Probabilistic rounding error analysis for large scale, low precision scientific computing

Theo Mary, CNRS
lecture based on joint work with Nick Higham (Univ. Manchester)

AFAE course, Sorbonne Université (2019 version)

An example: recursive summation

Let $s = \sum_{i=1}^{n} x_i$ be computed by recursive summation:

$s_1 = x_1$ for $i = 2$ to n $\quad s_i = s_{i-1} + x_i$ end $(s = s_n)$	$s_1 = x_1$ for $i = 2$ to n $\quad \widehat{s}_i = (\widehat{s}_{i-1} + x_i)(1 + \delta_i), \quad \lvert\delta_i\rvert \le u$ end $(\widehat{s} = \widehat{s}_n)$

We have the bound

$$\frac{|\widehat{s} - s|}{|s|} \le (n-1)\kappa u + O(u^2), \qquad \kappa = \frac{\sum_{i=1}^{n} |x_i|}{\left|\sum_{i=1}^{n} x_i\right|}$$

- Many efforts focus on getting small error even for large κ
- What about n?

Historical perspective

- Backward error analysis was developed by James Wilkison in the 1960s
- At that time, $n = 100$ was huge! Solving linear systems of $n = O(10)$ equations would take days

$\Rightarrow$ n was considered a "constant"

Hence traditional error analysis has paid little attention to n

> *The **constant** terms in an error bound are the least important parts of error analysis. It is not worth spending much effort to minimize constants because the achievable improvements are usually insignificant.*
>
> *Nick Higham, ASNA 2ed (2002)*

Today: large scale problems

- Since the 1990s, the TOP500 list ranks the world's most powerful supercomputers based on their ability to solve linear systems of equations $Ax = b$ as fast as possible

The Summit supercomputer

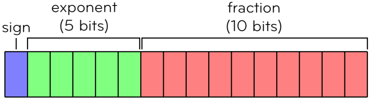
- In the latest ranking (Nov. 2019), the #1 computer solved a linear system of 16 million equations in 5.5 hours

Today: low precision arithmetics

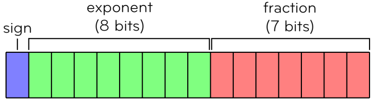
Type		Bits	Range	$u = 2^{-t}$
fp64	double	64	$10^{\pm 308}$	$2^{-53} \approx 1 \times 10^{-16}$
fp32	single	32	$10^{\pm 38}$	$2^{-24} \approx 6 \times 10^{-8}$
fp16	half	16	$10^{\pm 5}$	$2^{-11} \approx 5 \times 10^{-4}$
bfloat16	half	16	$10^{\pm 38}$	$2^{-8} \approx 4 \times 10^{-3}$

Half precision increasingly **supported by hardware**:

- Present: **NVIDIA** Pascal & Volta GPUs, **AMD** Radeon Instinct MI25 GPU, **Google** TPU, **ARM** NEON
- Near future: Fujitsu A64FX ARM, **IBM** AI chips, **Intel** Xeon Cooper Lake and Intel Nervana Neural Network

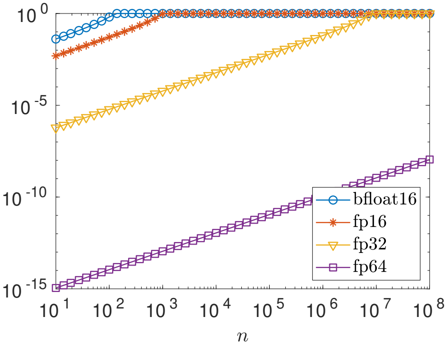


fp16



bfloat16

Backward stability guarantees are lost



In half precision, backward error of order 1 for $n > 1024$ (fp16) or $n > 128$ (bfloat16)

Not a single correct digit guaranteed when n and/or u are large, even if $\kappa = 1$!

For large scale and/or low precision computations, **bounds of the form *nu* are no longer acceptable**

A first summation error analysis

Recursive summation computes

$$\widehat{s}_i = (\widehat{s}_{i-1} + x_i)(1+\delta_i), \quad i = 2\colon n \qquad \text{with } \widehat{s}_1 = s_1 = x_1$$

$$\begin{aligned}
\widehat{s} - s = \widehat{s}_n - s_n &= \widehat{s}_{n-1} - s_{n-1} + (\widehat{s}_{n-1} + x_n)\delta_n \\
&= \sum_{i=2}^{n} (\widehat{s}_{i-1} + x_i)\delta_i = \sum_{i=2}^{n} \widehat{s}_i \delta_i/(1+\delta_i) = \sum_{i=2}^{n} s_i\delta_i + O(u^2)
\end{aligned}$$

First-order worst-case bound:

$$|\widehat{s} - s| \le \sum_{i=2}^{n} |s_i||\delta_i| \le \sum_{i=2}^{n}\sum_{j=1}^{i} |x_j|\, u \le (n-1)\sum_{j=1}^{n} |x_j|\, u$$

This bound is however attained only when

$$\forall i \ \delta_i = +u \text{ or } \forall i \ \delta_i = -u$$

$\Rightarrow$ intuitively seems very unlikely!

Modelling rounding errors as random variables

Since the 1960s, researchers have tried modelling the δ_i as independent random variables to translate the intuition that they are probably not all equal to $+u$ or $-u$

It is important to realize that rounding errors are **not** random:

- For a given input, order of computation, and deterministic rouding mode, rounding errors are entirely predetermined
- Successive δ_i are dependent:

$$\begin{aligned} \mathrm{fl}\big((a+b)+c\big) &= \mathrm{fl}\big(\mathrm{fl}(a+b)+c\big) = \mathrm{fl}\big((a+b)(1+\delta_1)+c\big) \\ &= \big((a+b)(1+\delta_1)+c\big)(1+\delta_2) \end{aligned}$$

> *There is no claim that ordinary rounding and chopping are random processes, or that successive errors are independent.* ***The question to be decided is whether or not these particular probabilistic models of the processes will adequately describe what actually happens.***
>
> *— Hull and Swenson, 1966*

A first basic probabilistic analysis

- Assume δ_i are independent random variables uniformly distributed on $[-u, +u]$
- Law of large numbers: $\widehat{s} - s = \sum_{i=2}^{n} s_i \delta_i$ converges towards $\mathbb{E}(\widehat{s} - s) = \sum_{i=2}^{n} s_i \, \mathbb{E}(\delta_i) = 0$ when $n \to \infty$

Central limit theorem (classical variant)

Let X_1, ..., X_n i.i.d. random variables of mean μ and variance σ^2, and let $S_n = \frac{1}{n} \sum_{i=1}^{n} X_i$. For large enough n, $\sqrt{n}(S_n - \mu)$ follows the normal distribution $\mathcal{N}(0, \sigma^2)$.

- For $X_i \leftarrow \delta_i$, $\mu = 0$ and $\sigma^2 = u^2/3 \Rightarrow \frac{\sum_{i=1}^{n} \delta_i}{\sqrt{n-1}} \sim \mathcal{N}(0, u^2/3)$
- Problem: $X_i \leftarrow s_i \delta_i$ are **not** identically distributed

Central limit theorem (Lyapunov's variant)

Let X_1, ..., X_n independent random variables of mean μ_i and variance σ_i^2. Let $\mathfrak{S}_n^2 = \sum_{i=1}^n \sigma_i^2$ and let $S_n = \sum_{i=1}^n X_i - \mu_i$. For large enough n, and under Lyapunov's condition, $S_n/\mathfrak{S}_n$ follows the normal distribution $\mathcal{N}(0,1)$.

