# Bounds on Non-Linear Errors for Variance Computation with Stochastic Rounding

El-Mehdi El ARAR
*el-mehdi.el-arar@uvsq.fr*

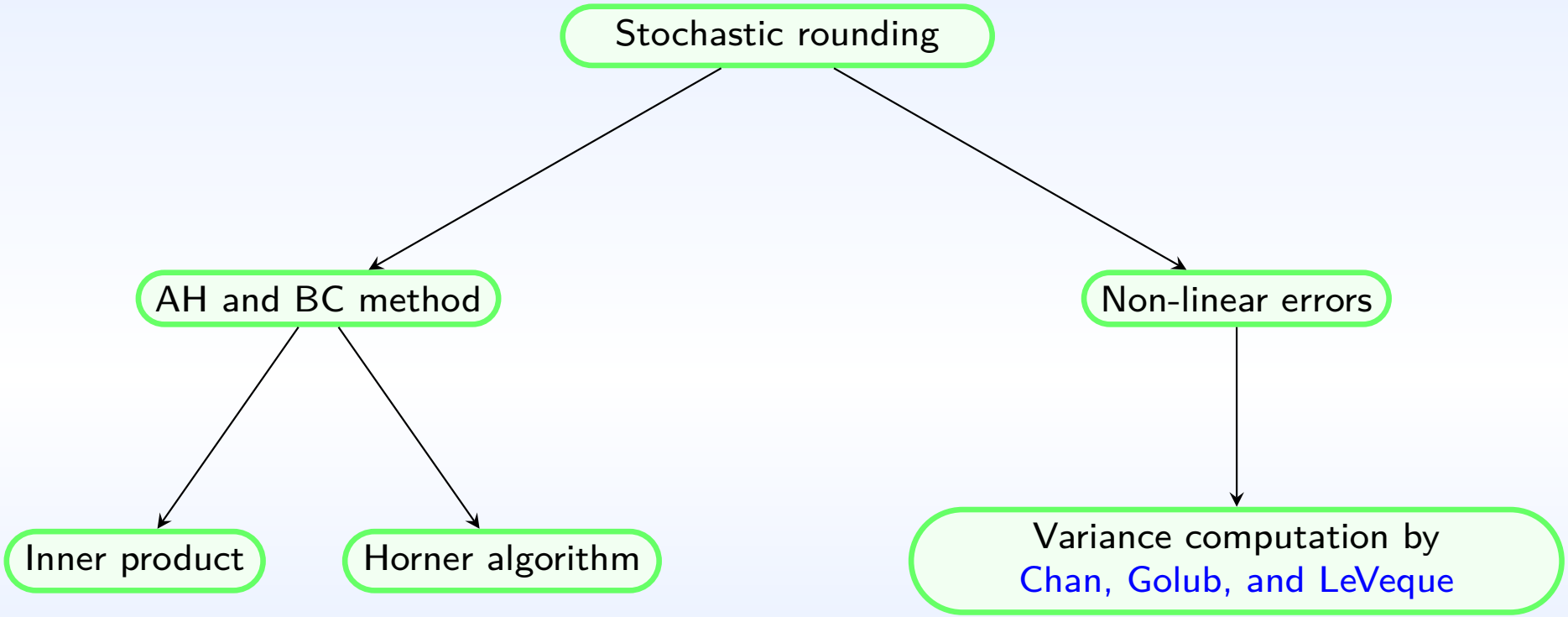
Devan SOHIER, Pablo DE OLIVEIRA CASTRO and Eric PETIT




ICIAM 2023

**Preprint submitted for publication (Available in Arxiv:2304.05177 )**

Bounds on Non-Linear Errors for Variance Computation with Stochastic Rounding.



# Plan

1. Recap of some stochastic rounding properties.
2. Textbook and Two-pass algorithms
   - Bias analysis
   - Probabilistic error analysis

Let us denote $\mathcal{F} \subset \mathbb{R}$ the set of floating-point numbers and $\hat{x} = \mathrm{fl}(x)$.

- For $x, y \in \mathcal{F}$ and $\mathrm{op} \in \{+, -, *, /\}$

$$\widehat{(x \,\mathrm{op}\, y)} = (x \,\mathrm{op}\, y)(1+\delta),\ |\delta| \leqslant u.$$

- IEEE-754 mode RN (round to nearest, ties to even) has the stronger property that $|\delta| \leqslant \frac{1}{2}\beta^{1-p} = \frac{1}{2}u$.
- $\varepsilon(x) = \beta^{e-p} = \lceil x \rceil - \lfloor x \rfloor$ and $p(x) = \frac{x - \lfloor x \rfloor}{\lceil x \rceil - \lfloor x \rfloor}$.

