MANCHESTER 1824

The University of Manchester

M\crNA
Manchester Numerical Analysis

Stochastic Rounding of Floating-Point Arithmetic

Michael Connolly
Department of Mathematics
The University of Manchester

michael.connolly-6@postgrad.manchester.ac.uk

NLA Group, 10 Oct 2019

Review of floating-point arithmetic I

For floating-point system F we have $y \in F \subset \mathbb{R}$ such that

$$y = \pm m \times \beta^{e-t}.$$

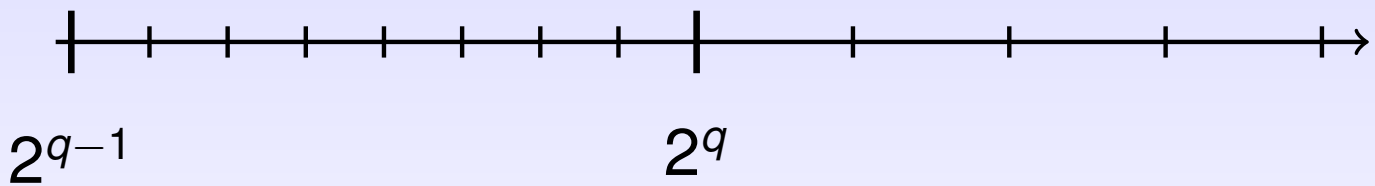
We have the integers:

- base β,
- precision t,
- exponent range $e_{\min} \leq e \leq e_{\max}$,
- significand m with $0 \leq m \leq \beta^t - 1$.

$fl(x \text{ op } y) = (x \text{ op } y)(1 + \delta)$, $|\delta| \leq u$, op $= +, -, *, /$.

Review of floating-point arithmetic II

- Spacing to right of 1 is $\epsilon_M = 2^{1-t}$ and unit roundoff $u = \epsilon_M/2 = 2^{-t}$.
- Spacing increases by factor of 2 after every power of 2.



Rounding modes

- Round to nearest (RTN)
- Round towards $\pm\infty$
- Round towards 0
- Stochastic rounding (SR)

Rounding modes

- Round to nearest (RTN)
- Round towards $\pm\infty$
- Round towards 0
- Stochastic rounding (SR)

Two possibilities for SR:

- Mode 1: Round up or down with equal probability
- Mode 2: Round with probability proportional to distance

All implemented in **`chop`** from Higham & Pranesh (2019).

Previous work

- Jézéquel & Chesneaux (2008) use SR in CADNA, a library for estimating propagation of rounding errors.
- Gupta et al (2015) show use of SR aids in the training of neural networks in low precision.
- Hopkins & Mikaitis et al (2019) improve accuracy of reduced precision fixed-point arithmetic with application to solving ODEs.

Stochastic rounding

We focus on Mode 2:

- $x_L < x < x_H$ with x_L, x_H adjacent floating-point numbers.
- $x_H - x_L = \epsilon$.

$$fl(x) = \begin{cases} x_L \text{ with probability } p = 1 - (x - x_L)/\epsilon, \\ x_H \text{ with probability } 1 - p = (x - x_L)/\epsilon. \end{cases}$$

Stochastic rounding

We focus on Mode 2:

- $x_L < x < x_H$ with x_L, x_H adjacent floating-point numbers.
- $x_H - x_L = \epsilon$.

$$fl(x) = \begin{cases} x_L \text{ with probability } p = 1 - (x - x_L)/\epsilon, \\ x_H \text{ with probability } 1 - p = (x - x_L)/\epsilon. \end{cases}$$

$$fl(x \text{ op } y) = (x \text{ op } y)(1 + \delta),\ |\delta| \leq 2u,\ \text{op} = +, -, *, /.$$

Some basic results

- $fl(x) \neq fl(x)$ under SR!

Some basic results

- $fl(x) \neq fl(x)$ under SR!
- $fl(x) = fl(x)$ if x a floating-point number.

