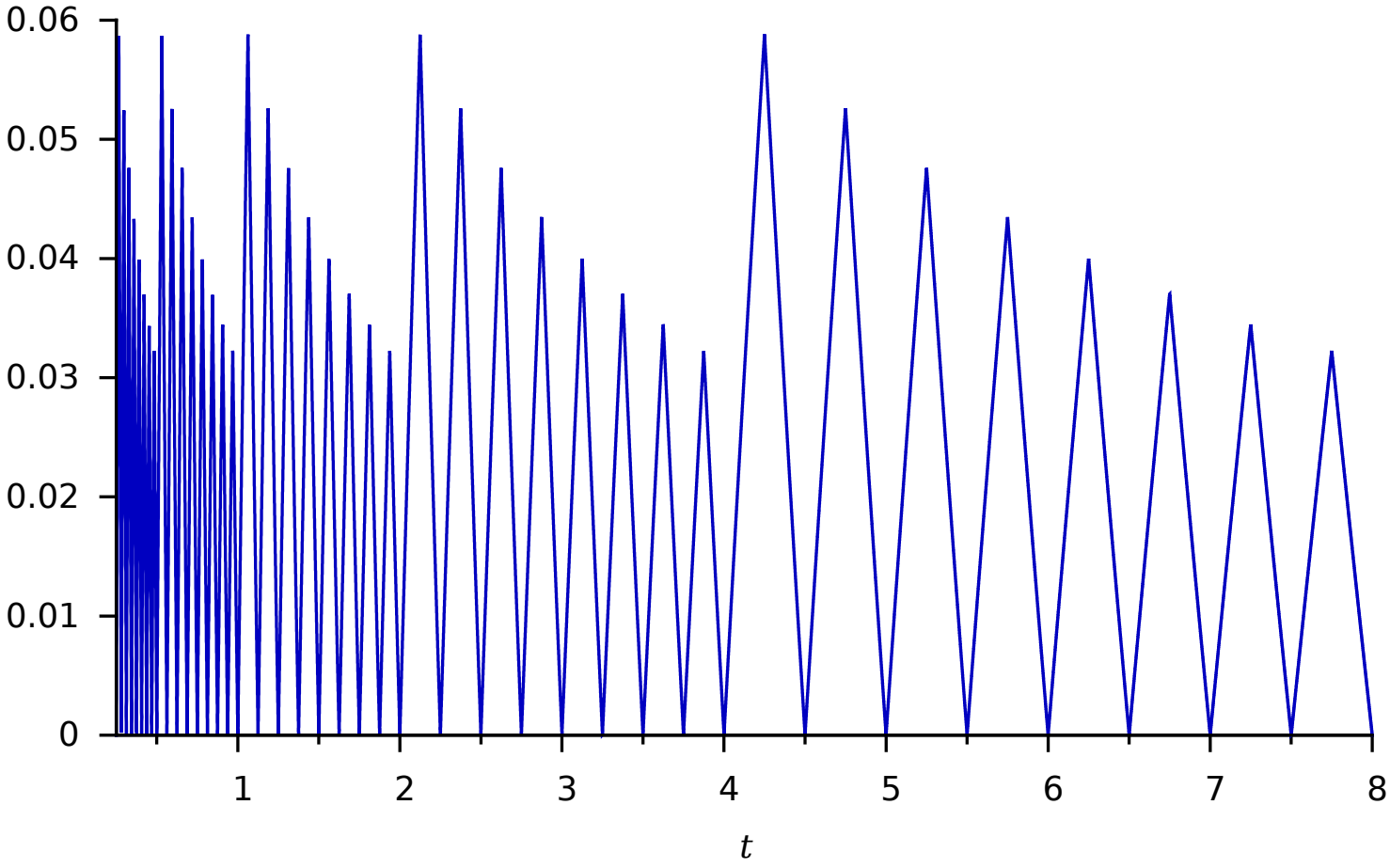

Figure 3: Relative error due to rounding, namely $|\mathrm{RN}(t) - t|/t$, for $\frac{1}{5} \leq t \leq 8$, and $p = 4$.

