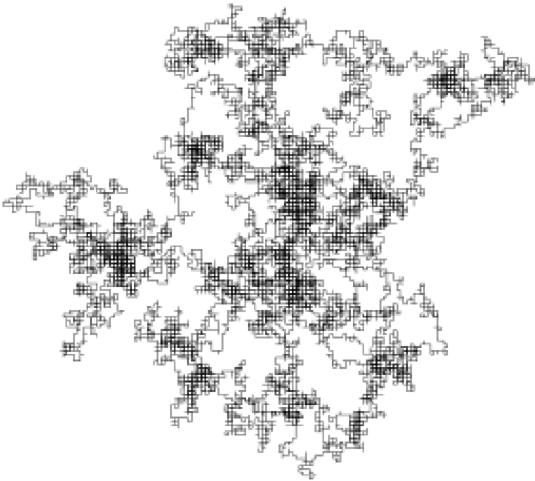
Harnessing inexactness in scientific computing

Lecture 4: probabilistic error analysis

Theo Mary (CNRS)
theo.mary@lip6.fr
https://perso.lip6.fr/Theo.Mary/

Elisa Riccietti (ENS Lyon)
elisa.riccietti@ens-lyon.fr
https://perso.ens-lyon.fr/elisa.riccietti/

M2 course at ENS Lyon, 2024–2025
Slides available on course webpage



Backward error analysis

- Backward error analysis recasts the rounding errors as perturbations of the input data
- Example for summation:

$$
\begin{aligned}
& s_2 &&= x_1 + x_2 \\
\Rightarrow \quad & \widehat{s}_2 &&= (x_1 + x_2)(1+\delta_1) = x_1(1+\delta_1) + x_2(1+\delta_1) \\
& s_3 &&= \widehat{s}_2 + x_3 \\
\Rightarrow \quad & \widehat{s}_3 &&= (\widehat{s}_2 + x_3)(1+\delta_2) \\
& &&= x_1(1+\delta_1)(1+\delta_2) + x_2(1+\delta_1)(1+\delta_2) + x_3(1+\delta_2) \\
& \dots && \\
\Rightarrow \quad & \widehat{s}_n &&= \sum_{i=1}^{n} \left[x_i \prod_{k=k_i}^{n} (1+\delta_k) \right]
\end{aligned}
$$

Backward error analysis

- Backward error analysis recasts the rounding errors as perturbations of the input data
- Example for summation:

$$
\begin{aligned}
& s_2 &&= x_1 + x_2 \\
\Rightarrow\quad & \widehat{s}_2 &&= (x_1 + x_2)(1+\delta_1) = x_1(1+\delta_1) + x_2(1+\delta_1) \\
& s_3 &&= \widehat{s}_2 + x_3 \\
\Rightarrow\quad & \widehat{s}_3 &&= (\widehat{s}_2 + x_3)(1+\delta_2) \\
& &&= x_1(1+\delta_1)(1+\delta_2) + x_2(1+\delta_1)(1+\delta_2) + x_3(1+\delta_2) \\
& \ldots && \\
\Rightarrow\quad & \widehat{s}_n &&= \sum_{i=1}^{n} \left[x_i \prod_{k=k_i}^{n} (1+\delta_k) \right]
\end{aligned}
$$

Worst-case fundamental lemma

Let δ_k, $k = 1 : n$, such that $|\delta_k| \le u$ and $nu < 1$. Then

$$\prod_{k=1}^{n} (1+\delta_k) = 1 + \theta_n, \quad |\theta_n| \le \gamma_n := \frac{nu}{1-nu}.$$

Backward stability in linear algebra

Most linear algebra computations have backward stable implementations:

- Inner products
$$\widehat{s} = (x + \Delta x)^T y, \qquad |\Delta x| \le \gamma_n |x|$$
- Matrix–vector products
$$\widehat{y} = (A + \Delta A)x, \qquad |\Delta A| \le \gamma_n |A|$$
- LU factorization*
$$\widehat{L}\widehat{U} = A + \Delta A, \qquad |\Delta A| \le \gamma_n |\widehat{L}||\widehat{U}|$$
- Triangular systems
$$(T + \Delta T)\widehat{x} = b, \qquad |\Delta T| \le \gamma_n |T|$$
- Linear systems*
$$(A + \Delta A)\widehat{x} = b, \qquad |\Delta A| \le (3\gamma_n + \gamma_n^2)|\widehat{L}||\widehat{U}|$$

(* backward stable only if $\||\widehat{L}||\widehat{U}|\| \approx \|A\|$, need stable pivoting strategy for Gaussian elimination)

Genesis of backward error analysis

- 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"

James Wilkinson

Nicholas Higham

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 problems

Since the 1990s, the TOP500 list ranks the world's most powerful supercomputers based on how fast they can solve a dense linear system of equations $Ax = b$

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,730,112	1,102.00	1,685.65	21,100
2	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442.01	537.21	29,899
3	**LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE EuroHPC/CSC Finland	1,110,144	151.90	214.35	2,942

November 2023: Frontier achieves 1.1 ExaFLOPS by solving system with $n =$ **22 millions**!

Today: low precisions

		number of bits			
		signif. (t)	exp.	range	$u = 2^{-t}$
fp128	quadruple	113	15	$10^{\pm 4932}$	1×10^{-34}
fp64	double	53	11	$10^{\pm 308}$	1×10^{-16}
fp32	single	24	8	$10^{\pm 38}$	6×10^{-8}
fp16	half	11	5	$10^{\pm 5}$	5×10^{-4}
bfloat16	half	8	8	$10^{\pm 38}$	4×10^{-3}
fp8 (e4m3)	quarter	4	4	$10^{\pm 2}$	6×10^{-2}
fp8 (e5m2)	quarter	3	5	$10^{\pm 5}$	1×10^{-1}

Half (16-bit) and quarter (8-bit) precision now in hardware, driven by AI

Can low precision extreme scale computations be accurate ?

Understanding rounding error accumulation

To limit it, we need to understand rounding error accumation: when and why does it happen?