Some basic results

- $fl(x) \neq fl(x)$ under SR!
- $fl(x) = fl(x)$ if x a floating-point number.
- With RTN, $fl(x * (1/x)) \in \{1 - \epsilon_M/2, 1\}$.

  With SR, $fl(x * (1/x)) \in \{1 - \epsilon_M, 1 - \epsilon_M/2, 1, 1 + \epsilon_M\}$.

Some basic results

- $fl(x) \neq fl(x)$ under SR!
- $fl(x) = fl(x)$ if x a floating-point number.
- With RTN, $fl(x * (1/x)) \in \{1 - \epsilon_M/2, 1\}$.

  With SR, $fl(x * (1/x)) \in \{1 - \epsilon_M, 1 - \epsilon_M/2, 1, 1 + \epsilon_M\}$.
- If x a floating-point number, with RTN, $fl(\sqrt{x^2}) = |x|$.

  With SR, $fl(\sqrt{x^2}) \in \{|x| - \epsilon_M, |x|, |x| + \epsilon_M\}$.

Expected values

Expected value of rounded result is true value:

$$\begin{aligned}
\mathbb{E}(fl(x)) &= \left(1 - \frac{x - x_L}{\epsilon}\right) x_L + \left(\frac{x - x_L}{\epsilon}\right) x_H, \\
&= x_L - \left(\frac{x - x_L}{\epsilon}\right) x_L + \left(\frac{x - x_L}{\epsilon}\right) x_H, \\
&= x_L + (x_H - x_L)\left(\frac{x - x_L}{\epsilon}\right), \\
&= x.
\end{aligned}$$

From $fl(x) = x(1 + \delta)$, we also have $\mathbb{E}(\delta) = 0$.

Probabilistic error analysis

Theorem (Higham & Mary, 2019)

Let δ_i, $i = 1 : n$, be independent random variables of mean zero such that $|\delta_i| \leq u$. For any $\lambda > 0$, the relation $\prod_{i=1}^{n}(1+\delta_i) = 1 + \theta_n$ holds with

$$|\theta_n| \leq \tilde{\gamma}_n(\lambda) := \exp\left(\lambda\sqrt{n}u + \frac{nu^2}{1-u}\right) - 1$$
$$\leq \lambda\sqrt{n}u + \mathcal{O}(u^2)$$

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

Probabilistic error analysis

Theorem (Higham & Mary, 2019)

Let δ_i, $i = 1 : n$, be independent random variables of mean zero such that $|\delta_i| \leq u$. For any $\lambda > 0$, the relation $\prod_{i=1}^{n}(1+\delta_i) = 1 + \theta_n$ holds with

$$|\theta_n| \leq \tilde{\gamma}_n(\lambda) := \exp\left(\lambda\sqrt{n}u + \frac{nu^2}{1-u}\right) - 1$$
$$\leq \lambda\sqrt{n}u + \mathcal{O}(u^2)$$

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

- Small λ are sufficient.

Assumptions of the theorem

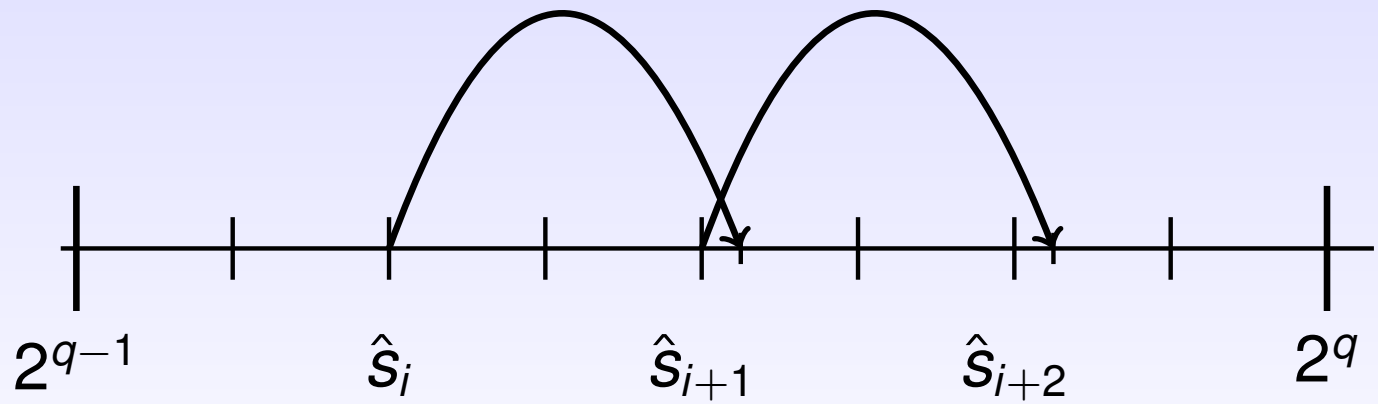
A key assumption was that the δ_i are random variables of mean zero.

