A New Approach to Probabilistic Rounding Error Analysis

Theo Mary, joint work with Nick Higham
University of Manchester, School of Mathematics

SIAM CSE19, Spokane, USA, Feb 25–Mar 1 2019

M\cr NA
Manchester Numerical Analysis

Context and motivation

Floating-point arithmetic model

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

	fp64 (double)	fp32 (single)	fp16 (half)	bfloat16 (half)	fp8 (quarter)
u	2^{-53} $\approx 10^{-16}$	2^{-24} $\approx 10^{-8}$	2^{-11} $\approx 10^{-4}$	2^{-8} $\approx 10^{-3}$	2^{-4} $\approx 10^{-2}$

- In many numerical linear algebra computations, traditional error bounds are proportional to nu, e.g., for LU factorization:

$$|A - LU| \le nu|L||U|$$

$\Rightarrow$ No guarantees if nu is large: issue of growing importance with the rise of large-scale, mixed-precision computations

- Yet, in practice errors are observed to be much smaller

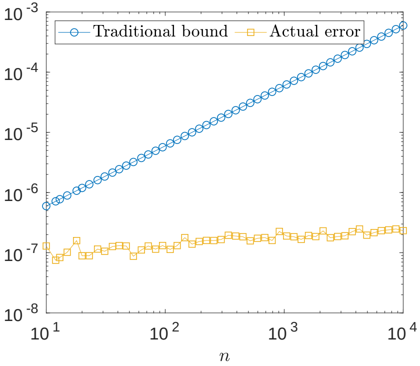
Traditional bounds are pessimistic

The issue is that traditional bounds are worst-case bounds, and are thus pessimistic on average

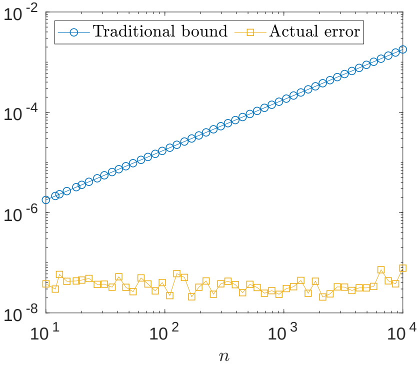
Traditional bounds are pessimistic

The issue is that traditional bounds are worst-case bounds, and are thus pessimistic on average

Matrix–vector product (fp32)

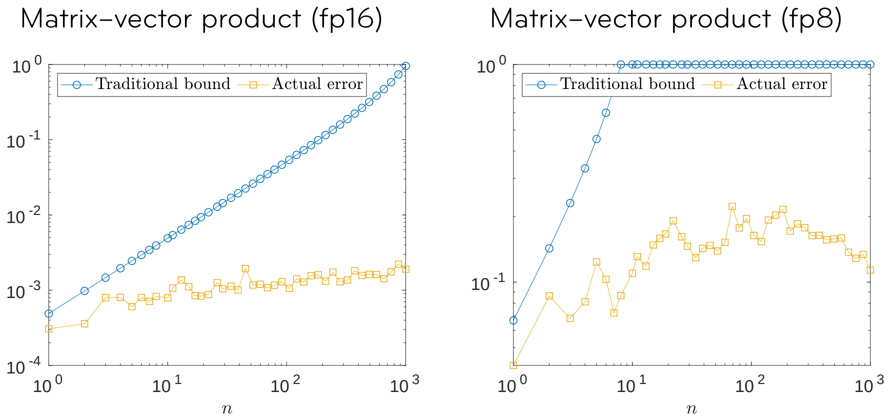


Solution of $Ax = b$ (fp32)



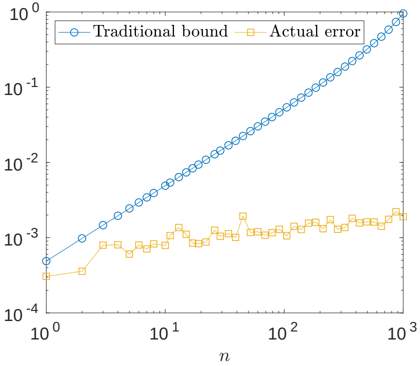
Traditional bounds are pessimistic

The issue is that traditional bounds are worst-case bounds, and are thus pessimistic on average