Local maximum error for x^6 as a function of x ($p = 53$)

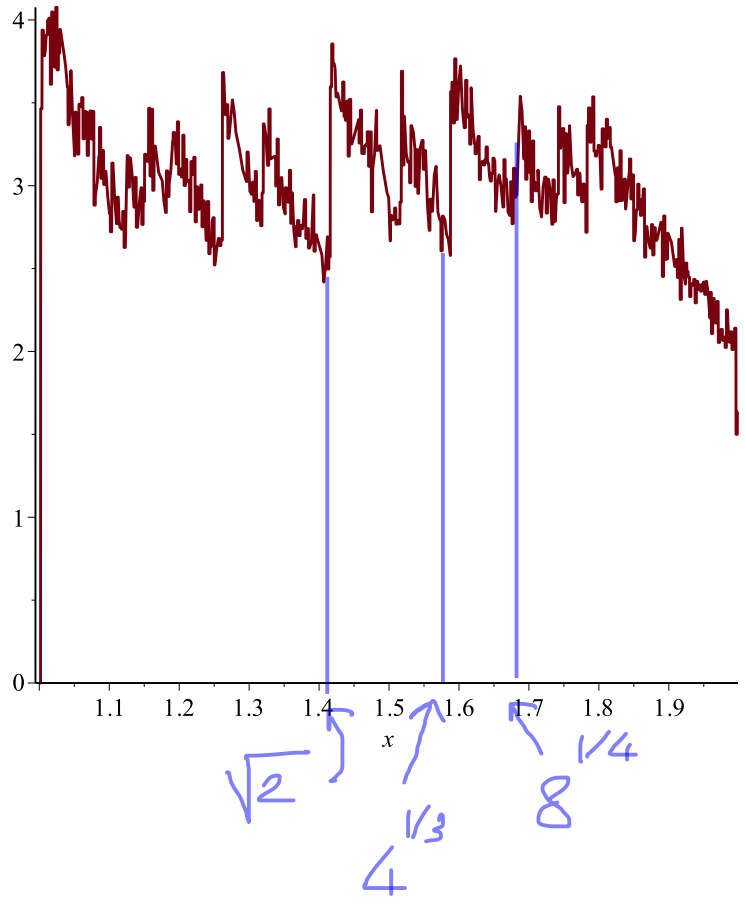

Figure 4: The input interval $[1, 2)$ is divided into 512 equal-sized subintervals. In each subinterval, we calculate x^6 for 5000 consecutive FP numbers x, compute the relative error, and plot the largest attained error.

Main idea behind the proof

At least once in the execution of the algorithm, $\overline{x \cdot y}$ is far enough from 1 to sufficiently reduce the error bound on the multiplication $y \leftarrow \mathrm{RN}(x \cdot y)$, so that the overall error bound becomes $\leq (n-1) \cdot u$.

$y \leftarrow x$
for $k = 2$ to n **do**
  $y \leftarrow \mathrm{RN}(x \cdot y)$
end for
return y

$$\psi_{n-1} = (1+u)^{n-1} - 1 = (n-1)\,u + \left(1/2\,n^2 - 3/2\,n + 1\right) u^2 + \cdots$$

$\rightarrow$ we have to save $\approx \frac{n^2}{2} u^2$, which requires one of the values $\overline{x \cdot y}$ to be larger than $\approx 1 + \frac{n^2}{2} u$.

What we are going to show

Unless x is very near 1, at least once $\overline{x \cdot y} \geq 1 + n^2 u$, so that in (4) the term $(1+u)^{n-1}$ can be replaced by

$$(1+u)^{n-2} \cdot \left(1 + \frac{u}{1+n^2 u}\right).$$

$\longrightarrow$ we need to bound this last quantity. We have,

Lemma 4

If $0 \leq u \leq 2/(3n^2)$ *and* $n \geq 3$ *then*

$$(1+u)^{n-2} \cdot \left(1 + \frac{u}{1+n^2 u}\right) \leq 1 + (n-1) \cdot u. \tag{7}$$

Proof of Lemma 4 (with the help of Bruno Salvy)