Let us measure the actual backward error, which is given by

$$\eta = \min\left\{\epsilon > 0 : \widehat{s} = (x+\Delta x)^T y, \quad |\Delta x| \le \epsilon |x|\right\} = \frac{|\widehat{s} - s|}{|x|^T |y|}$$

and compare it to its bound γ_n

Understanding rounding error accumulation

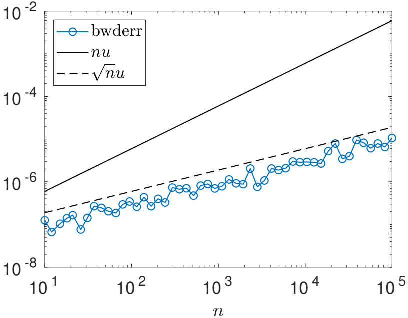
To limit it, we need to understand rounding error accumation: when and why does it happen?

Let us measure the actual backward error, which is given by

$$\eta = \min\left\{\epsilon > 0 : \widehat{s} = (x + \Delta x)^T y, \quad |\Delta x| \le \epsilon |x|\right\} = \frac{|\widehat{s} - s|}{|x|^T |y|}$$

and compare it to its bound γ_n

Inner product in single precision
with random uniform $[0, 1]$ vectors



Understanding rounding error accumulation

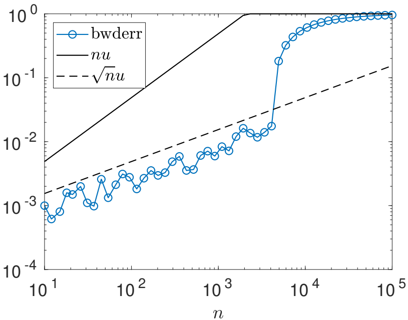
To limit it, we need to understand rounding error accumation: when and why does it happen?

Let us measure the actual backward error, which is given by

$$\eta = \min \left\{ \epsilon > 0 : \widehat{s} = (x + \Delta x)^T y, \quad |\Delta x| \le \epsilon |x| \right\} = \frac{|\widehat{s} - s|}{|x|^T |y|}$$

and compare it to its bound γ_n

Inner product in half precision
with random uniform $[0, 1]$ vectors



Understanding rounding error accumulation

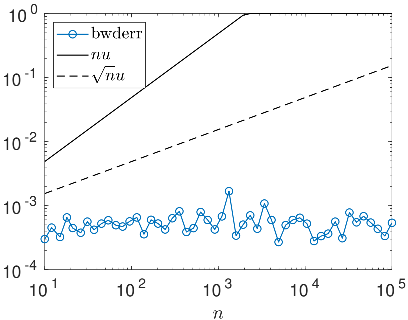
To limit it, we need to understand rounding error accumation: when and why does it happen?

Let us measure the actual backward error, which is given by

$$\eta = \min\left\{\epsilon > 0 : \widehat{s} = (x+\Delta x)^T y, \quad |\Delta x| \le \epsilon |x|\right\} = \frac{|\widehat{s} - s|}{|x|^T|y|}$$

and compare it to its bound γ_n

Inner product in half precision
with random uniform $[-1, 1]$ vectors



Understanding rounding error accumulation

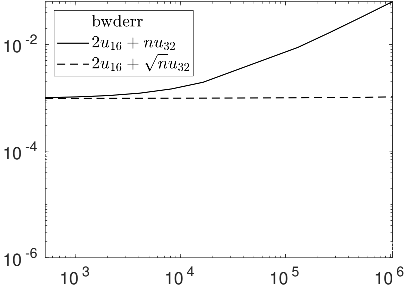
To limit it, we need to understand rounding error accumation: when and why does it happen?

Let us measure the actual backward error, which is given by

$$\eta = \min \left\{ \epsilon > 0 : \widehat{s} = (x + \Delta x)^T y, \quad |\Delta x| \leq \epsilon |x| \right\} = \frac{|\widehat{s} - s|}{|x|^T |y|}$$

and compare it to its bound γ_n

Inner product with tensor cores (lecture 16)
with random uniform $[-1, 1]$ vectors
Error bound $2u_{16} + nu_{32}$



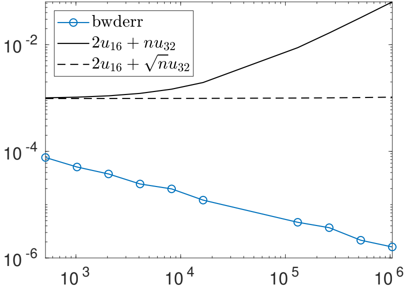
To limit it, we need to understand rounding error accumation: when and why does it happen?

Let us measure the actual backward error, which is given by

$$\eta = \min\left\{\epsilon > 0 : \widehat{s} = (x+\Delta x)^T y, \quad |\Delta x| \le \epsilon |x|\right\} = \frac{|\widehat{s} - s|}{|x|^T|y|}$$

and compare it to its bound γ_n

Inner product with tensor cores (lecture 16)
with random uniform $[-1,1]$ vectors
Error bound $2u_{16} + nu_{32}$



Probabilistic error analysis

Goal: develop probabilistic analyses to obtain improved bounds that are more realistic for the average computation

- Bounds for **random data**: specialize data-independent bounds to random data
- Bounds for **random errors**: improve bounds by modelling the rounding errors as random
- Bounds for **random data and errors**: both of the above

Basic properties of $\mathbb{E}$ operator

Let X, Y be random variables and a a constant.
The $\mathbb{E}$ operator satisfies the following properties:

- Linearity: $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$ and $\mathbb{E}(aX) = a\,\mathbb{E}(X)$
- Monotonocity: $X \leq Y \Rightarrow \mathbb{E}(X) \leq \mathbb{E}(Y)$
- Non-multiplicativity: in general, $\mathbb{E}(XY) \neq \mathbb{E}(X)\,\mathbb{E}(Y)$. If X and Y are independent, then equality holds.