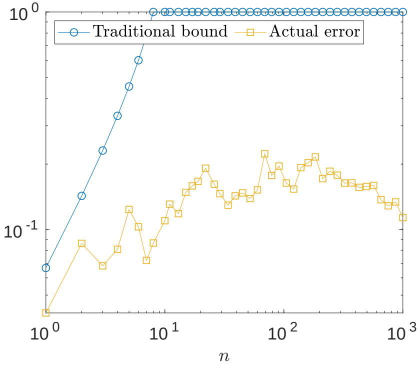


The issue is that traditional bounds are worst-case bounds, and are thus pessimistic on average

Matrix–vector product (fp16)



Matrix–vector product (fp8)



⇒ Traditional bounds do not provide a realistic picture of the typical behavior of numerical computations

Key intuition

- Consider the accumulated effect of n rounding errors

$$s = \sum_{i=1}^{n} \delta_i$$

- The worst-case bound $|s| \le nu$ is attained when all δ_i have identical sign and maximal magnitude, which intuitively seems very unlikely
- Assume δ_i are random independent variables of mean zero
- Then, the central limit theorem states that if n is sufficiently large, then

$$s/\sqrt{n} \sim \mathcal{N}(0, u)$$

$\Rightarrow$ $|s| \le \lambda\sqrt{n}u$, with λ a small constant, holds with high probability (e.g., 99.7% with $\lambda = 3$ by the 3-sigma rule)

The rule of thumb

This probabilistic approach had led to the following 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*

However, no rigorous result along these lines exists for a wide class of algorithms

The rule of thumb

This probabilistic approach had led to the following 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*

However, no rigorous result along these lines exists for a wide class of algorithms

Our contribution:
We provide the first rigorous foundation for this rule of thumb
by computing rigorous error bounds
that hold with probability at least a certain value
for a wide class of linear algebra algorithms

Objective and assumptions

Fundamental lemma in backward error analysis

If $|\delta_i| \le u$ for $i = 1 : n$ and $nu < 1$, then

$$\prod_{i=1}^{n}(1+\delta_i) = 1 + \theta_n, \quad |\theta_n| \le \gamma_n \le nu + O(u^2)$$

Objective and assumptions

Fundamental lemma in backward error analysis

If $|\delta_i| \leq u$ for $i = 1 : n$ and $nu < 1$, then

$$\prod_{i=1}^{n}(1+\delta_i) = 1 + \theta_n, \quad |\theta_n| \leq \gamma_n \leq nu + O(u^2)$$

We seek an anologous result by using the following model

Probabilistic model of rounding errors

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

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

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

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

Proof sketch

First step: transform the product in a sum by taking the logarithm

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

Proof sketch

First step: transform the product in a sum by taking the logarithm

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

Second step: apply Hoeffding's concentration inequality:

Hoeffding's inequality

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

$$\Pr(|S - \mathbb{E}(S)| \ge \xi) \le 2 \exp\left(-\frac{\xi^2}{2\sum_{i=1}^{n} c_i^2}\right)$$

to $X_i = \log(1+\delta_i)$ $\Rightarrow$ requires bounding $\log(1+\delta_i)$ and $\mathbb{E}\left(\log(1+\delta_i)\right)$ using Taylor expansions

Proof sketch

First step: transform the product in a sum by taking the logarithm

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

Second step: apply Hoeffding's concentration inequality:

Hoeffding's inequality

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

$$\Pr(|S - \mathbb{E}(S)| \ge \xi) \le 2\exp\left(-\frac{\xi^2}{2\sum_{i=1}^{n} c_i^2}\right)$$

to $X_i = \log(1+\delta_i)$ $\Rightarrow$ requires bounding $\log(1+\delta_i)$ and $\mathbb{E}\left(\log(1+\delta_i)\right)$ using Taylor expansions

Third step: retrieve the result by taking the exponential of S

Our 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

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

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

holds with probability of failure $P(\lambda) = 2\exp\left(-\lambda^2(1-u)^2/2\right)$

Our 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

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

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

holds with probability of failure $P(\lambda) = 2\exp\left(-\lambda^2(1-u)^2/2\right)$

Key features:

- Exact bound, not first order
- $nu < 1$ not required
- No "n is sufficiently large" assumption (achieved by replacing the central limit theorem by Hoeffding's inequality)
- Small values of λ suffice: $P(1) \approx 0.27$, $P(5) \le 10^{-5}$

Application to numerical linear algebra

Bounds for many numerical linear algebra algorithms are obtained by the repeated application of our main result. For example:

Probabilistic bound for LU factorization

Let $LU = A + \Delta A$ be the LU factors computed by Gaussian elimination of $A \in \mathbb{R}^{n\times n}$. Then, for any constant $\lambda > 0$, the relation

$$|\Delta A| \le \widetilde{\gamma}_n(\lambda)|L||U|, \quad |\widetilde{\gamma}_n(\lambda)| \le \lambda\sqrt{n}u + O(u^2)$$

holds with probability of failure $(n^3/3 + n^2/2 + 7n/6)P(\lambda)$

Application to numerical linear algebra

Bounds for many numerical linear algebra algorithms are obtained by the repeated application of our main result. For example:

Probabilistic bound for LU factorization

Let $LU = A + \Delta A$ be the LU factors computed by Gaussian elimination of $A \in \mathbb{R}^{n \times n}$. Then, for any constant $\lambda > 0$, the relation

$$|\Delta A| \le \widetilde{\gamma}_n(\lambda)|L||U|, \quad |\widetilde{\gamma}_n(\lambda)| \le \lambda\sqrt{n}u + O(u^2)$$

holds with probability of failure $(n^3/3 + n^2/2 + 7n/6)P(\lambda)$

We wish to keep the probabilities independent of n! Fortunately:

$$O(n^3)P(\lambda) = O(1) \quad \Rightarrow \quad \lambda = O(\sqrt{\log n})$$

$\Rightarrow$ error grows no faster than $\sqrt{n \log n}u$

Application to numerical linear algebra

Bounds for many numerical linear algebra algorithms are obtained by the repeated application of our main result. For example:

Probabilistic bound for LU factorization

Let $LU = A + \Delta A$ be the LU factors computed by Gaussian elimination of $A \in \mathbb{R}^{n\times n}$. Then, for any constant $\lambda > 0$, the relation

$$|\Delta A| \le \widetilde{\gamma}_n(\lambda)|L||U|, \quad |\widetilde{\gamma}_n(\lambda)| \le \lambda\sqrt{n}u + O(u^2)$$

holds with probability of failure $(n^3/3 + n^2/2 + 7n/6)P(\lambda)$

We wish to keep the probabilities independent of n! Fortunately:

$$O(n^3)P(\lambda) = O(1) \quad \Rightarrow \quad \lambda = O(\sqrt{\log n})$$

$\Rightarrow$ error grows no faster than $\sqrt{n \log n}u$

Moreover the constant hidden in the big O is small:

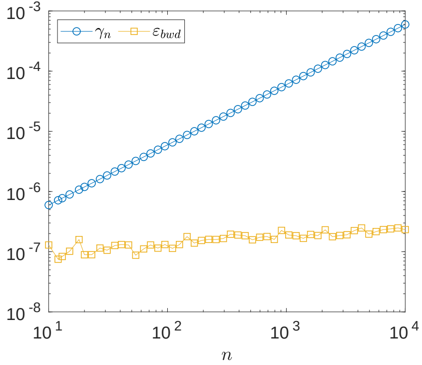
$$P(13) \le 10^{-5} \text{ for } n \le 10^{10}$$

Experimental setting

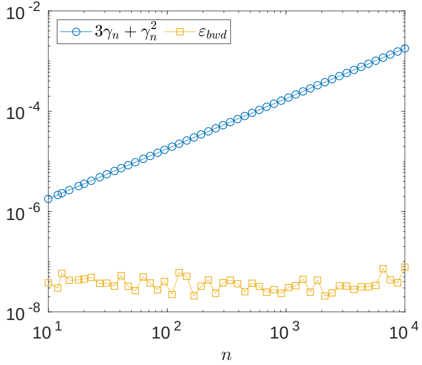
- We use MATLAB R2018b and set `rng(1)` for reproducibility
- fp16 and fp8 are simulated with the rounding function `chop.m` from the Matrix Computation Toolbox
- We use both random matrices with entries drawn from the uniform $[-1, 1]$ or $[0, 1]$ distribution and real-life matrices from the SuiteSparse collection
- We compare the bounds γ_n and $\widetilde{\gamma}_n(\lambda)$ with the componentwise backward error ε_{bwd} measured as (Oettli–Prager):
  - Matrix–vector product $y = Ax$: $\varepsilon_{bwd} = \max_i \frac{|\widehat{y}-y|_i}{(|A||x|)_i}$
  - Solution to $Ax = b$ via LU factorization: $\varepsilon_{bwd} = \max_i \frac{|A\widehat{x}-b|_i}{(|\widehat{L}||\widehat{U}||\widehat{x}|)_i}$
- Our codes are available online: https://gitlab.com/theo.andreas.mary/proberranalysis