Proving the Lemma reduces to proving that

$$P(u) = (1 + (n-1)u)(1 + n^2 u) - (1+u)^{n-2}(1 + n^2 u + u) \geq 0$$

for $0 \leq u \leq 2/(3n^2)$. We have

$$\ln(1+u) \leq u - \frac{u^2}{2} + \frac{u^3}{3}.$$

- $\ln(1+u) \leq u \Rightarrow (n-2)\ln(1+u) < 1/(2n) \leq 1/6$;
- For $0 \leq t \leq 1/6$, $e^t \leq 1 + t + \frac{3}{5}t^2$;

$\rightarrow$ for $0 \leq u \leq 2/(3n^2)$, to prove that $P(u) \geq 0$ it suffices to prove that

$$\begin{aligned} Q(n,u) &= (1 + (n-1)\,u)\left(n^2 u + 1\right) \\ &- \left(1 + (n-2)\left(u - \tfrac{1}{2}\,u^2 + \tfrac{1}{3}\,u^3\right) + \tfrac{3}{5}\,(n-2)^2\left(u - \tfrac{1}{2}\,u^2 + \tfrac{1}{3}\,u^3\right)^2\right) \\ &\times \left(n^2 u + u + 1\right) \geq 0. \end{aligned} \tag{8}$$

Proof of Lemma 4 (with the help of Bruno Salvy)

By defining $a = n^2 u$, $(5n^2/a^2) \cdot Q(n,u)$ is equal to

$$
\begin{aligned}
S(n,a) = -3\,a + 2 + \frac{\frac{29}{2}\,a + \frac{19}{2}}{n} + \frac{3\,a^2 - 17\,a - 7}{n^2} - \frac{1}{6}\,\frac{a(82\,a-5)}{n^3} \\
- \frac{1}{12}\,\frac{a\left(33\,a^2 - 187\,a + 20\right)}{n^4} + \frac{1}{3}\,\frac{a^2(33\,a-8)}{n^5} + \frac{1}{12}\,\frac{a^2\left(12\,a^2 - 153\,a + 52\right)}{n^6} \\
- \frac{a^3(4\,a-7)}{n^7} - \frac{1}{3}\,\frac{a^3\left(a^2 - 14\,a + 21\right)}{n^8} + \frac{4}{3}\,\frac{a^4(a-2)}{n^9} - \frac{1}{3}\,\frac{a^4(5\,a-8)}{n^{10}} \\
+ \frac{4}{3}\,\frac{a^5}{n^{11}} - \frac{4}{3}\,\frac{a^5}{n^{12}}
\end{aligned}
\tag{9}
$$

We wish to show that $S(n,a) \geq 0$ for $0 \leq a \leq 2/3$.

We examine the terms of $S(n,a)$ separately. For $a \in [0, 2/3]$ and $n \geq 3$:

- $-3\,a+2$ is always larger than 0;
- $\left(\frac{29}{2}\,a+\frac{19}{2}\right) n^{-1}$ is always larger than $19/(2n)$;
- $\frac{3\,a^2-17\,a-7}{n^2}$ is always larger than $-6/n$;
- $-\frac{1}{6}\,\frac{a(82\,a-5)}{n^3}$ is always larger than $-7/(10n)$;
- $-\frac{1}{12}\,\frac{a\left(33\,a^2-187\,a+20\right)}{n^4}$ is always larger than $-17/(10000n)$;
- $\frac{1}{3}\,\frac{a^2(33\,a-8)}{n^5}$ is always larger than $-3/(10000n)$;
- $\frac{1}{12}\,\frac{a^2\left(12\,a^2-153\,a+52\right)}{n^6}$ is always larger than $-69/(10000n)$;
- $-\frac{a^3(4\,a-7)}{n^7}$ is always larger than 0;
- $-\frac{1}{3}\,\frac{a^3\left(a^2-14\,a+21\right)}{n^8}$ is always larger than $-6/(10000n)$;
- $\frac{4}{3}\,\frac{a^4(a-2)}{n^9}$ is always larger than $-6/(100000n)$;
- $-\frac{1}{3}\,\frac{a^4(5\,a-8)}{n^{10}}$ and $\frac{4}{3}\,\frac{a^5}{n^{11}}$ are always larger than 0;
- $-\frac{4}{3}\,\frac{a^5}{n^{12}}$ is always larger than $-1/(1000000n)$.