Hoeffding's inequality

Let $X_0, \ldots, X_n$ be independent random variables satisfying $|X_k| \leq c_k$ for $k = 0 : n-1$. Then, the sum $S = \sum_{k=1}^{n} X_k$ satisfies, for any $\lambda > 0$,

$$\Pr\left(|S - \mathbb{E}(S)| \geq \lambda \left(\sum_{k=1}^{n} c_k^2 \right)^{1/2} \right) \leq 2 \exp(-\lambda^2/2)$$

- Better than the worst-case bound $\sum_{i=1}^{n} c_k$
- If $\forall k, c_k = c$: $nc \to \sqrt{nc}$
- Small values of λ suffice

Sum of random data

Let $S = \sum_{i=1}^{n} X_i$. Then by Hoeffding's inequality:

- If $X_i \sim U([0,1])$:
  - $|X_i| \leq 1$
  - $\mathbb{E}(S) = \frac{n}{2}$

  $\Rightarrow$ $|S| = \frac{n}{2} \pm \lambda\sqrt{n} \approx \frac{n}{2}$
- If $X_i \sim U([-1,1])$:
  - $|X_i| \leq 1$
  - $\mathbb{E}(S) = 0$

  $\Rightarrow$ $|S| \leq \lambda\sqrt{n}$, but no lower bound ($|S|$ can be close to 0 with non-small probability)

Let $T = \sum_{i=1}^{n} |X_i|$. Then by Hoeffding's inequality:

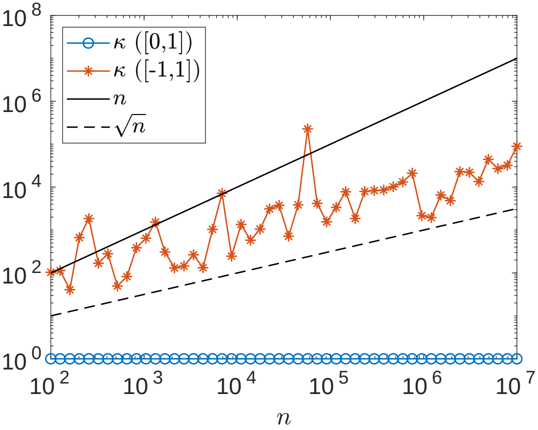
- If $X_i \sim U([0,1])$: no change, $T \approx \frac{n}{2}$
- If $X_i \sim U([-1,1])$: $|X_i| \sim U([0,1])$ and so $T \approx \frac{n}{2}$

Conditioning of random data

The conditioning of $S = \sum_{i=1}^{n} X_i$ is

$$\kappa = \frac{\sum_{i=1}^{n} |X_i|}{|\sum_{i=1}^{n} X_i|} = \frac{T}{|S|}$$

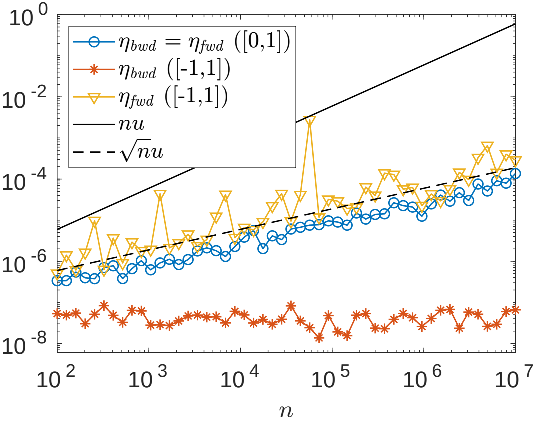
- If $X_i \sim U([0,1])$: $|S|, T \approx \frac{n}{2} \Rightarrow \kappa \approx 1$
- If $X_i \sim U([-1,1])$: $|S| \leq \lambda\sqrt{n}$ and $T \approx \frac{n}{2} \Rightarrow \kappa \geq \frac{\sqrt{n}}{2\lambda}$



Backward and forward errors for random data

$\eta_{\text{fwd}} = \kappa\eta_{\text{bwd}}$ and $\eta_{\text{bwd}} \le nu$:

- If $X_i \sim U([0,1])$: $\eta_{\text{fwd}} \approx nu$?
- If $X_i \sim U([-1,1])$: $\eta_{\text{fwd}} \approx n^{3/2}u$?



Modelling rounding errors as random variables

- Worst-case nu bound attained when all n errors are equal to $+u$, or all equal to $-u$ (i.e., they all accumulate in the same direction)
- Since the 1960s, numerical analysts have tried modelling the δ_i as independent random variables to translate the intuition that this does not seem very likely: we can hope the $+u$ and $-u$ to cancel each other

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 & Swenson, 1966

The fact that rounding errors are neither random nor uncorrelated will not in itself preclude the possibility of modelling them usefully by uncorrelated random variables

William Kahan, 1996

William Kahan

Wilkinson's conjecture

Wilkinson's conjecture (1961)

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

Wilkinson's conjecture (1961)

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

Why only a conjecture? Some heuristic arguments based on CLT, but:

- First-order analyses ("$+O(u^2)$")
- Asymptotic statements ("for sufficiently large n")
- Unspecified probabilities ("with high probability")
- Only applicable to specific algorithms
- Unable to explain diversity of behaviors previously observed

A naive model

Naive model

In the computation of interest, the rounding errors δ_k are independent random variables of mean zero: $\mathbb{E}(\delta_k) = 0$.

- Rounding errors are clearly not independent, so the model is not applicable
- We need a weaker assumption called mean independence

Conditional expectation and mean independence

We will use the conditional expectation $\mathbb{E}(X \mid Y)$:

- $\mathbb{E}(X \mid Y)$ is a random variable which takes the value $\mathbb{E}(X \mid Y = y)$ when $Y = y$
- The $\mathbb{E}(X \mid Y)$ operator satisfies the properties above
- $\mathbb{E}(X \mid Y) = \mathbb{E}(X)$ if X and Y are independent
- $\mathbb{E}(X \mid Y) = X$ if X is a function of Y

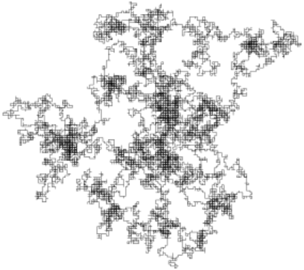
Model M

In the computation of interest, the rounding errors δ_k are mean independent random variables of mean zero: $\mathbb{E}(\delta_k \mid \delta_1, \dots, \delta_{k-1}) = \mathbb{E}(\delta_k) = 0$.