- For SR we always have $\mathbb{E}(\delta_i) = 0$.
- Not always true of RTN.

Assumptions of the theorem

A key assumption was that the δ_i are random variables of mean zero.

- For SR we always have $\mathbb{E}(\delta_i) = 0$.
- Not always true of RTN.



Numerical experiments I

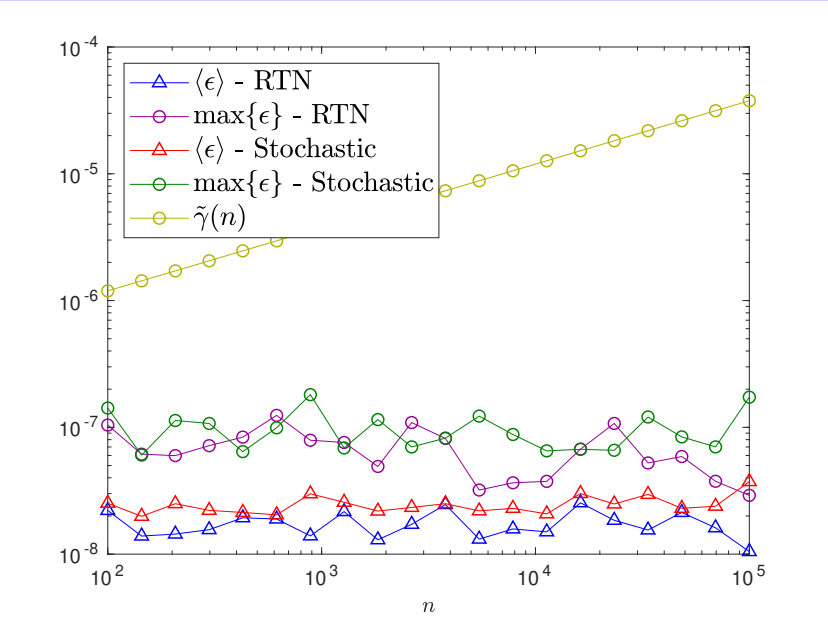
In the following:

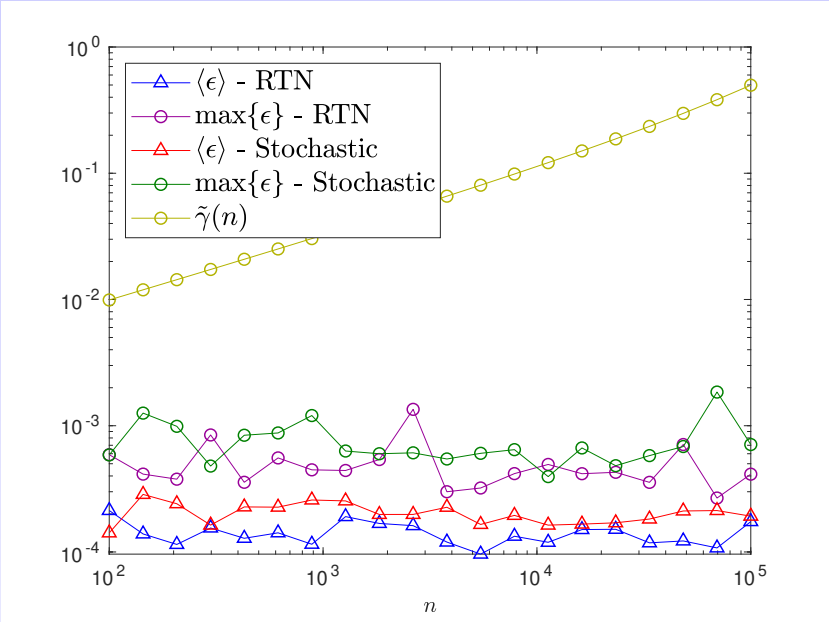
- Compute inner product $\hat{y} = a^T b$.
- Backward error given by