$\rightarrow$ for $0 \leq a \leq 2/3$ and $n \geq 3$, $S(n,a) \geq 2790439/(1000000n)$.

Two remarks

Remark 1

Assume $n \leq \sqrt{2/3} \cdot 2^{p/2}$. *If* $\exists k \leq n$ *s.t.* $\mathrm{RN}(x \cdot \hat{x}_{k-1}) \leq x \cdot \hat{x}_{k-1}$ *(i.e., if in the algorithm at least one rounding is done downwards), then*

$$\hat{x}_n \leq (1 + (n-1) \cdot u) x^n.$$

Proof.

We have

$$\hat{x}_n \leq (1+u)^{n-2} x^n.$$

Lemma 4 implies $(1+u)^{n-2} < 1 + (n-1) \cdot u$. Therefore,

$$\hat{x}_n \leq (1 + (n-1) \cdot u) x^n.$$

Two remarks

Remark 2

Assume $n \leq \sqrt{2/3} \cdot 2^{p/2}$. *If* $\exists k \leq n-1$, *s.t.* $\overline{x \cdot \hat{x}_k} \geq 1 + n^2 \cdot u$, *then*

$$\hat{x}_n \leq (1 + (n-1) \cdot u) x^n.$$

Proof.

By combining Lemmas 3 and 4, if there exists k, $1 \leq k \leq n-1$, such that

$$\overline{x \cdot \hat{x}_k} \geq 1 + n^2 \cdot u,$$

then

$$\hat{x}_n \leq (1+u)^{n-2} \cdot \left(1 + \frac{u}{1+n^2 u}\right) \cdot x^n \leq (1 + (n-1) \cdot u) \cdot x^n.$$

Proof of Theorem 2

We assume $n \geq 5$. Proof articulated as follows

- if x is close enough to 1, then when computing $\mathrm{RN}(x^2)$, the rounding is done downwards;
- in the other cases, $\exists k \leq n-1$ such that $\overline{x \cdot \hat{x}_k} \geq 1 + n^2 \cdot u$.

Lemma 5

If $1 < x < 1 + 2^{p/2} \cdot u$, *then* $\hat{x}_2 = \mathrm{RN}(x^2) < x^2$.

Proof.

$x < 1 + 2^{p/2} \cdot u \Rightarrow x = 1 + k \cdot 2^{-p+1} = 1 + 2ku$, with $k < 2^{p/2-1}$. We have $x^2 = 1 + 2k \cdot 2^{-p+1} + k^2 \cdot 2^{-2p+2}$, which gives $\mathrm{RN}(x^2) = 1 + 2k \cdot 2^{-p+1} < x^2$. □

In the following, we assume that no rounding is done downwards, which implies $x \geq 1 + 2^{p/2} \cdot u$.

Proof of Theorem 2: case $x^2 \leq 1 + n^2u$

- $x \geq 1 + 2^{p/2}u > 1 + nu \Rightarrow x^n > (1 + nu)^n > 1 + n^2u$;
- no downward rounding $\Rightarrow \hat{x}_{n-1} \cdot x > (1 + n^2u)$.

Therefore

- if $\hat{x}_{n-1}x < 2$, then $\overline{\hat{x}_{n-1}x} \geq (1 + n^2u)$, so that, from Remark 2, $x^n \leq (1 + (n-1) \cdot u) \cdot x^n$;
- if $\hat{x}_{n-1}x \geq 2$, let k be the smallest integer such that $\hat{x}_{k-1}x \geq 2$. $x^2 \leq 1 + n^2u \Rightarrow k \geq 3$. We have

$$\hat{x}_{k-1} \geq \frac{2}{x} \geq \frac{2}{\sqrt{1 + n^2u}},$$

hence

$$\hat{x}_{k-2} \cdot x \geq \frac{2}{\sqrt{1 + n^2u} \cdot (1 + u)}. \tag{10}$$

$$\hat{x}_{k-2} \cdot x \geq \frac{2}{\sqrt{1+n^2u} \cdot (1+u)}.$$

Define

$$\alpha_p = \sqrt{\left(\frac{2^{p+1}}{2^p+1}\right)^{2/3} - 1}.$$

For all $p \geq 5$, $\alpha_p \geq \alpha_5 = 0.745\cdots$, and $\alpha_p \leq \sqrt{2^{2/3}-1} = 0.766\cdots$. If