Martingales

- A sequence of random variables $E_0, \ldots, E_n$ is called a **martingale** with respect to $X_0, \ldots, X_n$ if, for all k,
  - E_k is a function of $X_0, \ldots, X_k$
  - $\mathbb{E}(|E_k|) < \infty$
  - $\mathbb{E}(E_{k+1} \mid X_0, \ldots, X_k) = E_k$

- 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's inequality

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

$$\Pr\left(|E_n - E_0| \geq \lambda \left(\sum_{k=1}^{n} c_k^2 \right)^{1/2} \right) \leq 2\exp(-\lambda^2/2)$$

Extends Hoeffding's inequality to mean independent variables

Sum of mean independent rouding errors

Model M

In the computation of interest, the rounding errors δ_k are mean independent random variables of mean zero: $\mathbb{E}(\delta_k \mid \delta_1, \ldots, \delta_{k-1}) = \mathbb{E}(\delta_k) = 0$.

- $E_n = \sum_{i=1}^{n} \delta_i$ (with $E_0 = 0$) is a martingale w.r.t. $\delta_1, \ldots, \delta_n$:
  E_k is a function of $\delta_1, \ldots, \delta_k$ and $|E_k| \leq ku \Rightarrow \mathbb{E}(|E_k|) < \infty$

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

- Azuma–Hoeffding: $|E_{k+1} - E_k| \leq u \Rightarrow |E_n - E_0| = |E_n| \leq \lambda\sqrt{n}u$ with probability $1 - 2\exp(-\lambda^2/2)$

Product of mean independent rouding errors

- We know how to bound $E_n = \sum_{i=1}^{n} \delta_i$, but what about $P_n = \prod_{i=1}^{n}(1+\delta_i)$?
- By the Taylor expansion

$$\log(1+x) = \sum_{i=1}^{\infty} (-1)^{i+1} \frac{x^i}{i}$$

we have

$$\delta_i - \frac{u^2}{1-u} \leq \log(1+\delta_i) \leq \delta_i + \frac{u^2}{1-u}$$

and thus

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

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

$$\frac{1}{\exp\left(\lambda\sqrt{n}u + \frac{nu^2}{1-u}\right)} \leq P_n \leq \exp\left(\lambda\sqrt{n}u + \frac{nu^2}{1-u}\right)$$

Probabilistic backward error analysis

Model M

In the computation of interest, the rounding errors δ_k are mean independent random variables of mean zero: $\mathbb{E}(\delta_k \mid \delta_1, \dots, \delta_{k-1}) = \mathbb{E}(\delta_k) = 0$.

Probabilistic fundamental lemma (Higham and M., 2019, 2020)

Let δ_k, $k = 1 : n$, satisfy Model M. Then, for any $\lambda > 0$, the relation

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

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

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

Probabilistic backward error analysis

Probabilistic fundamental lemma (Higham and M., 2019, 2020)

Let δ_k, $k = 1 : n$, satisfy Model M. Then, for any $\lambda > 0$, the relation

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

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

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

Key features:

- valid to all orders
- valid for all n
- explicit probability $P(\lambda)$ (but pessimistic)

Probabilistic backward error analysis

Probabilistic fundamental lemma (Higham and M., 2019, 2020)

Let δ_k, $k = 1 : n$, satisfy Model M. Then, for any $\lambda > 0$, the relation

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

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

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

Key features:

- can be applied **in a systematic way:** $\gamma_n \to \widetilde{\gamma}_n(\lambda)$

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

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

$$\widehat{L}\widehat{U} = A + \Delta A, \qquad |\Delta A| \le \widetilde{\gamma}_n(\lambda)|A|$$

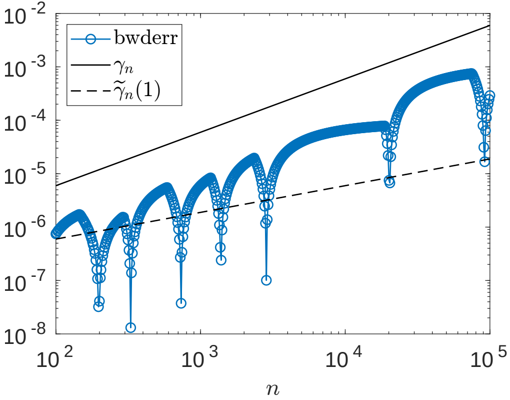
$$(A + \Delta A)\widehat{x} = b, \qquad |\Delta A| \le (3\widetilde{\gamma}_n(\lambda) + \widetilde{\gamma}_n(\lambda)^2)|A|$$

Example with dependent rounding errors

Summation with constant x_i:

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

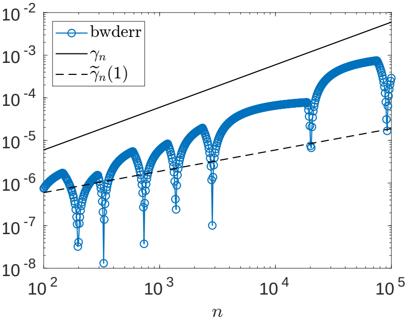
leads to an error growing as nu rather than $\sqrt{n}u$

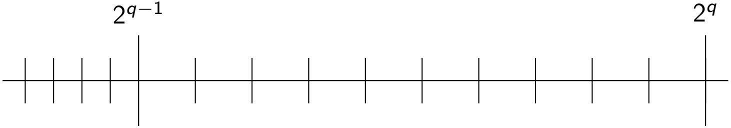


💬 Explain what is happening

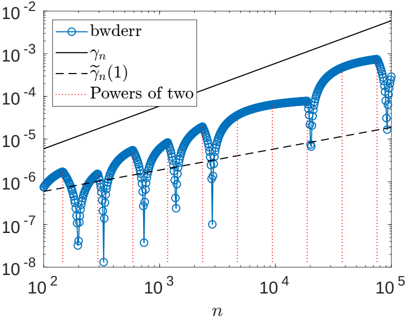
Example with dependent rounding errors (cont'd)

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

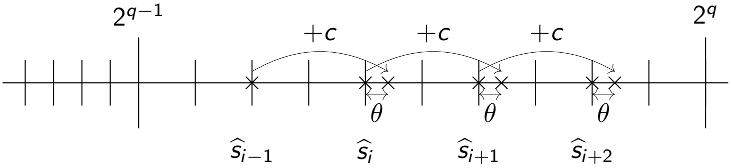




$$s_i = s_{i-1} + c \Rightarrow \quad \widehat{s}_i = (\widehat{s}_{i-1} + c)(1 + \delta_i)$$