Experimental results with $[-1, 1]$ entries

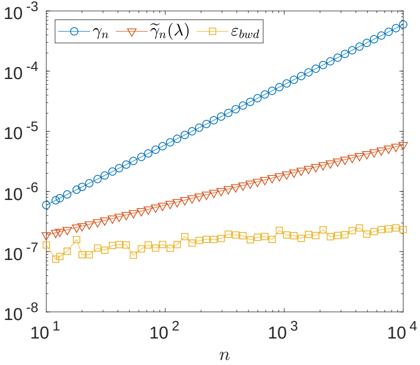
Matrix–vector product (fp32)



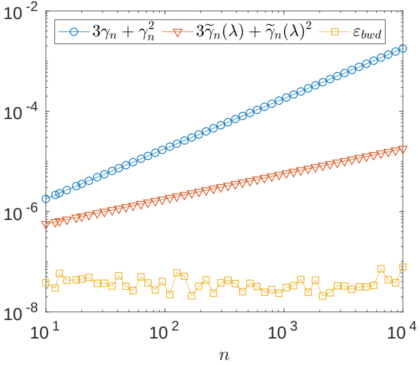
Solution of $Ax = b$ (fp32)



Matrix–vector product (fp32)

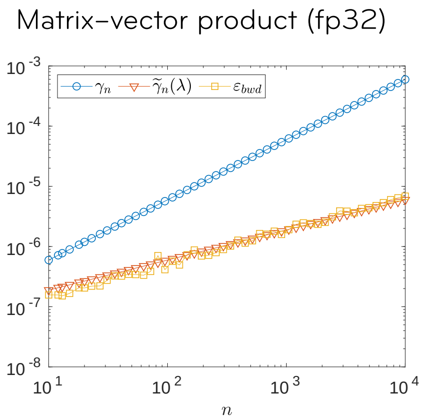


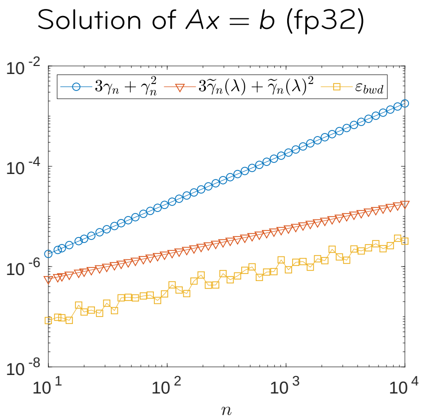
Solution of $Ax = b$ (fp32)



- The probabilistic bound is much closer to the actual error
- However for $[-1,1]$ entries it is still pessimistic

Experimental results with $[0,1]$ entries



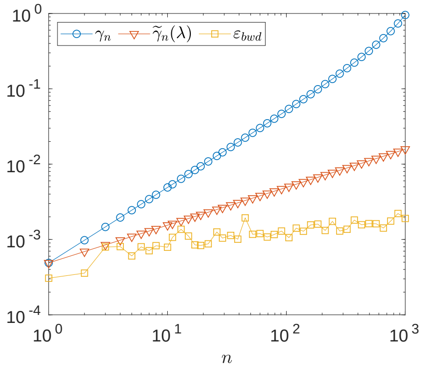


- Probabilistic bound is plotted with $\lambda = 1 \Rightarrow P(\lambda)$ is pessimistic...
- ...but $\widetilde{\gamma}_n$ bound itself can be sharp and successfully captures the $\sqrt{n}$ error growth

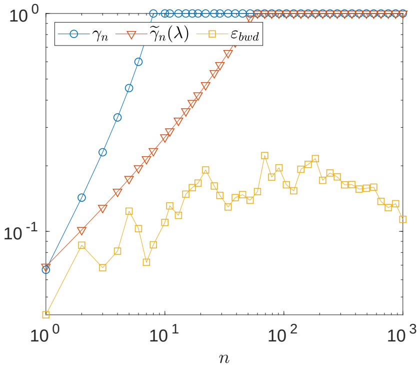
$\Rightarrow$ Therefore the bounds cannot be further improved without further assumptions

Experimental results with low precisions ($[-1, 1]$ entries)

Matrix–vector product (fp16)