Lyapunov's condition: for some $\epsilon > 0$,

$$\lim_{n\to\infty} \frac{1}{\mathfrak{S}_n^{2+\epsilon}} \sum_{i=1}^{n} \mathbb{E}\left(|X_i - \mu_i|^{2+\epsilon}\right) = 0$$

For $X_i \leftarrow s_i\delta_i$, $\mathfrak{S}_{n-1}^2 = \sum_{i=2}^n s_i^2\, u^2/3$ and so

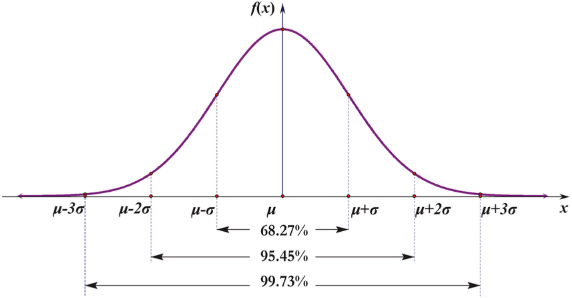
$$\mathfrak{S}_{n-1} \leq \sqrt{(n-1)/3} \sum_{j=1}^{n} |x_j|\, u$$

The 3σ rule

For a random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, the probability that $X \in [\mu - \lambda\sigma, \mu + \lambda\sigma]$ is given by $\Phi(\lambda) - \Phi(-\lambda)$ where

$$\Phi(\lambda) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\lambda} e^{-t^2/2} dt$$

In particular, $\Phi(3) - \Phi(-3) \approx 99.7\%$.



- $(\widehat{s} - s)/\mathfrak{S}_{n-1} \sim \mathcal{N}(0, 1)$ and so we have with 99.7% probability

$$|\widehat{s} - s| \leq 3\mathfrak{S}_{n-1} \leq \sqrt{3(n-1)} \sum_{j=1}^{n} |x_j|\, u + O(u^2)$$

Wilkinson's rule of thumb

In general, the statistical distribution of the rounding errors will reduce considerably the function of n occurring in the relative errors. We might expect in each case that this function should be replaced by something which is no bigger than its square root.

— James Wilkinson, 1961

Why is it a **"rule of thumb"**?

- Assumptions on δ_i: which are really needed? Which are true in practice?
- Use of central limit theorem requires $n \to \infty$
- Since $X_i = s_i\delta_i$ are not identically distributed Lyapunov's condition must also hold
- Only a first order bound
- Only for summation? Does it generalize to other algorithms?

A more general probabilistic model

We will show that the assumptions made by the following more general model are sufficient to obtain a $O(\sqrt{n}u)$ bound

Probabilistic model of rounding errors

In the computation of interest, the quantities δ in the model

$$\mathrm{fl}(a \text{ op } b) = (a \text{ op } b)(1+\delta), \quad |\delta| \le u, \quad \text{op} \in \{+,-,\times,\div\}$$

associated with every pair of operands are independent random variables of mean zero.

- No specific distribution (e.g., uniform) assumed
- Not necessarily identically distributed

Questions left for later:

- Are these assumptions as general as possible? (i.e., are they necessary? spoiler: no they're not)
- Are these assumptions satisfied in practice? (i.e., are they reasonable? spoiler: not always, but...)

Concentration inequalities

Central limit theorem is only valid asymptotically (for $n \to \infty$).
To obtain a rigorous result for a given n, we use the following concentration inequality

Hoeffding's inequality

Let X_1, ..., X_n be random independent variables satisfying $|X_i| \leq c$.
Then the sum $S = \sum_{i=1}^{n} X_i$ satisfies

$$\Pr(|S - \mathbb{E}(S)| \geq \lambda\sqrt{n}c) \leq 2\exp(-\lambda^2/2)$$

Two key assumptions:

- Independence $\to$ OK with our probabilistic model
- Boundedness $\to$ OK with standard FP model $|\delta_i| \leq u$

Resulting bound for summation

Recall that $\widehat{s} - s = \sum_{i=2}^{n} s_i \delta_i$
Exercise: apply Hoeffding's inequality to $X_i \leftarrow s_i \delta_i$

Resulting bound for summation

Recall that $\widehat{s} - s = \sum_{i=2}^{n} s_i \delta_i$

Exercise: apply Hoeffding's inequality to $X_i \leftarrow s_i \delta_i$

- Independence $\rightarrow$ OK because independence of δ_i
- Boundedness $\rightarrow$ OK $|X_i| \leq c_i = \sum_{j=1}^{i} |x_j|\, u$

Apply Hoeffding's inequality to $S = \sum_{i=2}^{n} X_i$ with
$c = \max c_i = \sum_{j=1}^{n} |x_j|\, u$ to obtain with probability $\exp(-\lambda^2/2)$

$$|S - \mathbb{E}(S)| = |S| = |\widehat{s} - s| \leq \lambda \sqrt{n-1} \sum_{j=1}^{n} |x_j|\, u$$

This bound is for summation only... what about inner products? matrix–matrix multiplication? solution to linear systems? eigenvalue decompositions? polynomial computations? etc. etc.

Systematizing the analysis: backward error analysis

$$\widehat{s}_2 = (x_1 + x_2)(1 + \delta_2)$$
$$\widehat{s}_k = (\widehat{s}_{k-1} + x_k)(1 + \delta_k) = x_1 \prod_{j=2}^{k} (1 + \delta_j) + \ldots + x_k(1 + \delta_k)$$
$$\widehat{s}_n = \widehat{s} = \sum_{i=1}^{n} x_i \prod_{j=\max(i,2)}^{n} (1 + \delta_j), \qquad |\delta_j| \le u$$

Fundamental lemma in backward error analysis

If $|\delta_i| \le u$ for $i = 1 : n$ and $nu < 1$, then
$$\prod_{i=1}^{n} (1 + \delta_i) = 1 + \theta_n, \quad |\theta_n| \le \gamma_n := \frac{nu}{1 - nu} = nu + O(u^2)$$

This fundamental lemma can be applied to essentially any numerical algorithm $\Rightarrow$ can we obtain an analogous probabilistic lemma?