$$\epsilon_{\text{bwd}} = \frac{|\hat{y} - y|}{|a|^T |b|}.$$

- $u \to 2u$ in $\tilde{\gamma}(n)$.
- Set $\lambda = 1$ in all experiments.
- We expect $\epsilon_{\text{bwd}} < \tilde{\gamma}(n)$.

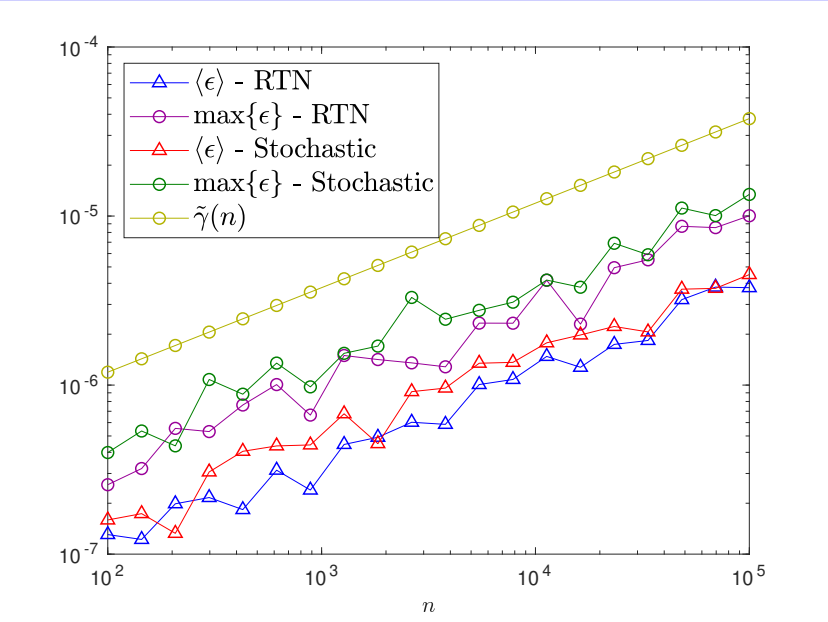
Numerical experiments II

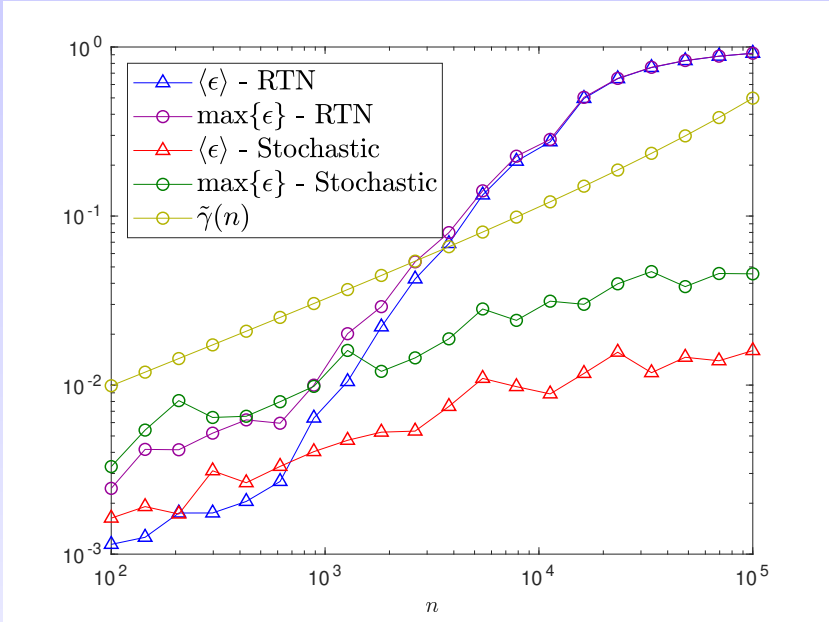




Data sampled from $U([-1, 1])$. Results shown for fp32 (left) and fp16 (right).