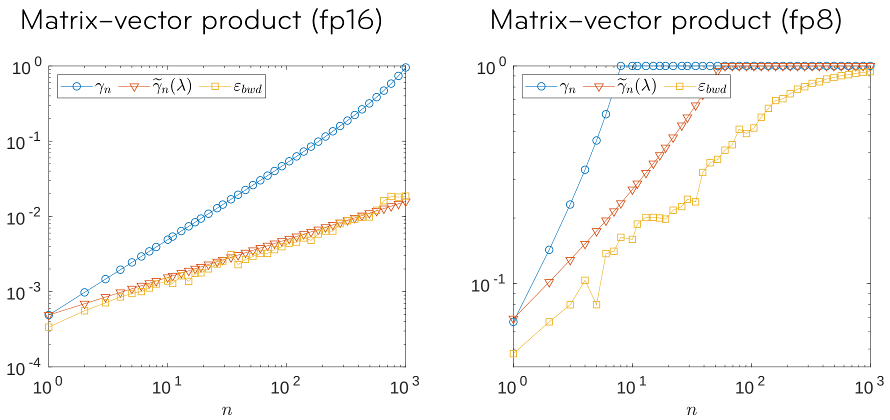


Matrix–vector product (fp8)



- Importance of the probabilistic bound becomes even clearer for lower precisions

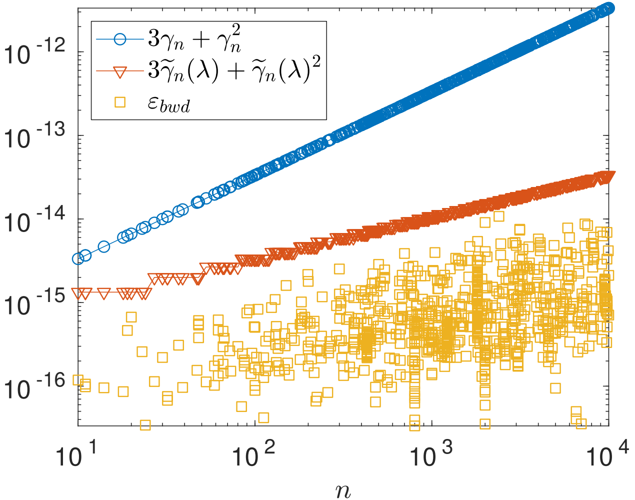
Experimental results with low precisions ($[0,1]$ entries)



- Importance of the probabilistic bound becomes even clearer for lower precisions

Experimental results with real-life matrices

Solution of $Ax = b$ (fp64),
for 943 matrices from the SuiteSparse collection



An example where rounding errors are not independent

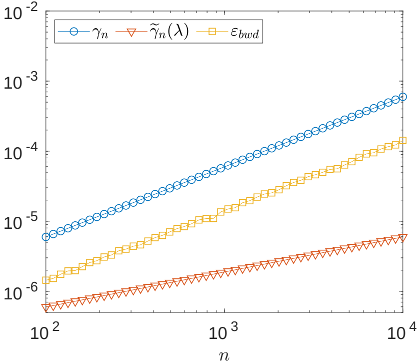
Inner product of two constant vectors:

$$
\begin{aligned}
& s_{i+1} = s_i + a_i b_i = s_i + c \\
\Rightarrow \quad & \widehat{s}_{i+1} = (\widehat{s}_i + c)(1 + \delta_i)
\end{aligned}
$$

Inner product of two constant vectors:

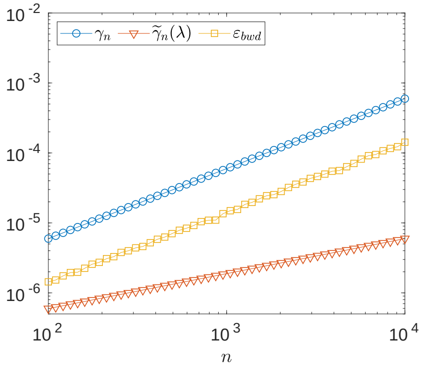
$$s_{i+1} = s_i + a_i b_i = s_i + c$$

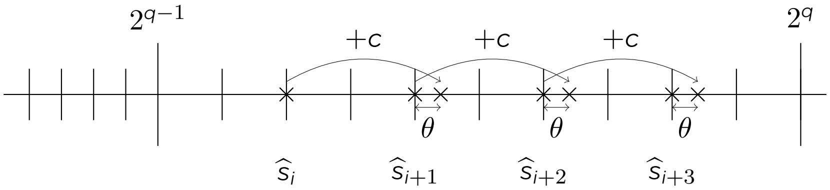
$$\Rightarrow \quad \widehat{s}_{i+1} = (\widehat{s}_i + c)(1 + \delta_i)$$



Inner product of two constant vectors:

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

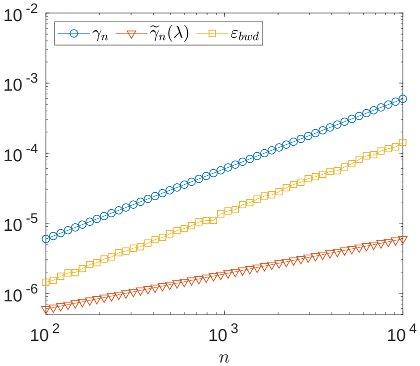


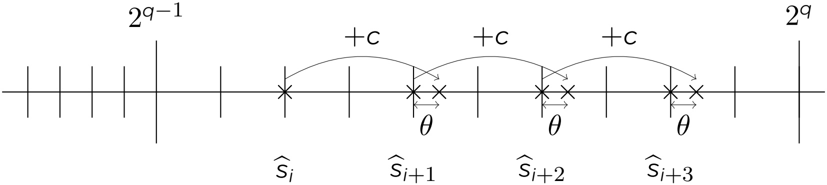


Inner product of two constant vectors:

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

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





An example where rounding errors have nonzero mean

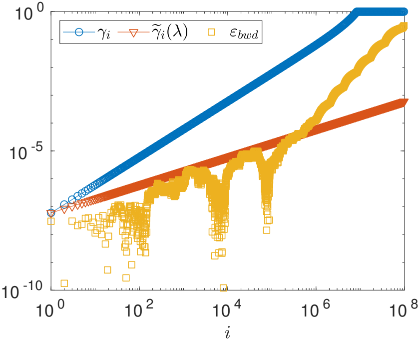
Inner product of two very large nonnegative vectors:

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

An example where rounding errors have nonzero mean

Inner product of two very large nonnegative vectors:

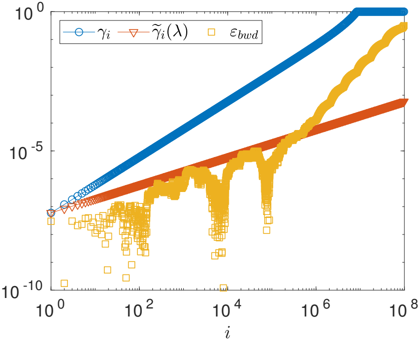
$$s_{i+1} = s_i + a_i b_i \quad \Rightarrow \quad \widehat{s}_{i+1} = (\widehat{s}_i + a_i b_i)(1+\delta_i)$$



An example where rounding errors have nonzero mean

Inner product of two very large nonnegative vectors:

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

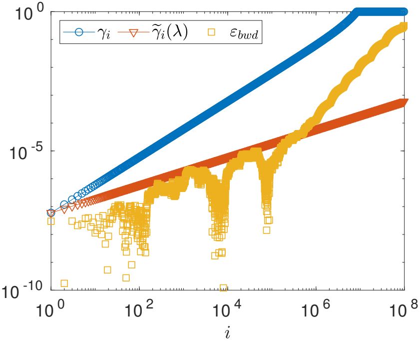


Explanation: s_i keeps increasing, at some point, it becomes so large that $\widehat{s}_{i+1} = \widehat{s}_i \Rightarrow \delta_i = -a_i b_i / (\widehat{s}_i + a_i b_i) < 0$

An example where rounding errors have nonzero mean

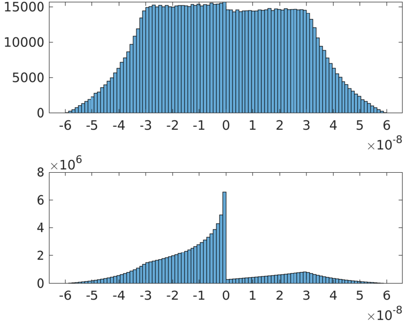
Inner product of two very large nonnegative vectors:

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



Top: $1 \le n \le 10^6$

Bottom: $10^6 \le n \le 10^8$



Explanation: s_i keeps increasing, at some point, it becomes so large that $\widehat{s}_{i+1} = \widehat{s}_i \Rightarrow \delta_i = -a_i b_i / (\widehat{s}_i + a_i b_i) < 0$

Conclusion

- Our analysis provides the first rigorous justification of the rule of thumb that one can take the square root of the constant in traditional error bounds to obtain a more realistic bound

- Our experiments show that the probabilistic bounds are in good agreement with the actual error for both random and real-life matrices, except in two very special situations, justifying that

  > *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*

  and answering Hull and Swenson's question

Slides and paper available here

bit.ly/theomary