Probabilistic backward error analysis

- Main difficulty: $\prod_{i=1}^{n}(1+\delta_i)$ is a product, but probabilistic tools (CLT, concentration ineq) apply to sums

- Transform the product in a sum by taking the logarithm

$$S = \log \prod_{i=1}^{n}(1+\delta_i) = \sum_{i=1}^{n} \log(1+\delta_i)$$

- Taylor: $\log(1+\delta_i) = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}\delta_i^k}{k}$
  - $\log(1+\delta_i) \le \delta_i + \sum_{k=2}^{\infty} |\delta_i|^k = \delta_i + \frac{\delta_i^2}{1-|\delta_i|}$
  - $\log(1+\delta_i) \ge \delta_i - \sum_{k=2}^{\infty} |\delta_i|^k = \delta_i - \frac{\delta_i^2}{1-|\delta_i|}$

$\Rightarrow$ $|\log(1+\delta_i)| \le u + \frac{u^2}{1-u} = \frac{u}{1-u} := c$

Probabilistic backward error analysis (cont'd)

- **Exercise:** use Hoeffding's ineq to $X_i \leftarrow \log(1 + \delta_i)$ to bound $|S|$

Probabilistic backward error analysis (cont'd)

- **Exercise:** use Hoeffding's ineq to $X_i \leftarrow \log(1+\delta_i)$ to bound $|S|$

$$|S - \mathbb{E}(S)| \le \lambda\sqrt{n}c = \frac{\lambda\sqrt{n}u}{1-u}$$

with probability $\exp(-\lambda^2/2)$

  - $\mathbb{E}(S) = \sum_{i=1}^{n} \mathbb{E}(\log(1+\delta_i))$
  - $|\mathbb{E}(\log(1+\delta_i))| \le \frac{u^2}{1-u}$

$\Rightarrow$ $|\mathbb{E}(S)| \le \frac{nu^2}{1-u}$

$$|S| \le \frac{\lambda\sqrt{n}u + nu^2}{1-u}$$

- Retrieve the result by taking the exponential of S:

$$\prod_{i=1}^{n}(1+\delta_i) = 1+\theta_n, \quad |\theta_n| \le \exp\left(\frac{\lambda\sqrt{n}u + nu^2}{1-u}\right) - 1 := \widetilde{\gamma}_n(\lambda)$$

For $0 < t < 1$, $e^t \le 1 + t/(1-t)$ gives

$$\widetilde{\gamma}_n(\lambda) \le \lambda\sqrt{n}u + O(u^2)$$

Probabilistic backward error analysis: main result

Main result

Let δ_i, $i = 1 : n$, be independent random variables of mean zero such that $|\delta_i| \le u$. Then, for any constant $\lambda > 0$, the relation

$$\prod_{i=1}^{n}(1+\delta_i) = 1 + \theta_n, \quad |\theta_n| \le \widetilde{\gamma}_n(\lambda) := \exp\left(\frac{\lambda\sqrt{n}u + nu^2}{1-u}\right) - 1$$

$$\le \lambda\sqrt{n}u + O(u^2)$$

holds with probability $P(\lambda) = 1 - 2\exp(-\lambda^2/2)$

Key features:

- Exact bound, not first order (and $nu < 1$ not required)
- No "$n \to \infty$" assumption (CLT $\to$ Hoeffding's inequality)
- Small values of λ suffice: $P(1) \approx 0.73$, $P(5) \ge 1 - 10^{-5}$
- Can be applied **in a nearly systematic way:** $\gamma_n \to \widetilde{\gamma}_n(\lambda)$

Application to summation, inner products

$$\widehat{s} = \sum_{i=1}^{n} x_i \prod_{j=\max(i,2)}^{n} (1+\delta_j) = \sum_{i=1}^{n} x_i (1+\theta_i)$$

$$\forall i \le n,\ \Pr\big(|\theta_i| \le \widetilde{\gamma}_{n-\max(i,2)+1}(\lambda)\big) \ge 1 - 2\exp(-\lambda^2/2)$$

$$\Rightarrow \Pr\big(\forall i \le n,\ |\theta_i| \le \widetilde{\gamma}_{n-\max(i,2)+1}(\lambda)\big) \ge 1 - 2n\exp(-\lambda^2/2)$$

$$\Rightarrow \Pr\big(\forall i \le n,\ |\theta_i| \le \widetilde{\gamma}_{n-1}(\lambda)\big) \ge 1 - 2n\exp(-\lambda^2/2)$$

Similarly, if $s = x^T y = \sum_{i=1}^{n} x_i y_i$, then

$$\widehat{s} = \sum_{i=1}^{n} x_i y_i (1+\epsilon_i) \prod_{j=\max(i,2)}^{n} (1+\delta_j) = \sum_{i=1}^{n} x_i y_i (1+\theta_i)$$

where

$$\Pr\big(\forall i \le n,\ |\theta_i| \le \widetilde{\gamma}_{n}(\lambda)\big) \ge 1 - 2n\exp(-\lambda^2/2)$$

Componentwise expression:

$$\widehat{s} = (x+\Delta x)^T y = x^T (y + \Delta y), \qquad |\Delta x| \le \widetilde{\gamma}_n(\lambda), \quad |\Delta y| \le \widetilde{\gamma}_n(\lambda)$$

Application to numerical linear algebra

Similarly to summation, repeated application of our main result yields the following bounds with probability $1 - 2F\exp(-\lambda^2/2)$

Algorithm	Bound	F
Matrix–vector product $y = Av$, $A \in \mathbb{R}^{m\times n}$, $v \in \mathbb{R}^n$	$\widetilde{\gamma}_n(\lambda)$	mn
Matrix–matrix product $C = AB$, $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{n\times p}$	$\widetilde{\gamma}_n(\lambda)$	mnp
LU factorization $A = LU$, $A \in \mathbb{R}^{n\times n}$	$\widetilde{\gamma}_n(\lambda)$	$n^3/3 + n^2/2 + n/6$
Triangular system $Tx = b$, $T \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^n$	$\widetilde{\gamma}_n(\lambda)$	$n^2/2 + n/2$
Linear system $Ax = b$, $A \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^n$	$3\widetilde{\gamma}_n(\lambda) + 2\widetilde{\gamma}_n(\lambda)^2$	$n^3/3 + 3n^2/2 + 7n/6$
Cholesky factorization $A = LL^T$, $A \in \mathbb{R}^{n\times n}$	$\widetilde{\gamma}_{n+1}(\lambda)$	$n^3/6 + n^2/2 + n/3$

$\Rightarrow$ n (in $\widetilde{\gamma}_n(\lambda)$) is the maximal number of products of terms $(1+\delta_i)$

$\Rightarrow$ $2F$ (in the prob.) $\approx$ equals the flops required by the algorithm

Keeping the probabilities independent of n

- Let the probability of failure $P_F(\lambda, n) = 2n \exp(-\lambda^2/2)$

- $P_F(5, 10) \approx 0.00007$ ... but $P_F(5, 10^5) \approx 0.7$!