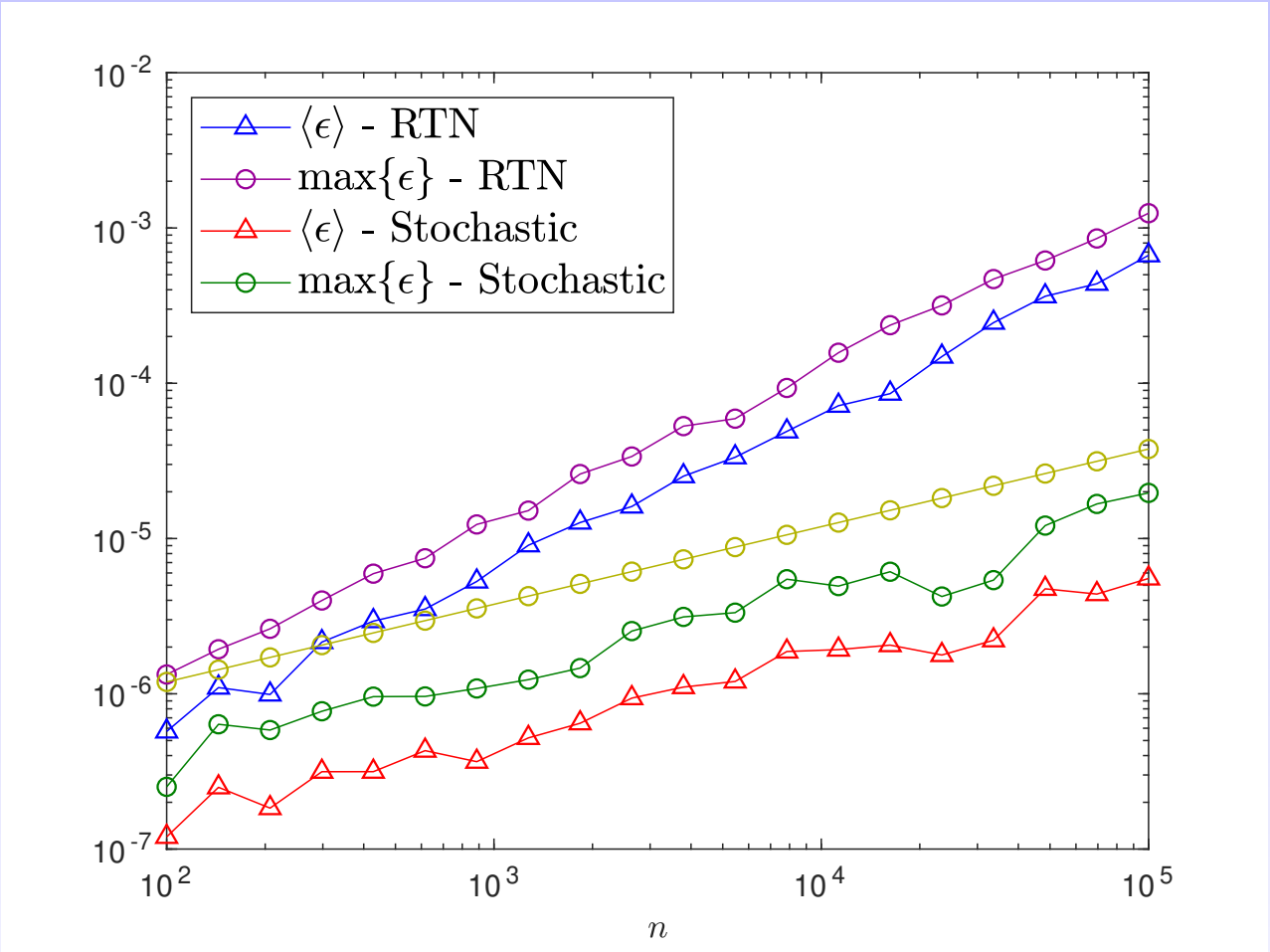
Numerical experiments III





Data sampled from $U([0,1])$. Results shown for fp32 (left) and fp16 (right).

Numerical experiments IV