$$n \leq \alpha_p \cdot 2^{p/2}, \tag{11}$$

then

$$\frac{2}{\sqrt{1+n^2u} \cdot (1+u)} \geq 1 + n^2u.$$

$\rightarrow \hat{x}_{k-2} \cdot x \geq 1 + n^2u$. Also, $\hat{x}_{k-2} \cdot x < 2$ since k is the smallest integer such that $\hat{x}_{k-1}x \geq 2$. Therefore

$$\overline{\hat{x}_{k-2} \cdot x} \geq 1 + n^2u.$$

Which implies $x^n \leq (1 + (n-1) \cdot u) \cdot x^n$.

Proof of Theorem 2: case $x^2 > 1 + n^2u$

- if $x^2 < 2$ then $\overline{x^2} > 1 + n^2u \Rightarrow x^n \leq (1 + (n-1) \cdot u)$;
- $x^2 = 2$ impossible (x is rational);

$\longrightarrow$ we assume $x^2 > 2$ we also assume $x^2 < 2 + 2n^2u$ (otherwise, $\overline{x^2} \geq 1 + n^2u$). This gives

$$x^{n-1} < (2 + 2n^2u)^{\frac{n-1}{2}},$$

therefore, using the classical bound (Theorem 1),

$$\hat{x}_{n-1} < (2 + 2n^2u)^{\frac{n-1}{2}} \cdot (1 + u)^{n-2},$$

which implies

$$x \cdot \hat{x}_{n-1} < (2 + 2n^2u)^{\frac{n}{2}} \cdot (1 + u)^{n-2}. \tag{12}$$

Reminder:

$$x \cdot \hat{x}_{n-1} < (2 + 2n^2u)^{n/2} \cdot (1 + u)^{n-2} \text{ and } n \geq 5$$

Define

$$\beta = \sqrt{2^{1/3} - 1}.$$

If $n \leq \beta \cdot 2^{p/2}$ then $2 + 2n^2u \leq 2^{4/3}$, so that

$$(2 + 2n^2u)^{n/2} \cdot (1 + u)^{n-2} \leq 2^{2n/3} \cdot (1 + u)^{n-2}. \tag{13}$$

The function

$$g(t) = 2^{t-1} - 2^{2t/3} \left(1 + \frac{1}{2^p}\right)^{t-2} = 2^{2t/3} \left[2^{t/3-1} - \left(1 + \frac{1}{2^p}\right)^{t-2}\right].$$

is continuous, goes to $+\infty$ as $t \to +\infty$, has one root only:

$$\frac{\log(2) + 2\log\left(1 + \frac{1}{2^p}\right)}{\frac{1}{3}\log(2) - \log\left(1 + \frac{1}{2^p}\right)},$$

which is < 4 as soon as $p \geq 5 \Rightarrow$ if $p \geq 5$ then $x \cdot \hat{x}_{n-1} < 2^{n-1}$.

Reminder: if $p \geq 5$ then $x \cdot \hat{x}_{n-1} < 2^{n-1}$.

- define k as the smallest integer for which $x \cdot \hat{x}_{k-1} < 2^{k-1}$,
- $3 \leq k \leq n$ (we have assumed $x^2 > 2$),
- $x \cdot \hat{x}_{k-2} \geq 2^{k-2} \Rightarrow \hat{x}_{k-1} = \mathrm{RN}(x \cdot \hat{x}_{k-2}) \geq 2^{k-2}$.

Therefore, $\hat{x}_{k-1}$ and $x \cdot \hat{x}_{k-1}$ belong to the same binade, therefore,

$$\overline{x \cdot \hat{x}_{k-1}} \geq x > \sqrt{2}. \tag{14}$$

The constraint $n \leq \beta \cdot 2^{p/2}$ implies

$$1 + n^2 u \leq 1 + \beta^2 = 2^{1/3} < \sqrt{2}. \tag{15}$$

By combining (14) and (15) we obtain

$$\overline{x \cdot \hat{x}_{k-1}} \geq 1 + n^2 u.$$

Therefore, using Remark 2, we deduce that $\hat{x}_n \leq (1 + (n-1) \cdot u) \cdot x^n$.