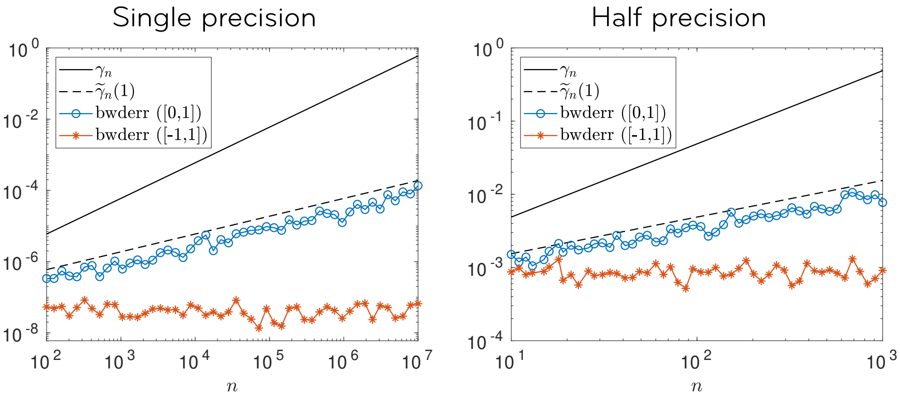
$\Rightarrow$ Crucial to keep the probabilities independent of n!
Fortunately:

$$P_F(\lambda, O(n^p)) = O(1) \quad \Leftrightarrow \quad \lambda = O(\sqrt{p \log n})$$

$\Rightarrow$ Error bound grows no faster than $\sqrt{n \log n}u$ for algorithms of polynomial complexity

- Moreover the constant hidden in the big O is small:
$$P_F(10, 10^{10}) \approx 4 \times 10^{-12}$$

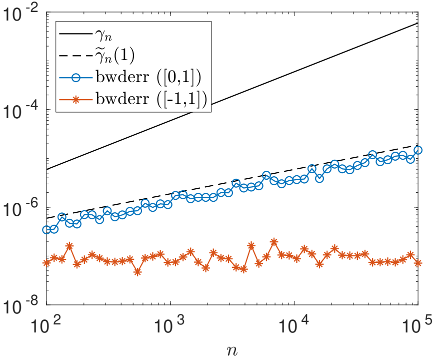
Numerical experiments with summation



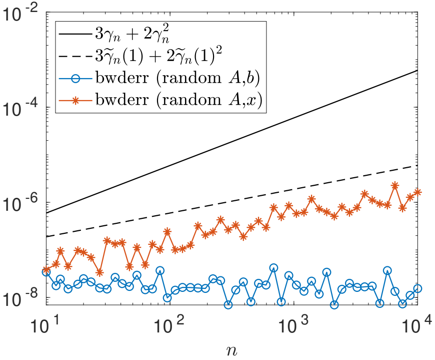
- Able to guarantee backward stability for a wider range of problems in a probabilistic sense
- $\widetilde{\gamma}_n$ is not always asymptotically sharp: error does not grow with n for $[-1, 1]$ data (we will come back to this observation later)

Experimental results with NLA algorithms

Matrix–vector product
(single precision)

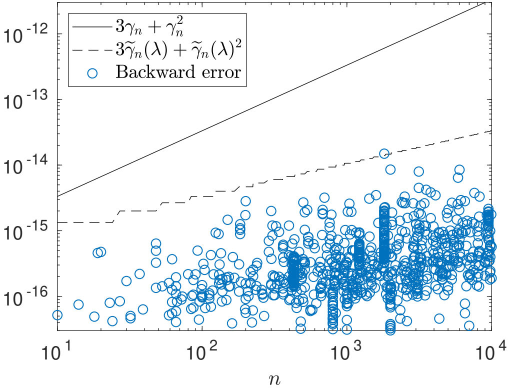


Solution to $Ax = b$
(single precision)



Experimental results with real-life matrices

Solution of $Ax = b$ (double precision),
for 943 matrices from the SuiteSparse collection



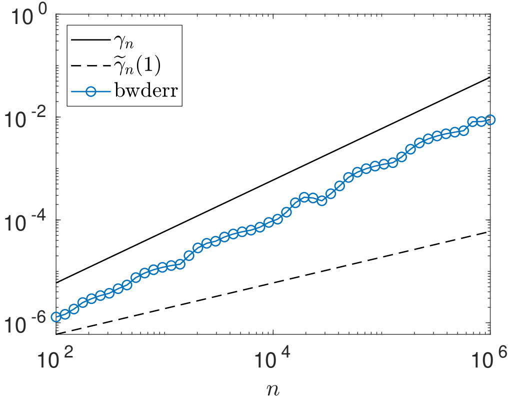
$\Rightarrow$ Probabilistic bound is thus satisfied in many important cases...
but there are counter-examples

Example with dependent rounding errors

Summation with constant x_i:

$$s_i = s_{i-1} + c, \qquad i = 2\colon n$$

leads to an error growing as nu rather than $\sqrt{n}u$

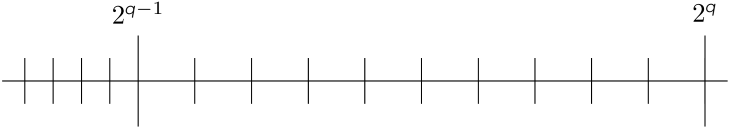


Exercise: explain what is happening

Example with dependent rounding errors (cont'd)

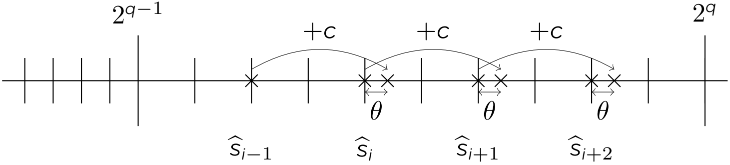
$$s_i = s_{i-1} + c \Rightarrow \quad \widehat{s}_i = (\widehat{s}_{i-1} + c)(1 + \delta_i)$$

Example with dependent rounding errors (cont'd)



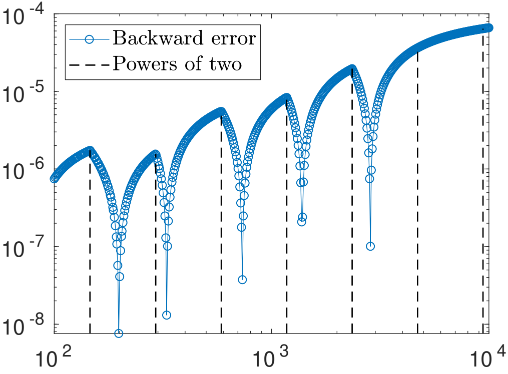
$$s_i = s_{i-1} + c \Rightarrow \quad \widehat{s}_i = (\widehat{s}_{i-1} + c)(1 + \delta_i)$$

Example with dependent rounding errors (cont'd)



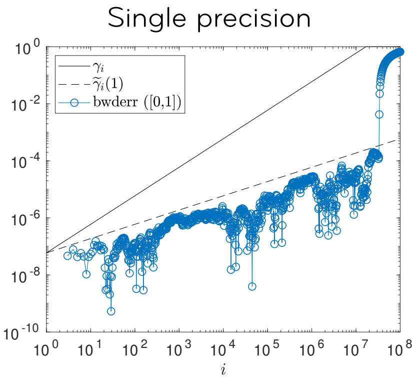
$$s_i = s_{i-1} + c \Rightarrow \quad \widehat{s}_i = (\widehat{s}_{i-1} + c)(1 + \delta_i)$$

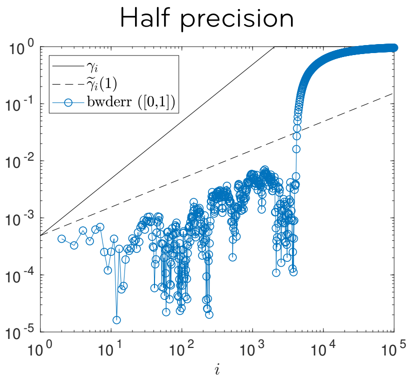
$\Rightarrow$ $\delta_i = \theta$ is constant within intervals $[2^{q-1}; 2^q]$