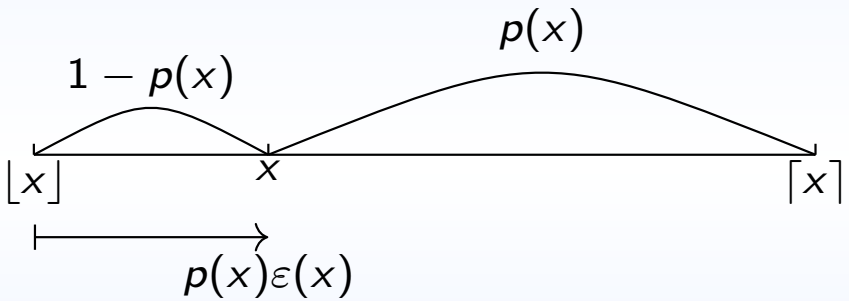



Figure: **SR-nearness**.

- $E(\hat{x}) = p(x)\lceil x \rceil + (1 - p(x))\lfloor x \rfloor = x$.

# Mean independence

- For $x_1, x_2, x_3 \in \mathbb{R}$, such that $c = x_1 \,\mathrm{op}\, x_2 \,\mathrm{op}\, x_3$, and

$$\widehat{c} = ((x_1 \,\mathrm{op}\, x_2)(1+\delta_1) \,\mathrm{op}\, x_3)(1+\delta_2),$$

  obtained from SR-nearness. $\delta_1, \delta_2$ are random variables such that $\mathbb{E}(\delta_1) = \mathbb{E}(\delta_2) = 0$.
- Mean independence: $X_1, X_2, ...$ are mean independent if $\mathbb{E}[X_k/X_1, ..., X_{k-1}] = \mathbb{E}(X_k)$ for all $k$.
- $X$ and $Y$ are independents $\implies$ $X$ is mean independent from $Y$ $\implies$ $X$ and $Y$ are uncorrelated.

**Lemma 1 (M. P. CONNOLLY, N. J. HIGHAM, AND T. MARY).**

*For some $\delta_1, \delta_2, ...$, in that order obtained from SR-nearness, the $\delta_k$ are random variables with mean zero such that $\mathbb{E}[\delta_k/\delta_1, ..., \delta_{k-1}] = \mathbb{E}(\delta_k) = 0$.*

**Definition 1 (Martingale).**

*A sequence of random variables $M_1, ..., M_n$ is a martingale with respect to the sequence $X_1, ..., X_n$ if, for all $k$,*

- *$M_k$ is a function of $X_1, ..., X_k$,*
- *$\mathbb{E}(|M_k|) < \infty$, and*
- *$\mathbb{E}[M_k / X_1, ..., X_{k-1}] = M_{k-1}$.*

**Definition 2 (Azuma-Hoeffding inequality).**

*Let $M_0, ..., M_n$ be a martingale with respect to a sequence $X_1, ..., X_n$. We assume that $-b_k \leqslant M_k - M_{k-1} \leqslant b_k$ for $k = 1 : n$*

$$\mathbb{P}\left( |M_n - M_0| \geqslant \sqrt{\sum_{k=1}^{n} b_k^2} \sqrt{2 \ln(2/\lambda)} \right) \leqslant \lambda,$$

*where $0 < \lambda < 1$.*

# Textbook and two-pass algorithms bias

For $x \in \mathbb{R}^n$, let $s = \sum_{i=1}^{n} x_i$ and $m = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}s$.

## Textbook algorithm

The textbook algorithm computes the variance using the formula $y = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}s^2$. Under SR-nearness, we have

- $\widehat{s} = \sum_{i=1}^{n} x_i \prod_{k=\max(2,i)}^{n} (1+\delta_{k-1})$.
- $\widehat{y} = \sum_{i=1}^{n} x_i^2 \psi_i - \frac{1}{n}\widehat{s}^2 \psi_{n+1}$.
- $E(\widehat{y}) = y - \frac{1}{n} V(\widehat{s})$.

# Textbook and two-pass algorithms bias

For $x \in \mathbb{R}^n$, let $s = \sum_{i=1}^n x_i$ and $m = \frac{1}{n}\sum_{i=1}^n x_i = \frac{1}{n}s$.

## Textbook algorithm

The textbook algorithm computes the variance using the formula $y = \sum_{i=1}^n x_i^2 - \frac{1}{n}s^2$. Under SR-nearness, we have

- $\widehat{s} = \sum_{i=1}^n x_i \prod_{k=\max(2,i)}^n (1+\delta_{k-1})$.
- $\widehat{y} = \sum_{i=1}^n x_i^2 \psi_i - \frac{1}{n}\widehat{s}^2 \psi_{n+1}$.
- $E(\widehat{y}) = y - \frac{1}{n}V(\widehat{s})$.