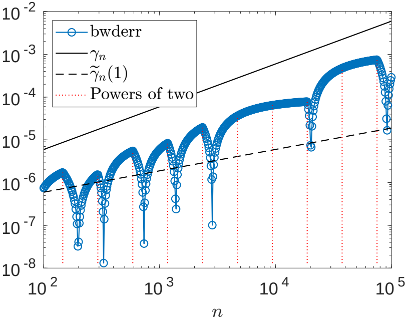


Example with dependent rounding errors (cont'd)



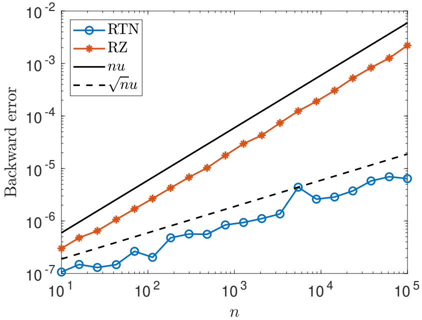
$$s_i = s_{i-1} + c \Rightarrow \quad \widehat{s}_i = (\widehat{s}_{i-1} + c)(1 + \delta_i)$$

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



Example with round to zero mode

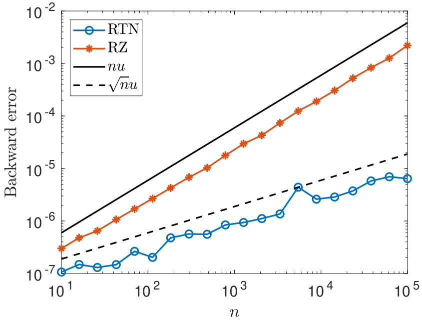
Error for round to nearest (RTN) and round to zero (RZ)



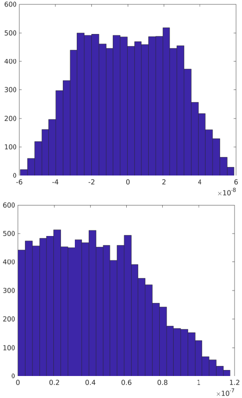
💬 Explain what is happening

Example with round to zero mode

Error for round to nearest (RTN) and round to zero (RZ)



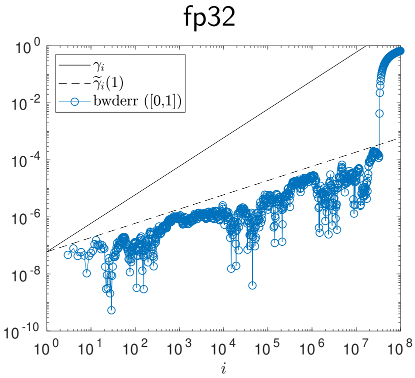
Distribution of δ_i: RTN (top) vs RZ (bottom)

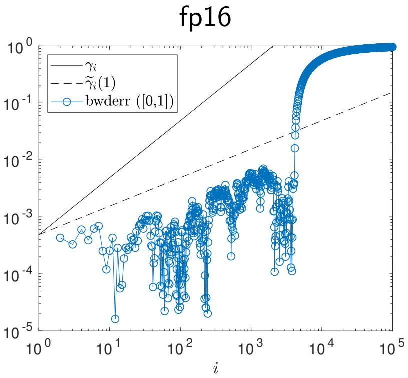


💬 Explain what is happening

Example with rounding errors of nonzero mean

Summation of a very large number of nonnegative terms ($n \gg 10^3$ in fp16, $n \gg 10^7$ in fp32) leads to an error eventually growing like $O(nu)$





Explain what is happening

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

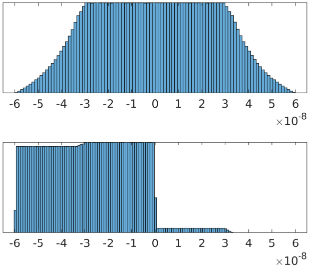
$$s_i = s_{i-1} + x_i \quad \Rightarrow \quad \widehat{s}_i = (\widehat{s}_{i-1} + x_i)(1 + \delta_i)$$

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

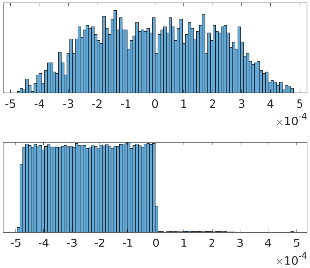
fp32



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

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

fp16



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

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

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

fp32 | fp16

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

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

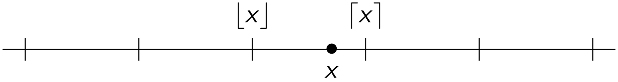
Validity of the probabilistic bound

- 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**

Stochastic rounding: definition



With round to nearest

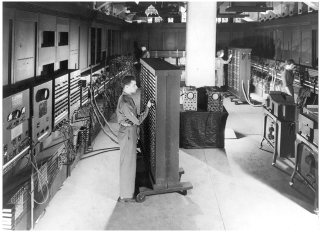
$$\mathrm{fl}(x) = \begin{cases} \lceil x \rceil & \text{if } x - \lfloor x \rfloor > \lceil x \rceil - x \\ \lfloor x \rfloor & \text{otherwise} \end{cases}$$

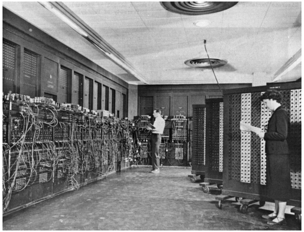
Instead, with stochastic rounding

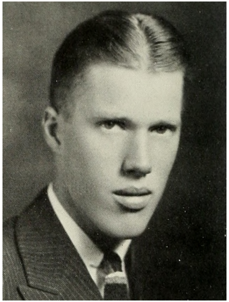
$$\mathrm{fl}(x) = \begin{cases} \lceil x \rceil & \text{with probability } p = \frac{x - \lfloor x \rfloor}{\lceil x \rceil - \lfloor x \rfloor} \\ \lfloor x \rfloor & \text{with probability } 1 - p = \frac{\lceil x \rceil - x}{\lceil x \rceil - \lfloor x \rfloor} \end{cases}$$

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

ENIAC (1947–1955)



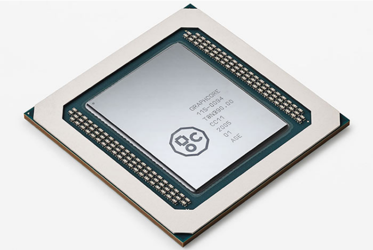




The individual rounding-off errors ϵ are really not random variables. In certain interval the ϵ's had a biased distribution which caused unexpectedly large accumulations of the rounding-off error. To circumvent this difficulty the present writer has proposed a random rounding-off procedure whichs make ϵ a true random variable.