Example with rounding errors of nonzero mean

Summation of a very large number of nonnegative terms ($n \gg 10^3$ in half precision, $n \gg 10^7$ in single precision) leads to an error eventually growing like $O(nu)$





Exercise: explain what is happening

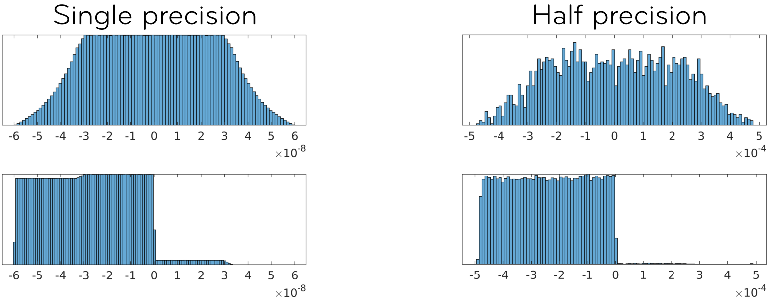
Example with rounding errors of nonzero mean (cont'd)

$$s_i = s_{i-1} + x_i \quad \Rightarrow \quad \widehat{s}_i = (\widehat{s}_{i-1} + x_i)(1 + \delta_i)$$

Example with rounding errors of nonzero mean (cont'd)

$$s_i = s_{i-1} + x_i \quad \Rightarrow \quad \widehat{s}_i = (\widehat{s}_{i-1} + x_i)(1 + \delta_i)$$

Distribution of the δ_i



Top: $1 \le i \le 3 \times 10^3$
Bottom: $3 \times 10^3 \le i \le 10^5$

Top: $1 \le i \le 3 \times 10^7$
Bottom: $3 \times 10^7 \le i \le 10^8$

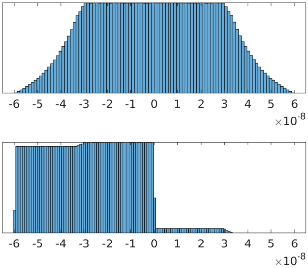
Example with rounding errors of nonzero mean (cont'd)

$$s_i = s_{i-1} + x_i \quad \Rightarrow \quad \widehat{s}_i = (\widehat{s}_{i-1} + x_i)(1+\delta_i)$$

Explanation: s_i keeps increasing, at some point, it becomes so large that $\widehat{s}_{i-1} \geq x_i/u$ and the computed sum **stagnates**: $\widehat{s}_i = \widehat{s}_{i-1}$.
Stagnation produces negative δ_i: indeed $\delta_i = -x_i/(\widehat{s}_{i-1} + x_i) < 0$

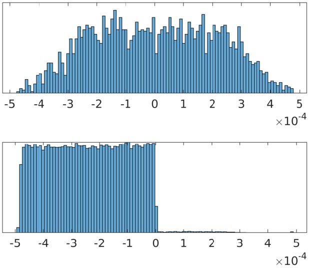
Distribution of the δ_i

Single precision



Top: $1 \leq i \leq 3 \times 10^3$
Bottom: $3 \times 10^3 \leq i \leq 10^5$

Half precision



Top: $1 \leq i \leq 3 \times 10^7$
Bottom: $3 \times 10^7 \leq i \leq 10^8$

Validity of the probabilistic bound

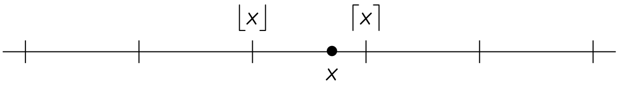
- The previous examples reveal situations in which the probabilistic bound is not valid, because the assumptions in the model are not satisfied

- Even though the analysis gives useful predictions, care is required in applying and interpreting the bound

... at least with a deterministic rounding mode such as **round to nearest** (which we have used so far)

⇒ what about **stochastic** rounding modes?

Stochastic rounding: definition



With round to nearest

$$\mathrm{fl}(x) = \begin{cases} \lceil x \rceil & \text{if } x - \lfloor x \rfloor > \lceil x \rceil - x \\ \lfloor x \rfloor & \text{otherwise} \end{cases}$$

Instead, with stochastic rounding

$$\mathrm{fl}(x) = \begin{cases} \lceil x \rceil & \text{with probability } p = \frac{x - \lfloor x \rfloor}{\lceil x \rceil - \lfloor x \rfloor} \\ \lfloor x \rfloor & \text{with probability } 1 - p = \frac{\lceil x \rceil - x}{\lceil x \rceil - \lfloor x \rfloor} \end{cases}$$

where $\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$ denote the operators that round down and up

Stochastic rounding $\Rightarrow$ zero mean δ_i

- Let $a, b \in \mathbb{R}$ and op $\in \{+, -, \times, \div\}$ such that

$$\mathrm{fl}(a \text{ op } b) = (a \text{ op } b)(1 + \delta)$$

- **Exercise:** prove that with stochastic rounding, $\mathbb{E}(\delta) = 0$

Stochastic rounding $\Rightarrow$ zero mean δ_i

- Let $a, b \in \mathbb{R}$ and op $\in \{+, -, \times, \div\}$ such that

$$\mathrm{fl}(a \text{ op } b) = (a \text{ op } b)(1+\delta)$$

- **Exercise:** prove that with stochastic rounding, $\mathbb{E}(\delta) = 0$
- Let $x := a \text{ op } b$; with stochastic rounding,

$$\mathbb{E}(\mathrm{fl}(x)) = \frac{\lceil x \rceil \big(x - \lfloor x \rfloor\big) + \lfloor x \rfloor \big(\lceil x \rceil - x\big)}{\lceil x \rceil - \lfloor x \rfloor}$$

$$= \frac{x\big(\lceil x \rceil - \lfloor x \rfloor\big)}{\lceil x \rceil - \lfloor x \rfloor} = x$$

- The expected value of the computed result is the exact result

$$\begin{aligned} & \mathbb{E}(\mathrm{fl}(a \text{ op } b)) = a \text{ op } b \\ \Rightarrow \quad & \mathbb{E}((a \text{ op } b)(1+\delta)) = a \text{ op } b \\ \Rightarrow \quad & (a \text{ op } b)\,\mathbb{E}(\delta) = 0 \\ \Rightarrow \quad & \mathbb{E}(\delta) = 0 \quad \text{if } a \text{ op } b \neq 0 \end{aligned}$$

$\Rightarrow$ **Stochastic rounding enforces zero mean rounding errors**

Stochastic rounding $\not\Rightarrow$ independent δ_i

- Consider the computation of $s := (a+b)+c$

$$\widehat{s} = \mathrm{fl}(\mathrm{fl}(a+b)+c) = \big((a+b)(1+\delta_1)+c\big)(1+\delta_2)$$