Final steps

$\forall p \geq 5$, $\alpha_p \geq \beta \rightarrow$ combining the conditions found in the cases $x^2 \leq 1 + n^2 u$ and $x^2 > 1 + n^2 u$, we deduce

If $p \geq 5$ *and* $n \leq \beta \cdot 2^{p/2}$, *then for all* x,

$$(1 - (n-1) \cdot u) \cdot x^n \leq \hat{x}_n \leq (1 + (n-1) \cdot u) \cdot x^n.$$

where $\beta = \sqrt{2^{1/3} - 1} = 0.5098245285339 \cdots$

Q.E.D.

Questions:

- is the restriction $n \leq \beta \cdot 2^{p/2}$ problematic?
- is the bound sharp?
- any hope of generalizing to iterated products?

On the restriction $n \leq \beta \cdot 2^{p/2}$

format	p	$n_{\max}$
binary32/single	24	2088
binary64/double	53	48385542
binary128/quad	113	5195358025846195 9

With the first n larger than the bound, x^n under- or overflows, unless

- in single precision, $0.95905406 \leq x \leq 1.0433863$,
- in double precision, $0.999985359 \leq x \leq 1.000014669422$,

and nobody will use the "naive" algorithm for a huge n.

On the restriction $n \leq \beta \cdot 2^{p/2}$

Furthermore, that restriction is not just a "proof artefact". For very big n, the bound does not hold:

If $p = 10$ and $x = 891$, when computing x^{2474}, relative error $2473.299u$.

Notice that:

- for $p = 10$, $n_{\max} = \beta \cdot 2^{p/2} = 16.31$;
- 2474 is the smallest exponent for which the bound does not hold when $p = 10$.

The case of huge values of n

- $\hat{x}_n$ computed approximation to x^n;
- $\overline{\hat{x}_n} = \hat{x}_n / 2^{\lfloor \log_2 \hat{x}_n \rfloor}$;
- one can build examples for which $\exists m$ s.t. $\overline{\hat{x}_m} = 1$ (and $\overline{x^m} \neq 1$);

$\rightarrow$ for all i, $\hat{x}_{m+i} = \hat{x}_i$;

- let α be the relative error on x_m:

$$\hat{x}_m = (1 + \alpha) \cdot x^m,$$

- relative error on x^{mk} ?

$$\hat{x}_{mk} = (1 + \alpha)^k \cdot x^{mk},$$

$\rightarrow$ the relative error grows exponentially with k

$\rightarrow$ ultimately it will be larger that $(mk - 1) \cdot u$.

Tightness of the bound $(n-1)\cdot u$

Small p and not-too-large n: an exhaustive test is possible.

Table 1: Actual maximum relative error assuming $p = 8$, compared with γ_{n-1} and our bound $(n-1)u$.

n	actual maximum	γ_{n-1}	our bound
4	$1.73903u$	$3.0355u$	$3u$
5	$2.21152u$	$4.06349u$	$4u$
6	$2.53023u$	$5.099601u$	$5u$
7	$2.69634u$	$6.1440u$	$6u$
$8 = n_{max}$	$3.42929u$	$7.1967u$	$7u$

$\longrightarrow$ our bound seems to be quite poor... however...

Tightness of the bound $(n-1)\cdot u$

For larger values of p:

- single precision ($p = 24$), exhaustive search still possible, largest error $4.328005619u$ for $n = 6$, and $7.059603149u$ for $n = 10$;
- double precision ($p = 53$), we have a case with error $4.7805779u$ for $n = 6$ and $7.8618\cdots u$ for $n = 10$;
- quad precision ($p = 113$), case with error $4.8827888185u$ for $n = 6$;

$\rightarrow$ we seem to get close to $(n-1)\cdot u$ for large p.