George Forsythe, 1950

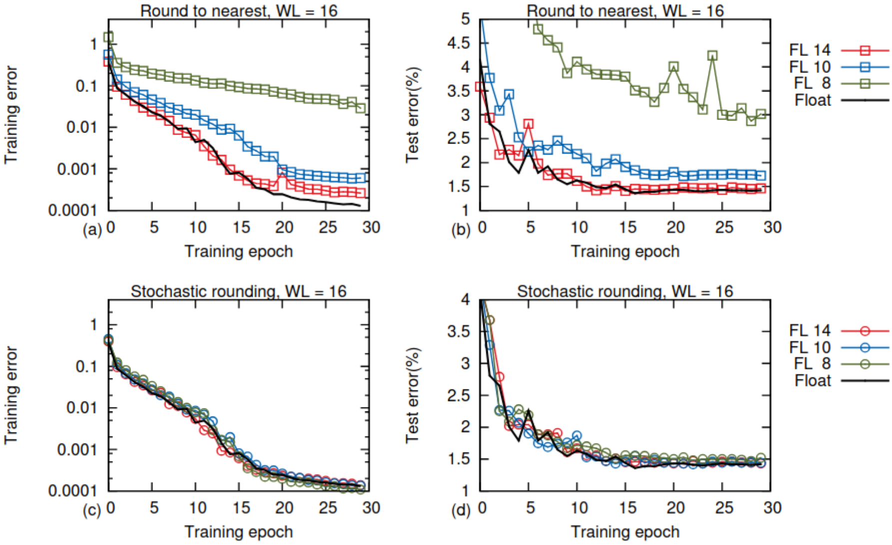
The last decade has seen a resurgence of interest in SR, with use in many applications and growing hardware support



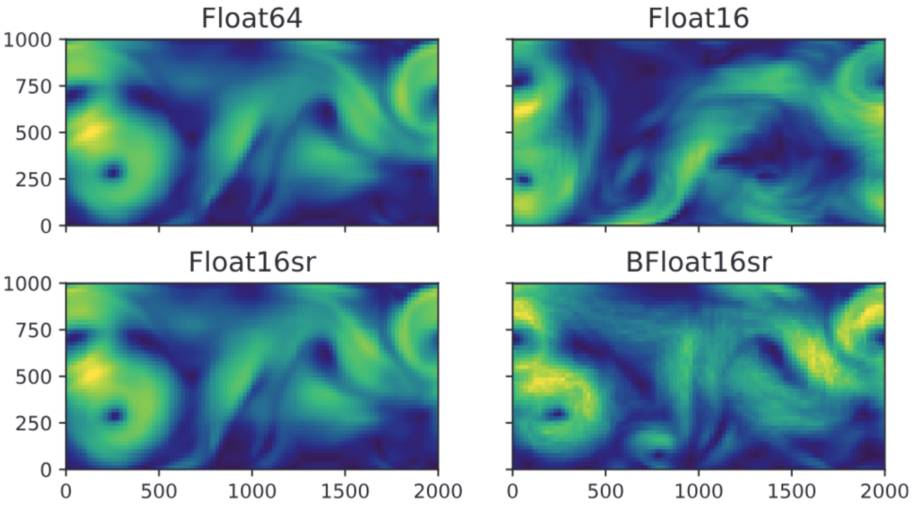
Graphcore IPU

Intel Loihi2

SR in deep learning (Gupta et al., 2015)



SR in climatology (Paxton et al., 2022)



How to explain the success of SR in modern low precision computing?

SR $\Rightarrow$ zero mean δ_i

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

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

SR $\Rightarrow$ zero mean δ_i

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

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

- Let $x := a \,\mathrm{op}\, b$; 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

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

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

SR $\not\Rightarrow$ 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, $\delta_2 = (\widehat{s} - \widehat{s}_1)/\widehat{s}_1$ is entirely determined by

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

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

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

SR $\Rightarrow$ 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 $\delta_{k+1} = (\widehat{s} - s)/s$ (which depends on $\delta_1, \dots, \delta_k$) is given by

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

- 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

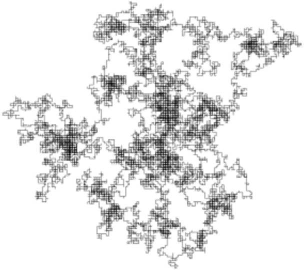
- Therefore we obtain

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

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

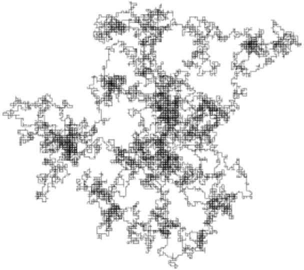
$$\mathbb{E}(\delta_i \mid \delta_1, \dots, \delta_{i-1}) = \mathbb{E}(\delta_i) \; (= 0)$$

SR is a 1D random walk!



The position at each step depends on previous positions, but we have an equal chance to take any direction at any given step $\Rightarrow$ martingale

SR is a 1D random walk!



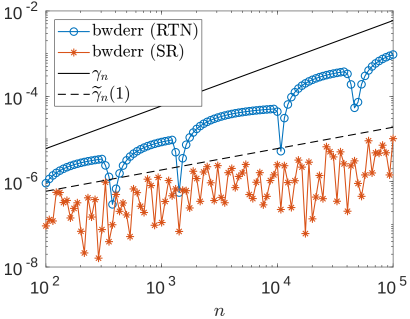
The position at each step depends on previous positions, but we have an equal chance to take any direction at any given step $\Rightarrow$ martingale

- SR transforms finite-precision computations into random walks
  - Rounding errors at a given step depend on previous errors, but this dependence is weak: their expectation remains zero by construction

Backward error bound with SR (Connolly, Higham and M., 2021)

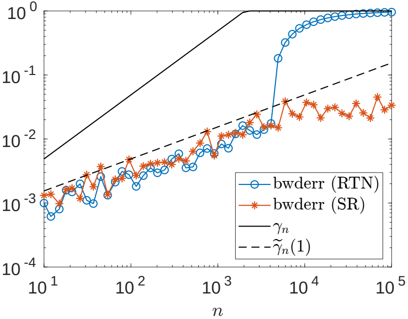
SR enforces Model M. Therefore, the $\widetilde{\gamma}_n$ bound holds unconditionally with SR.