- Define $\widehat{s}_1 = \mathrm{fl}(a+b)+c = (a+b)(1+\delta_1)+c$

- Then, $\delta_2 = \widehat{s} - \widehat{s}_1$ is entirely determined by

$$\delta_2 = \begin{cases} \lceil \widehat{s}_1 \rceil - \widehat{s}_1 \text{ with probability } p = (\widehat{s}_1 - \lfloor \widehat{s}_1 \rfloor)/(\lceil \widehat{s}_1 \rceil - \lfloor \widehat{s}_1 \rfloor), \\ \lfloor \widehat{s}_1 \rfloor - \widehat{s}_1 \text{ with probability } 1-p \end{cases}$$

which clearly depends on $\widehat{s}_1$ and so on δ_1

$\Rightarrow$ **Even with stoch. rounding, rounding errors may be dependent**

Stochastic rounding $\Rightarrow$ mean independent δ_i

- Consider the computation of $s = \widehat{a} \text{ op } \widehat{b}$, where the computation of $\widehat{a}$ and $\widehat{b}$ has already produced k rounding errors $\delta_1, \dots, \delta_k$
- Then, $\widehat{s} = \mathrm{fl}(\widehat{a} \text{ op } \widehat{b}) = (\widehat{a} \text{ op } \widehat{b})(1 + \delta_{k+1})$ and $\delta_{k+1} = \widehat{s} - s$ (which depends on $\delta_1, \dots, \delta_k$) is given by

$$\delta_{k+1} = \begin{cases} \lceil s \rceil - s \text{ with probability } p = \frac{s - \lfloor s \rfloor}{\lceil s \rceil - \lfloor s \rfloor} \\ \lfloor s \rfloor - s \text{ with probability } 1 - p = \frac{\lceil s \rceil - s}{\lceil s \rceil - \lfloor s \rfloor} \end{cases}$$

- Since $\lceil s \rceil - s$ and $\lfloor s \rfloor - s$ are entirely determined by $\delta_1, \dots, \delta_k$

$$\mathbb{E}(\lceil s \rceil - s \mid \delta_1, \dots, \delta_k) = \lceil s \rceil - s$$
$$\mathbb{E}(\lfloor s \rfloor - s \mid \delta_1, \dots, \delta_k) = \lfloor s \rfloor - s$$

  where $\mathbb{E}(X \mid Y)$ denotes the conditional expectation of X given Y
- Therefore we obtain

$$\begin{aligned} \mathbb{E}\left(\delta_{k+1} \mid \delta_1, \dots, \delta_k\right) &= p\, \mathbb{E}\left(\lceil s \rceil - s \mid \delta_1, \dots, \delta_k\right) + (1-p)\, \mathbb{E}\left(\lfloor s \rfloor - s \mid \delta_1, \dots, \delta_k\right) \\ &= p(\lceil s \rceil - s) + (1-p)(\lfloor s \rfloor - s) = 0 \end{aligned}$$

$\Rightarrow$ **Stochastic rounding enforces mean independence:**

$$\mathbb{E}(\delta_i \mid \delta_1, \dots, \delta_{i-1}) = \mathbb{E}(\delta_i) \ (= 0)$$

Martingales and Azuma–Hoeffding inequality

- A sequence of random variables E_0, ..., E_n is called a **martingale** if, for all k, $\mathbb{E}(|E_k|) < \infty$ and $\mathbb{E}(E_{k+1} \mid E_0, \ldots, E_k) = E_k$
- Example: random walks are martingales. Position at step $k+1$ depends on previous positions but, if all directions have equal probabilities, its expected value is the position at step k

Azuma–Hoeffding inequality

Let E_0, ..., E_n be a martingale such that $|E_{k+1} - E_k| \leq c$, for $k = 0\colon n-1$. Then, for any $\lambda > 0$,

$$\Pr\left(|E_n - E_0| \geq \lambda\sqrt{n}c\right) \leq 2\exp(-\lambda^2/2)$$

- Azuma–Hoeffding generalizes Hoeffding's inequality to possibly dependent random variables

$\Rightarrow$ Can we relax our model to mean independent δ_i?

An even more general probabilistic model

Probabilistic model of rounding errors

Let the computation of interest generate rounding errors δ_1, δ_2, ... in that order. The δ_k are (possibly dependent) random variables of mean zero and mean independent of the previous δ_1, ..., δ_{k-1}, i.e., $\mathbb{E}(\delta_k \mid \delta_1, \ldots, \delta_{k-1}) = \mathbb{E}(\delta_k) = 0$.

- **Exercise:** prove that with this model, we recover the probabilistic bound $\prod_{i=1}^{n}(1+\delta_i) = 1+\theta_n, \quad |\theta_n| \le \widetilde{\gamma}_n(\lambda)$

An even more general probabilistic model

Probabilistic model of rounding errors

Let the computation of interest generate rounding errors δ_1, δ_2, ... in that order. The δ_k are (possibly dependent) random variables of mean zero and mean independent of the previous δ_1, ..., δ_{k-1}, i.e., $\mathbb{E}(\delta_k \mid \delta_1, \ldots, \delta_{k-1}) = \mathbb{E}(\delta_k) = 0$.

- **Exercise:** prove that with this model, we recover the probabilistic bound $\prod_{i=1}^{n}(1+\delta_i) = 1 + \theta_n, \quad |\theta_n| \le \widetilde{\gamma}_n(\lambda)$
- $E_n = \sum_{i=1}^{n} \delta_i$ (with $E_0 = 0$) is a martingale

An even more general probabilistic model

Probabilistic model of rounding errors

Let the computation of interest generate rounding errors δ_1, δ_2, ... in that order. The δ_k are (possibly dependent) random variables of mean zero and mean independent of the previous δ_1, ..., δ_{k-1}, i.e., $\mathbb{E}(\delta_k \mid \delta_1, \dots, \delta_{k-1}) = \mathbb{E}(\delta_k) = 0$.

- **Exercise:** prove that with this model, we recover the probabilistic bound $\prod_{i=1}^{n}(1+\delta_i) = 1+\theta_n, \quad |\theta_n| \le \widetilde{\gamma}_n(\lambda)$
- $E_n = \sum_{i=1}^{n} \delta_i$ (with $E_0 = 0$) is a martingale
  $\Rightarrow$ Clearly $|E_k| \le ku \Rightarrow \mathbb{E}(|E_k|) < \infty$ and

$$\begin{aligned}\mathbb{E}(E_{k+1} \mid E_0, \dots, E_k) &= \mathbb{E}(E_k + \delta_{k+1} \mid \delta_1, \dots, \delta_k) \\ &= \mathbb{E}(E_k \mid \delta_1, \dots, \delta_k) + \mathbb{E}(\delta_{k+1} \mid \delta_1, \dots, \delta_k) = E_k\end{aligned}$$

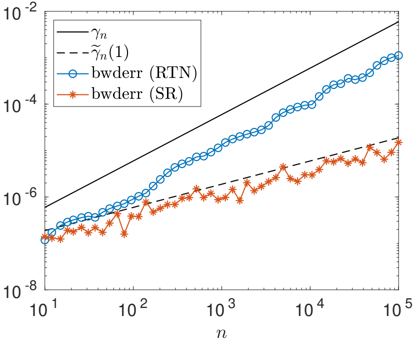
- Azuma–Hoeffding: $|E_{k+1} - E_k| \le u \Rightarrow |E_n - E_0| = |E_n| \le \lambda\sqrt{n}u$
- By Taylor expansions

$$E_n - \frac{nu^2}{1-u} \le \sum_{i=1}^{n} \log(1+\delta_i) \le E_n + \frac{nu^2}{1-u}$$