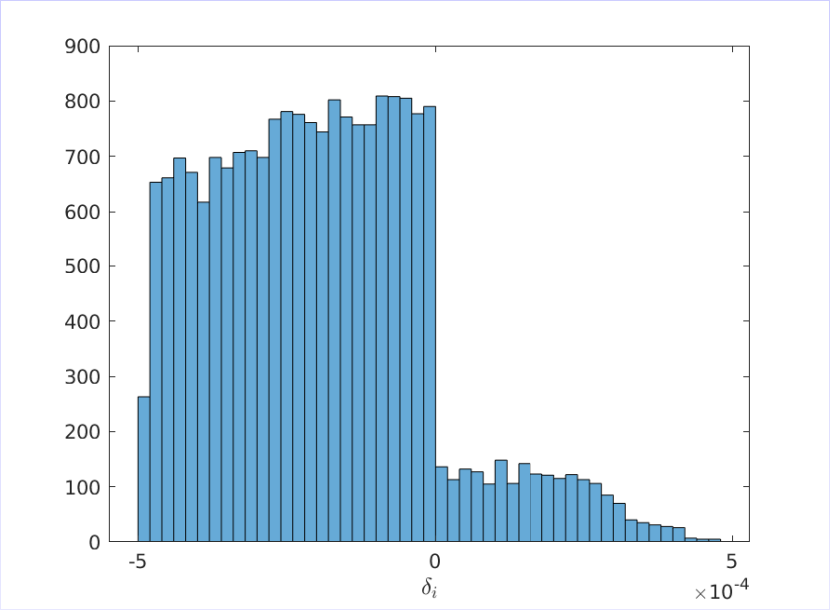


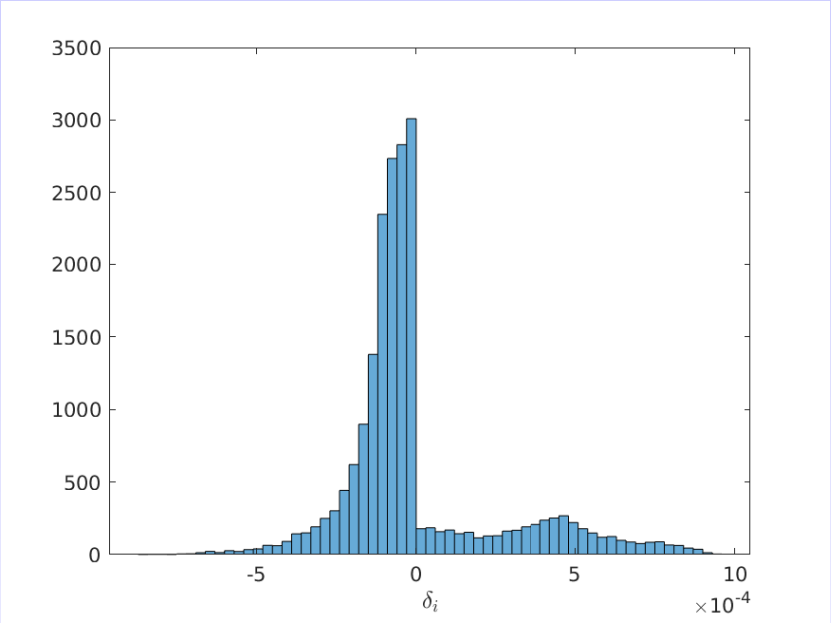
Computed backward errors of inner products for random constant vectors in fp32.