Constant x_i (fp32)



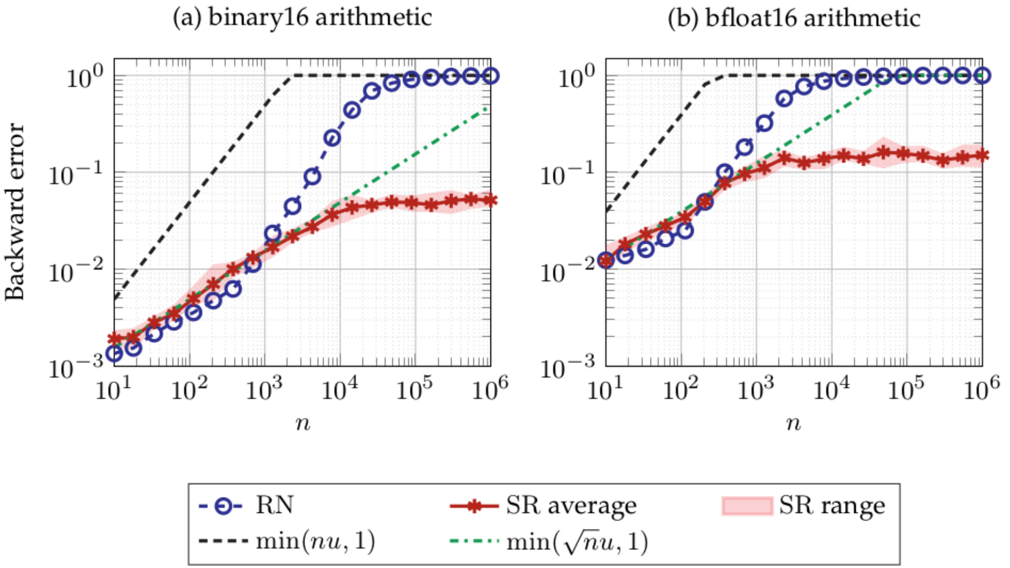
$\Rightarrow$ stochastic rounding produces nonconstant δ_i

Nonzero x_i (fp16)



$\Rightarrow$ stochastic rounding overcomes stagnation

Variance of SR



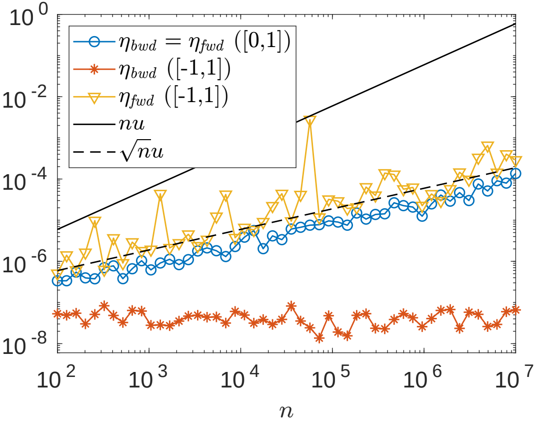
Back to random data

$\eta_{\text{fwd}} = \kappa \eta_{\text{bwd}}$

$\cancel{\eta_{\text{bwd}} \approx nu} \to \eta_{\text{bwd}} \approx \sqrt{n}u$:

- If $X_i \sim U([0,1])$: $\eta_{\text{fwd}} \approx \sqrt{n}u$?
  SHARP!
- If $X_i \sim U([-1,1])$: $\eta_{\text{fwd}} \approx nu$?
  Still not sharp!

$\Rightarrow$ Why do we have $\eta_{\text{bwd}} \approx u$ for $U([-1,1])$ data??



Intuitive explanation

- Recall that under Model M

$$\widehat{s} = \sum_{i=1}^{n} x_i y_i (1+\theta_i), \quad |\theta_i| \le \widetilde{\gamma}_n(\lambda)$$

  and thus

$$\eta_{\mathrm{bwd}} = \frac{|\widehat{s} - s|}{|x|^T |y|} = \frac{|\sum_{i=1}^n x_i y_i \theta_i|}{\sum_{i=1}^n |x_i y_i|}$$

- Under Model M, $|\widehat{s} - s| \le \lambda \sqrt{n} u \max_k |s_k|$, where s_k is the partial inner product of the first k elements of x and y
- Without any assumption on x_i, y_i, the best bound we have on $|\sum_{i=1}^n x_i y_i \theta_i|$ is $\widetilde{\gamma}_n(\lambda) \sum_{i=1}^n |x_i y_i|$. But what about for specific x_i, y_i?

$\Rightarrow$ If $\mathbb{E}(x_i y_i) = 0$, then we can expect the variables $z_i = x_i y_i \theta_i$ to also cancel each other!

Formal explanation

Model M'

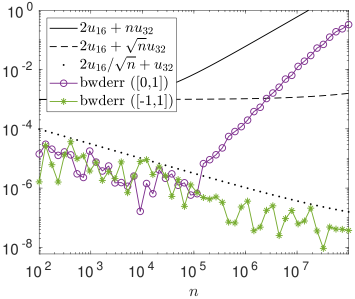
In addition to the assumptions of Model M, assume that in the inner product $s = x^T y$, x_i and y_i are random independent variables such that $\mathbb{E}(x_i y_i) = \mu$, $\mathbb{E}(|x_i y_i|) = \mu_+$, and $|x_i y_i| \le C$.