Stochastic rounding enforces probabilistic bound

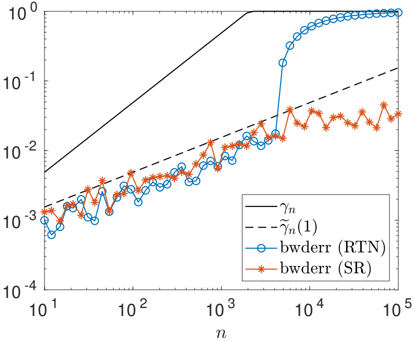
Conclusion: with stochastic rounding,
the probabilistic bound holds rigorously, with no exceptions

Constant x_i (single precision)



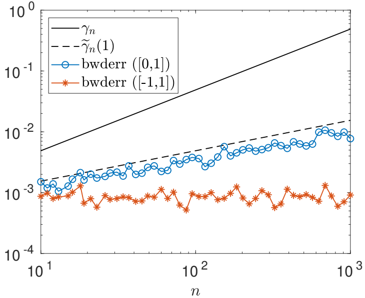
⇒ stochastic rounding produces nonconstant δ_i

Nonzero x_i (half precision)



⇒ stochastic rounding overcomes stagnation

Probabilistic model of the data



We now come back to the observed difference between $[0,1]$ and $[-1,1]$ data

We seek to obtain a **sharper** backward error bound by taking into account the distribution of the x_i

Probabilistic model of the data

The x_i, $i = 1:n$, are independent random variables sampled from a given distribution of mean μ_x and satisfy $|x_i| \le C_x$. In summary, we assume δ_k, x_k, and $x_k\delta_k$ to be mean independent of previous δ_i and x_i:

$$\mathbb{E}(\delta_k \mid \delta_1, \dots, \delta_{k-1}, x_1, \dots, x_{k-1}) = \mathbb{E}(\delta_k) = 0$$

$$\mathbb{E}(x_k \mid \delta_1, \dots, \delta_{k-1}, x_1, \dots, x_{k-1}) = \mathbb{E}(x_k) = \mu_x$$

$$\mathbb{E}(x_k\delta_k \mid \delta_1, \dots, \delta_{k-1}, x_1, \dots, x_{k-1}) = \mathbb{E}(x_k\delta_k) = 0$$

Note that $\mathbb{E}(x_k\delta_k) = 0$ is a consequence of the law of total expectation:

$$\mathbb{E}(x_k\delta_k) = \mathbb{E}(\mathbb{E}(x_k\delta_k|x_k)) = \mathbb{E}(x_k\,\mathbb{E}(\delta_k|x_k))\,\mathbb{E}(x_k\,\mathbb{E}(\delta_k)) = \mathbb{E}(x_k)\,\mathbb{E}(\delta_k) = 0$$

Sharper probabilistic bound for summation

Exercise: compute a bound on the backward error $\frac{|\widehat{s}-s|}{\sum_{i=1}^{n}|x_i|}$

1. Compute an upper bound on $|s_j| = |\sum_{i=1}^{j} x_i|$

2. Compute an upper bound on $|E_n| = |\sum_{j=2}^{n} s_j \delta_j|$

3. Compute a lower bound $\sum_{i=1}^{n} |x_i|$ as a function of $\mu_{|x|}$

Sharper probabilistic bound for summation

Exercise: compute a bound on the backward error $\frac{|\widehat{s}-s|}{\sum_{i=1}^{n}|x_i|}$

1. Compute an upper bound on $|s_j| = |\sum_{i=1}^{j} x_i|$
   $\Rightarrow$ Hoeffding: $|s_j| \le \mu_x j + \lambda C_x \sqrt{j}$ with prob. $1 - 2\exp(-\lambda^2/2)$
2. Compute an upper bound on $|E_n| = |\sum_{j=2}^{n} s_j \delta_j|$
3. Compute a lower bound $\sum_{i=1}^{n}|x_i|$ as a function of $\mu_{|x|}$

Sharper probabilistic bound for summation

Exercise: compute a bound on the backward error $\frac{|\widehat{s}-s|}{\sum_{i=1}^{n}|x_i|}$

1. Compute an upper bound on $|s_j| = |\sum_{i=1}^{j} x_i|$

   $\Rightarrow$ Hoeffding: $|s_j| \le \mu_x j + \lambda C_x \sqrt{j}$ with prob. $1 - 2\exp(-\lambda^2/2)$

2. Compute an upper bound on $|E_n| = |\sum_{j=2}^{n} s_j \delta_j|$
   - Clearly $|E_k| \le k(k-1)C_x u \Rightarrow \mathbb{E}(|E_k|) < \infty$
   - Let $\mathcal{S}_k = \{\delta_1, \ldots, \delta_{k-1}, x_1, \ldots, x_{k-1}\}$, we have

$$\begin{aligned}\mathbb{E}(E_k \mid E_0, \ldots, E_{k-1}) &= \mathbb{E}(E_{k-1} + s_k\delta_k \mid \mathcal{S}_k) \\ &= \mathbb{E}(E_{k-1} \mid \mathcal{S}_k) + \mathbb{E}(s_{k-1}\delta_k \mid \mathcal{S}_k) + \mathbb{E}(x_k\delta_k \mid \mathcal{S}_k) \\ &= E_{k-1} + s_{k-1}\,\mathbb{E}(\delta_k \mid \mathcal{S}_k) + \mathbb{E}(x_k\delta_k \mid \mathcal{S}_k) = E_{k-1}\end{aligned}$$

   - $E_0, \ldots, E_n$ is a martingale and

$$\forall k \le n-1,\ \Pr\Big(|E_{k+1} - E_k| \ge (\mu_x n + \lambda C_x \sqrt{n})u\Big) \le 2\exp(-\lambda^2/2)$$

$$\Rightarrow \Pr\Big(\forall k \le n-1,\ |E_{k+1} - E_k| \ge (\mu_x n + \lambda C_x \sqrt{n})u\Big) \le 2(n-1)\exp(-\lambda^2/2)$$

   $\Rightarrow$ By Azuma–Hoeffding, we obtain with prob. $1 - 2n\exp(-\lambda^2/2)$

$$|\widehat{s} - s| = |E_n - E_0| \le \big(\lambda\mu_x n^{3/2} + \lambda^2 C_x n\big)u$$

3. Compute a lower bound $\sum_{i=1}^{n}|x_i|$ as a function of $\mu_{|x|}$

Sharper probabilistic bound for summation

Exercise: compute a bound on the backward error $\frac{|\widehat{s}-s|}{\sum_{i=1}^{n}|x_i|}$

1. Compute an upper bound on $|s_j| = |\sum_{i=1}^{j} x_i|$

   $\Rightarrow$ Hoeffding: $|s_j| \le \mu_x j + \lambda C_x \sqrt{j}$ with prob. $1-2\exp(-\lambda^2/2)$