Stagnation

- RTN loses accuracy as our sum grows.
- Small summands and large spacing means our sum won't update.
- With SR, we make a "large" jump every so often to offset this.

Rounding error distributions





- Here we compute recursively in fp16 the sum of data sampled from $U([0,1])$ with $n = 2 \times 10^4$.
- For RTN (left) we have $\langle \delta_i \rangle = -1.93 \times 10^{-4}$.
- For SR (right) we have $\langle \delta_i \rangle = -7.13 \times 10^{-7}$.

Bernoulli random variables

- Define probability mass function (PMF) of a discrete random variable

$$f_X(x) = Pr(X = x),$$

the probability that random variable X takes the value x.

Bernoulli random variables

- Define probability mass function (PMF) of a discrete random variable

$$f_X(x) = Pr(X = x),$$

  the probability that random variable X takes the value x.
- Recall Bernoulli random variables with success/failure outcomes:

$$Pr(X = 1) = p = 1 - Pr(X = 0) = 1 - q.$$

- Binomial PMF given by

$$B(k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Poisson binomial PMF

- If we allow the probability to change after each trial we get

$$\xi_n(k) = \sum_{F \in \mathcal{F}_k} \prod_{i \in F} p_i \prod_{j \in F^c} (1 - p_j), \quad p_j = Pr(X_j = 1).$$

- $\mathcal{F}_k$ all subsets of k integers that can be selected from $\{1, \ldots, n\}$.
- Direct computation of $\xi_n(k)$ infeasible even for modest n.
- We can approximate

$$\xi_n(k) \approx \frac{\mu^k \exp(-\mu)}{k!}, \quad \mu = \sum p_i.$$

A toy problem I

Consider the following:

- Some large +ve value x_0.
- Some +ve summands x_i, $i = 1 : n$, which are small relative to the spacing in the vicinity of x_0, denoted ϵ.
- We then compute $\hat{s} = \sum_{i=0}^{n} x_i$.
- Assume spacing is constant throughout the computation.

Under RTN, $\hat{s} = x_0$.