Probabilistic bwd error bound for random inner products (Higham and M., 2020)

Let $s = x^T y$. Under Model M', for any $\lambda > 0$, the backward error bound

$$\eta_{\mathrm{bwd}} = \frac{|\widehat{s} - s|}{|x|^T |y|} \le \frac{\lambda \mu \sqrt{n} + \lambda^2 C}{\mu_+ - \lambda C / \sqrt{n}} \cdot u + O(u^2)$$

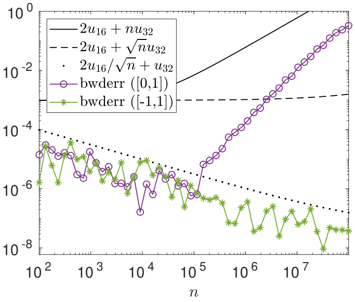
holds with probability $P(\lambda) = 1 - 2(n+1)\exp\left(-\lambda^2/2\right)$



Convert x and y to fp16, then compute $s = x^T y$ in fp32 arithmetic

$$\eta = \eta_{\text{convert}} + \eta_{\text{compute}}$$
$$\leq \frac{\left|\sum_{i=1}^{n} x_i y_i \epsilon_i\right|}{|x|^T |y|} + \eta_{\text{compute}}, \quad |\epsilon_i| \lesssim 2u_{16}$$

- Worst-case bound: $\eta \lesssim 2u_{16} + nu_{32}$
  $\Rightarrow$ starts growing for $n \geq 2u_{16}/u_{32} = 2^{14} \approx 10^4$



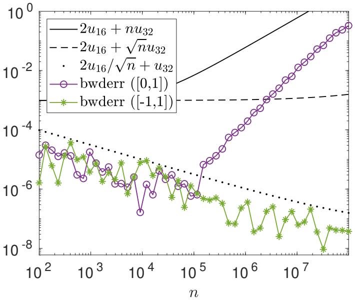
Convert x and y to fp16, then compute $s = x^T y$ in fp32 arithmetic

$$\eta = \eta_{\text{convert}} + \eta_{\text{compute}}$$

$$\leq \frac{\left|\sum_{i=1}^{n} x_i y_i \epsilon_i\right|}{|x|^T|y|} + \eta_{\text{compute}}, \qquad |\epsilon_i| \lesssim 2u_{16}$$

- Worst-case bound: $\eta \lesssim 2u_{16} + nu_{32}$
  $\Rightarrow$ starts growing for $n \geq 2u_{16}/u_{32} = 2^{14} \approx 10^4$
- Under Model M: $\eta \lesssim 2u_{16} + \sqrt{n}u_{32}$
  $\Rightarrow$ starts growing for $n \geq 2^{28} \approx 10^8$

Tensor cores



Convert x and y to fp16, then compute $s = x^T y$ in fp32 arithmetic

$$\eta = \eta_{\text{convert}} + \eta_{\text{compute}}$$

$$\leq \frac{\left|\sum_{i=1}^{n} x_i y_i \epsilon_i\right|}{|x|^T |y|} + \eta_{\text{compute}}, \quad |\epsilon_i| \lesssim 2u_{16}$$

- Worst-case bound: $\eta \lesssim 2u_{16} + nu_{32}$
  $\Rightarrow$ starts growing for $n \geq 2u_{16}/u_{32} = 2^{14} \approx 10^4$
- Under Model M: $\eta \lesssim 2u_{16} + \sqrt{n}u_{32}$
  $\Rightarrow$ starts growing for $n \geq 2^{28} \approx 10^8$
- Under Model M' for zero-mean vectors: $\eta \lesssim \frac{u_{16}}{\sqrt{n}} + cu_{32}$
  $\Rightarrow$ decreases until $n \gtrsim 10^8$

Shifting to zero mean for accuracy

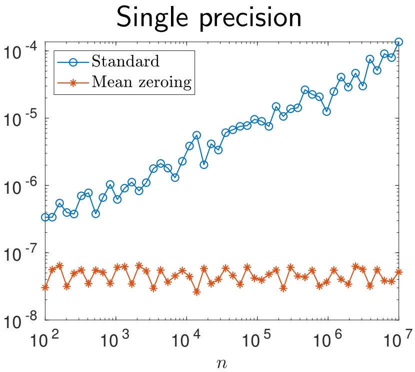
Idea: given x_i, y_i of mean $\mu \neq 0$, let $z_i = x_i - \mu$ and compute $s = z^T y + n\mu$, then $\eta \leq cu$ for some c independent of n

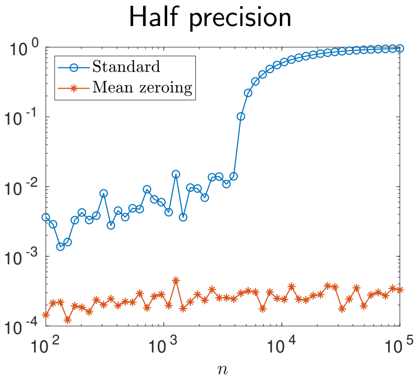
Cost: $2n$ flops but for $C = AB$, where $A, B, C \in \mathbb{R}^{n\times n}$ the cost of the algorithm below is in $O(n^2)$ instead of $O(n^3)$

$$\widetilde{A} \leftarrow A - xe^T$$
$$C \leftarrow \widetilde{A}B + x(e^T B)$$

where $x_i =$ mean of ith row of A and e is the vector full of ones

Backward error (for $[0,1]$ data)





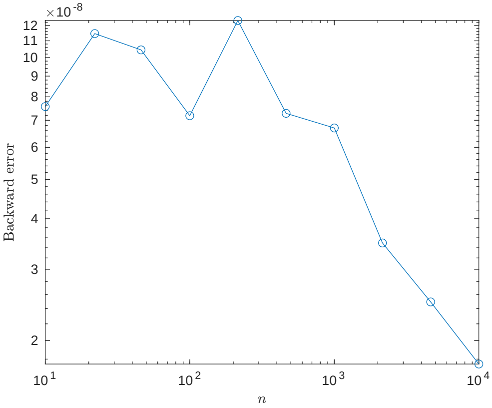
Open problem: matrix factorization and linear systems

Doolittle's formula for $A = LU$

$$\ell_{ik} = \left(a_{ik} - \sum_{j=1}^{k-1} \ell_{ij} u_{jk}\right) / u_{kk},$$

$$u_{kj} = a_{kj} - \sum_{i=1}^{k-1} \ell_{ki} u_{ij}$$

The inner products arising in LU factorization are not random! And yet...



Internship/PhD available (see here for details)