2. Compute an upper bound on $|E_n| = |\sum_{j=2}^{n} s_j \delta_j|$
   - Clearly $|E_k| \le k(k-1)C_x u \Rightarrow \mathbb{E}(|E_k|) < \infty$
   - Let $\mathcal{S}_k = \{\delta_1, \dots, \delta_{k-1}, x_1, \dots, x_{k-1}\}$, we have

$$\begin{aligned}\mathbb{E}(E_k \mid E_0, \dots, E_{k-1}) &= \mathbb{E}(E_{k-1} + s_k\delta_k \mid \mathcal{S}_k) \\ &= \mathbb{E}(E_{k-1} \mid \mathcal{S}_k) + \mathbb{E}(s_{k-1}\delta_k \mid \mathcal{S}_k) + \mathbb{E}(x_k\delta_k \mid \mathcal{S}_k) \\ &= E_{k-1} + s_{k-1}\,\mathbb{E}(\delta_k \mid \mathcal{S}_k) + \mathbb{E}(x_k\delta_k \mid \mathcal{S}_k) = E_{k-1}\end{aligned}$$

   - $E_0, \dots, E_n$ is a martingale and

$$\forall k \le n{-}1,\ \Pr\Big(|E_{k+1} - E_k| \ge (\mu_x n + \lambda C_x\sqrt{n})u\Big) \le 2\exp(-\lambda^2/2)$$

$$\Rightarrow \Pr\Big(\forall k \le n{-}1,\ |E_{k+1} - E_k| \ge (\mu_x n + \lambda C_x\sqrt{n})u\Big) \le 2(n{-}1)\exp(-\lambda^2/2)$$

   $\Rightarrow$ By Azuma–Hoeffding, we obtain with prob. $1-2n\exp(-\lambda^2/2)$

$$|\widehat{s} - s| = |E_n - E_0| \le \big(\lambda\mu_x n^{3/2} + \lambda^2 C_x n\big)u$$

3. Compute a lower bound $\sum_{i=1}^{n}|x_i|$ as a function of $\mu_{|x|}$

   $\Rightarrow$ Hoeffding: $\sum_{i=1}^{n}|x_i| \ge n\mu_{|x|} - \lambda C_x\sqrt{n}$ with prob. $1-2\exp(-\lambda^2/2)$

Sharper probabilistic bound for summation (cont'd)

Main result

Under the previously stated models of rounding errors and data,

$$\varepsilon_{bwd} = \frac{|\widehat{s} - s|}{\sum_{i=1}^n |x_i|} \le \frac{\lambda \mu_x \sqrt{n} + \lambda^2 C_x}{\mu_{|x|} - \lambda C_x/\sqrt{n}} \cdot u + O(u^2)$$

holds with probability $P(\lambda) = 1 - 2(n+1)\exp\left(-\lambda^2/2\right)$

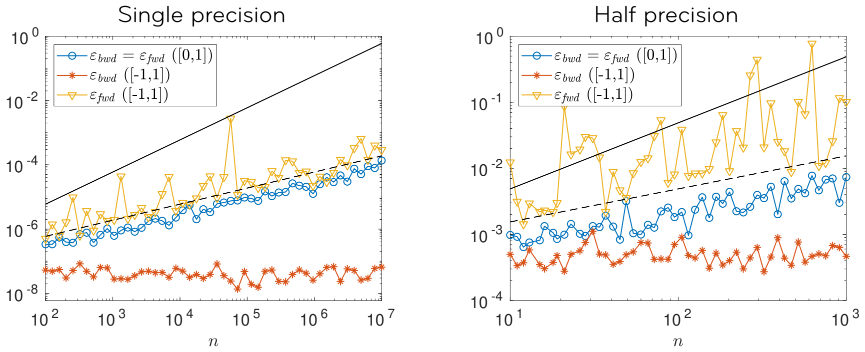
- $\mu_x = O(1) \Rightarrow \varepsilon_{bwd} = O(\sqrt{n}u)$
- $\mu_x = 0$ or $\mu_x \ll 1 \Rightarrow \varepsilon_{bwd} = O(u)$

$$\varepsilon_{fwd} = \kappa \cdot \varepsilon_{bwd}, \qquad \kappa = \frac{\sum_{i=1}^n |x_i|}{\left|\sum_{i=1}^n x_i\right|}$$

What probabilistic bound on κ for random x_i?

- If $\mu_x = \Theta(1)$, $|\sum_{i=1}^n x_i| = \Theta(n)$ and $\sum_{i=1}^n |x_i| = \Theta(n) \Rightarrow \kappa = \Theta(1)$
- If $\mu_x = o(1)$, $|\sum_{i=1}^n x_i| = O(\sqrt{n})$ and $\sum_{i=1}^n |x_i| = \Theta(n) \Rightarrow \kappa = \Omega(\sqrt{n})$

Summary of probabilistic bounds for summation



	General δ_i	Probabilistic model on δ_i (enforced by SR)		
		General x_i	Probabilistic model on x_i	
			$\mu_x \neq 0$	$\mu_x = 0$
κ	—	—	$\Theta(1)$	$\Omega(\sqrt{n})$
ε_{bwd}	nu	$O(\sqrt{n})u$	$\Theta(\sqrt{n})u$	$\Theta(1)u$
ε_{fwd}	$n\kappa u$	$O(\sqrt{n})\kappa u$	$\Theta(\sqrt{n})u$	$\Omega(\sqrt{n})u$

$\sqrt{n}u$ is too large for large n and u! $\Rightarrow$ **we need new algorithms with smaller error bounds** (another story for another time)

Conclusion

Take home message

With the emergence of **large scale computations** and **low precision arithmetics**, **classical analyses can no longer guarantee the numerical stability of classical algorithms**
$\Rightarrow$ **Probabilistic analyses are a powerful and timely tool to obtain sharper error bounds**

Lecture slides available on my webpage

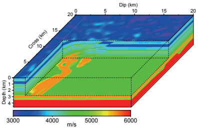
`bit.ly/tmaryLIP6`

To go further

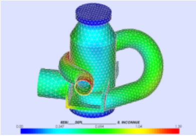
Nicholas J. Higham and Theo Mary, *A New Approach to Probabilistic Rounding Error Analysis*, SIAM Journal on Scientific Computing, 41(5):A2815-A2835, 2019 (PDF available here).

Sujet de stage+thèse: résolution parallèle de problèmes industriels avec techniques de compression multiprécisions (contact: theo.mary@lip6.fr)

- Contexte: les grands défis actuels de la simulation numérique de problèmes industriels de très grande taille requièrent des solveurs linéaires efficaces. Des algorithmes de compression Block Low-Rank (BLR) ont permis de réduire le côut de cette résolution sur architectures parallèles (jusqu'à 2400 coeurs). Exemple de facteurs de compression:



Imagerie sismique
Taille: 130 millions
Précision cible: 10^{-3}
Facteur de compression: 41



Mécanique struct.
Taille: 31 millions
Précision cible: 10^{-9}
Facteur de compression: 26

- Objectif: exploiter les architectures parallèles émergentes conçues pour l'IA (ex: GPU tensor cores) pour décupler ces gains de performance en concevant des algorithmes multiprécisions efficaces qui combinent compression et précisions faibles
- Environnement: le stage (qui pourra se poursuivre en thèse) s'effectuera à Paris (LIP6 et EDF Paris-Saclay) en collaboration avec le groupe MUMPS (Lyon et Toulouse)