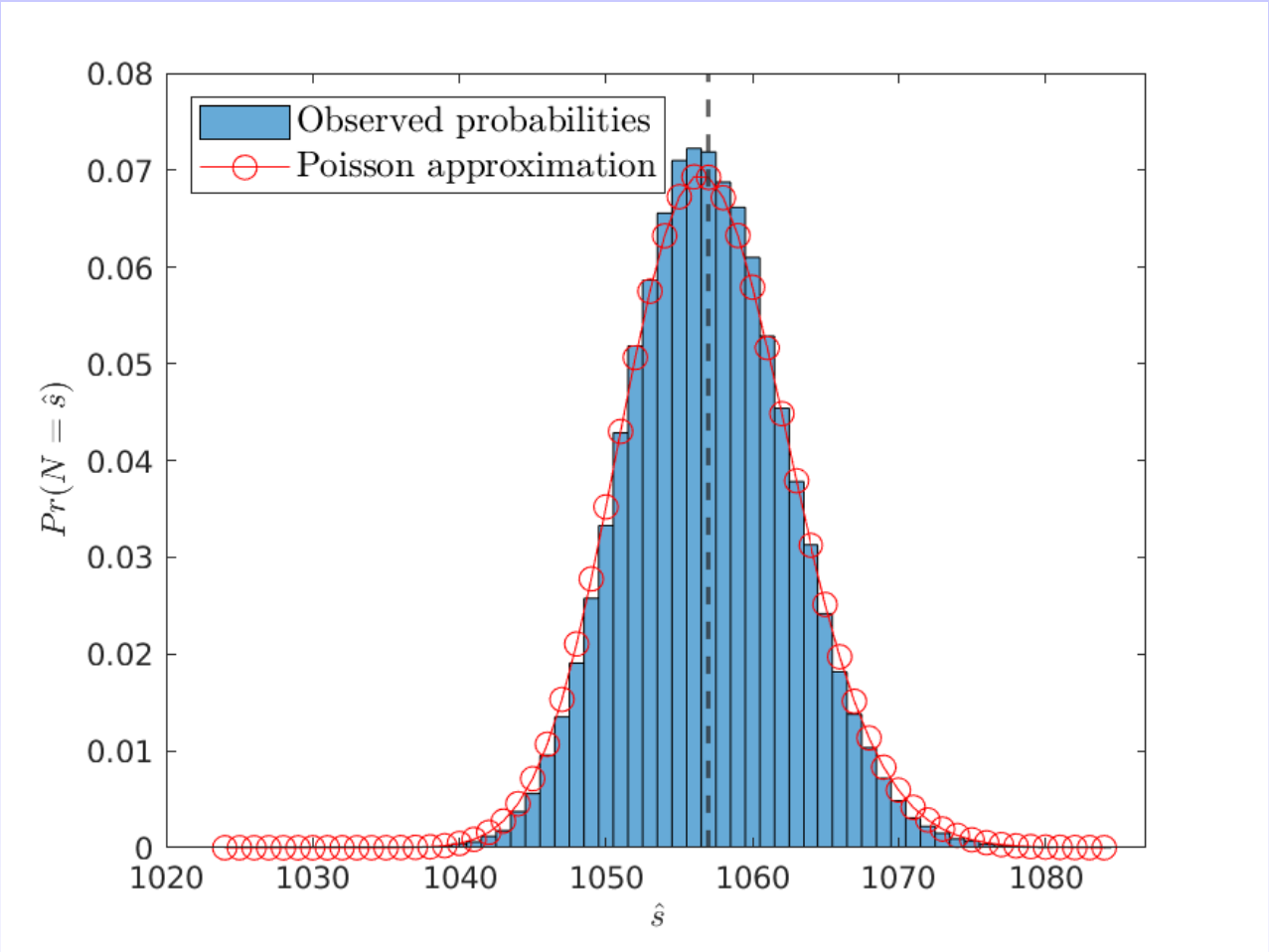
A toy problem II

- Under SR the probability we increment $\hat{s}_{i-1}$ by one floating-point number is x_i/ϵ. Let's call incrementing a "success".

- We can then approximate the PMF of $\hat{s}$ with

$$\xi_n(k) \approx \frac{\mu^k \exp(-\mu)}{k!}, \quad \mu = \sum x_i/\epsilon.$$

Results



Here we take $x_0 = 1024$, work in fp16 so $\epsilon = 1$ and generate the x_i from $U([0, 1/8])$ with $n = 512$.

Current and future work

Compute bound on variance of rounding errors. We could then:

- Use a different concentration inequality to produce a tighter $\tilde{\gamma}(n)$.
- By CLT show mean rounding error converges to a normal random variable. Can use this to approximate backward error and provide a bounding PDF of backward error.

Questions?

References I

N. J. Higham and S. Pranesh
Simulating low precision floating point arithmetic.
MIMS EPrint 2019.4, Manchester Institute for Mathematical Sciences, The University of Manchester, UK, 2019.

N. J. Higham and T. Mary
A new approach to probabilistic rounding error analysis.
SIAM J. Sci. Comput., 41(5):A2815-A2835, 2019.

References II

S. Gupta and A. Agrawal and K. Gopalakrishnan and P. Narayanan

Deep learning with limited numerical precision.

In Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37, ICML'15, pages 1737–1746. JMLR.org, 2015.

M. Hopkins and M. Mikaitis and D. R. Lester and S. Furber

Stochastic rounding and reduced-precision fixed-point arithmetic for solving neural ODEs.

arXiv:1904.11263, 2019.

References III

F. Jézéquel and J. Chesneaux
CADNA: a library for estimating round-off error propagation
Computer Physics Communications, 178(12):933 – 955, 2008.