Rough explanation

- n is not too large
- the $\overline{x \cdot \hat{x}_k}$ are close to 1;
- we assume that each elementary relative rounding error ϵ_i is uniformly distributed in $[-u, +u]$.

$$\hat{x}_n = x^n \cdot (1+\epsilon_1)(1+\epsilon_2)\cdots(1+\epsilon_{n-1}) \approx x^n \cdot (1+\epsilon_1+\epsilon_2+\cdots\epsilon_{n-1}).$$

Define $\alpha_i = (\epsilon_i + u)/(2u)$. The α_i are uniform in $[0,1] \rightarrow$ cumulative distribution function of $\alpha_1 + \alpha_2 + \cdots + \alpha_{n-1}$:

$$F(x) = \frac{1}{(n-1)!} \sum_{k=0}^{\lfloor x \rfloor} (-1)^k \binom{n-1}{k} (x-k)^{n-1}.$$

For a given x, probability that $|\hat{x}_{10} - x^{10}|/x^{10} \geq 8.9$: 5.38×10^{-18}.
$\rightarrow$ *There are just not enough possible single precision significands for that to happen!*

Repartition of relative error

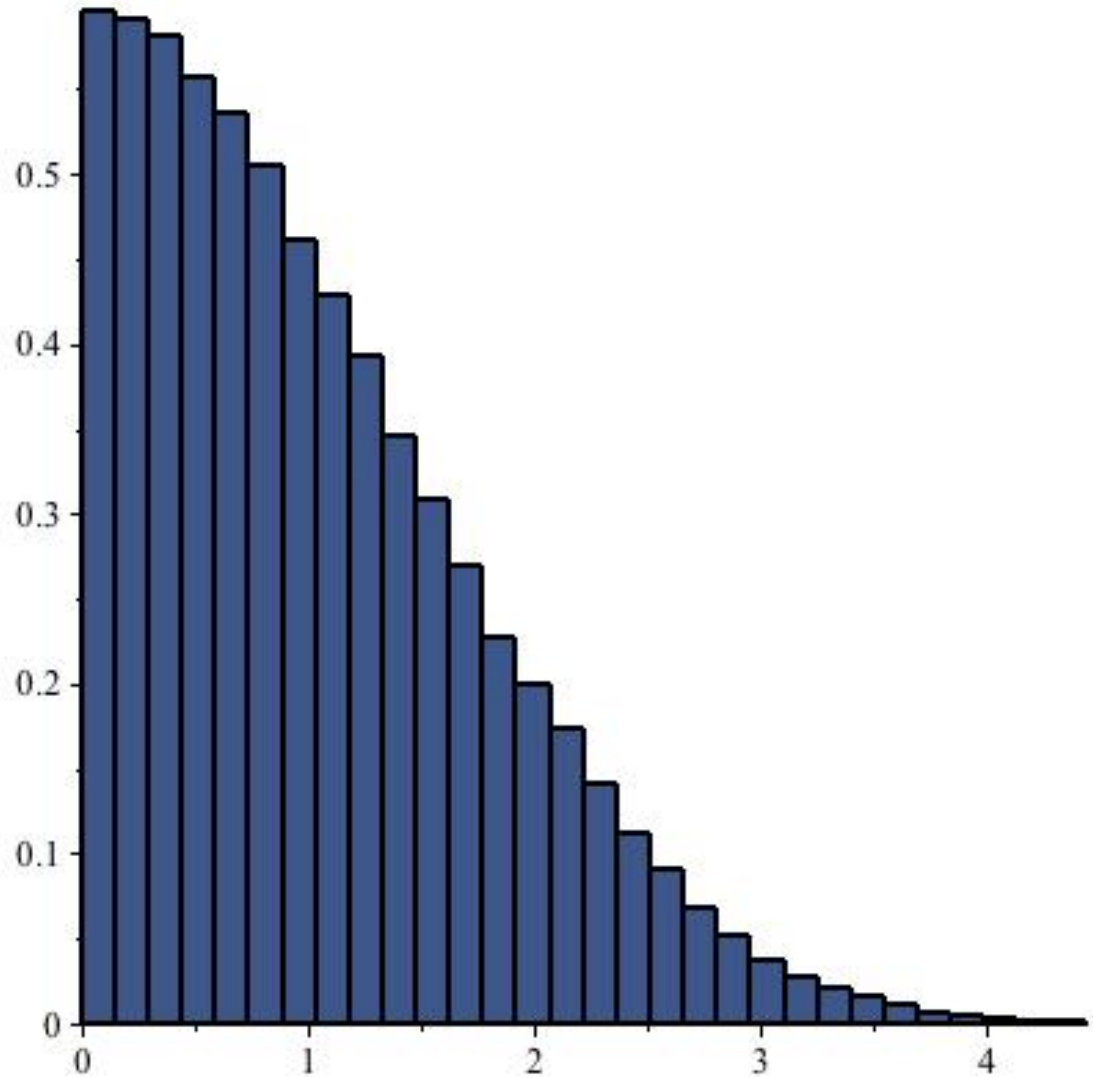

Figure 5: Repartition of the relative error (divided by u), for $p = 53$ and $n = 6$, for a sample of 100000 random values of x uniformly chosen between 1 and 2.

Building "bad cases" for the iterated product

Still in precision-p binary FP arithmetic, we approximate

$$a_1 \cdot a_2 \cdots \cdots \cdot a_n,$$

by

$$\mathrm{RN}(\cdots \mathrm{RN}(\mathrm{RN}(a_1 \cdot a_2) \cdot a_3) \cdot \cdots) \cdot a_n)$$

- $\pi_k = a_1 \cdots a_k$,
- $\hat{\pi}_k =$ computed value,
- relative error $|\pi_n - \hat{\pi}_n|/\pi_n$ upper-bounded by γ_{n-1},
- conjecture: if n is "not too large" it is bounded by $(n-1)u$.

Let us now show how to build a_1, a_2, ..., a_n so that the relative error becomes extremely close to $(n-1) \cdot u$.

Building "bad cases" for the iterated product

- define $a_1 = 1 + k_1 \cdot 2^{-p+1}$, and $a_2 = 1 + k_2 \cdot 2^{-p+1}$. We have

$$\pi_2 = a_1 a_2 = 1 + (k_1 + k_2) \cdot 2^{-p+1} + k_1 k_2 \cdot 2^{-2p+2}.$$

  If k_1 and k_2 are not too large, $1 + (k_1 + k_2) \cdot 2^{-p+1}$ is a FP number $\longrightarrow$ we wish $k_1 + k_2$ to be as small as possible, while $k_1 k_2 \cdot 2^{-2p+2}$ is as close as possible (yet ess than) to 2^{-p}. Hence a natural choice is

$$k_1 = k_2 = \left\lfloor 2^{\frac{p}{2}-1} \right\rfloor,$$

  which gives $\hat{\pi}_2 < \pi_2$.
- Now, if at step $i-1$ we have

$$\hat{\pi}_i = 1 + g_i \cdot 2^{-p+1}, \text{ with } \hat{\pi}_i < \pi_i,$$

  we choose a_{i+1} of the form $1 + k_{i+1} 2^{-p+1}$, with
  - $k_{i+1} = \left\lceil \frac{2^{p-2}}{g_i} - 1 \right\rceil$ if $g_i \leq 2^{\frac{p}{2}-1}$;
  - $k_{i+1} = -\left\lfloor \frac{2^{p-2}}{g_i} + 1 \right\rfloor$ otherwise.

Building "bad cases" for the iterated product

Table 2: Relative errors achieved with the values a_i generated by our method.

p	n	relative error
24	10	$8.99336984\cdots u$
24	100	$98.9371972591\cdots u$
53	10	$8.99999972447\cdots u$
53	100	$98.9999970091\cdots u$
113	10	$8.999999999999973119\cdots u$
113	100	$98.999999999999701662\cdots u$

Conclusion

- error bound $(n-1)\cdot u$ for computation of x^n by the naive algorithm;
- valid for $n \leq \sqrt{2^{1/3}-1}\cdot 2^{p/2} \longrightarrow$ all practical cases;
- small improvement: the main interest lies in the simplicity of the bound;
- seems to be "asymptotically sharp" (as $p \longrightarrow \infty$) but not sure;
- unsolved issue: iterated products and n "not too large";
- if this is the case, it is very sharp.

Thank you for your attention.