## Two-pass algorithm

The two-pass algorithm computes the variance using the formula $z = \sum_{i=1}^n (x_i - m)^2$. Under SR-nearness, we have

- $\widehat{m} = \frac{1}{n}\sum_{i=1}^n x_i \prod_{k=\max(2,i)}^{n+1} (1+\delta_{k-1})$.
- $\widehat{z} = \sum_{i=1}^n (x_i - \widehat{m})^2 (1+\varepsilon_i)^2 (1+\eta_i) \prod_{k=\max(2,i)}^n (1+\theta_k)$.
- $E(\widehat{z}) = z + \frac{1}{n}V(\widehat{s}) + O(u^2)$.

# Error analysis for algorithms with non-linear error

For the textbook algorithm, we have

$$\left|\widehat{y} - y\right| = \left|\sum_{i=1}^{n} x_i^2(\psi_i - 1) - \frac{1}{n}(\widehat{s^2}\psi_{n+1} - s^2)\right|$$

$$\leqslant \left|\sum_{i=1}^{n} x_i^2(\psi_i - 1)\right| + \frac{1}{n}\left|\widehat{s^2}\psi_{n+1} - s^2\right|.$$

# Error analysis for algorithms with non-linear error

For the textbook algorithm, we have

$$\left|\widehat{y} - y\right| = \left|\sum_{i=1}^{n} x_i^2(\psi_i - 1) - \frac{1}{n}(\widehat{s}^2\psi_{n+1} - s^2)\right|$$

$$\leqslant \left|\sum_{i=1}^{n} x_i^2(\psi_i - 1)\right| + \frac{1}{n}\left|\widehat{s}^2\psi_{n+1} - s^2\right|.$$

For example, we know that $\widehat{s} - s$ constructs a martingale:

- $\frac{1}{n}\left|\left((\widehat{s} - s) + s\right)^2\psi_{n+1} - s^2\right| \leqslant \frac{1}{n}\left(\left|(\widehat{s} - s)^2\psi_{n+1}\right| + 2\left|s(\widehat{s} - s)\psi_{n+1}\right| + \left|s^2(\psi_{n+1} - 1)\right|\right)$.
- $\frac{1}{n}\left|\widehat{s}^2\psi_{n+1} - \widehat{s}s + \widehat{s}s - s^2\right| = \frac{1}{n}\left|\widehat{s}(\widehat{s}\psi_{n+1} - s) + s(\widehat{s} - s)\right|$.
- $Z_n = \widehat{s} - s$ and $(Z_n + s)^2 = X_n + s^2 + A_n$. Therefore

$$\frac{1}{n}\left|(Z_n + s)^2\psi_{n+1} - s^2\right| = \frac{1}{n}\left|\psi_{n+1}(X_n + A_n) - s^2(\psi_{n+1} - 1)\right|.$$

- The key idea:

$$\widehat{y} - y = M + A.$$

| | $nu \ll 1$ | $nu \gg 1$ and $nu^2 \ll 1$ |
|---|---|---|
| Det | $(\mathcal{K}_2^2 + 2\mathcal{K}_1^2)nu$ | $(\mathcal{K}_2^2 + \mathcal{K}_1^2)e^{(2n+1)u}$ |
| BC | $(\mathcal{K}_2^2 + 2\mathcal{K}_1^2)\sqrt{2/\lambda}\sqrt{nu}$ | $(\mathcal{K}_2^2 + 2\mathcal{K}_1^2)\sqrt{2/\lambda}\sqrt{nu}$ |
| AH | $(\mathcal{K}_2^2 + 2\mathcal{K}_1^2)\sqrt{\ln(4/\lambda)}\sqrt{nu}$ | $(\mathcal{K}_2^2 + \mathcal{K}_1^2\sqrt{u\ln(4/\lambda)})\sqrt{u\ln(4/\lambda)}e^{(2n+1)u}$ |
| DM | $(\mathcal{K}_2^2 + \sqrt{8}\mathcal{K}_1^2)\sqrt{\ln(4/\lambda)}\sqrt{nu}$ | $\left(\sqrt{u\ln(4/\lambda)}(\mathcal{K}_2^2 + \sqrt{2}\mathcal{K}_1^2) + \mathcal{K}_1^2\frac{u}{2}\right)e^{(2n+1)u}$ |

Table: The asymptotic behavior of the textbook forward error bounds for a fixed probability $\lambda$ and over $n$ up to a constant. $\mathcal{K}_1 = \frac{\|x\|_1}{\sqrt{ny}}$ and $\mathcal{K}_2 = \frac{\|x\|_2}{\sqrt{y}}$.

| | $nu \ll 1$ | $nu \gg 1$ and $nu^2 \ll 1$ |
|---|---|---|
| Det | $(\mathcal{K}_2^2 + 2\mathcal{K}_1^2)nu$ | $(\mathcal{K}_2^2 + \mathcal{K}_1^2)e^{(2n+1)u}$ |
| BC | $(\mathcal{K}_2^2 + 2\mathcal{K}_1^2)\sqrt{2/\lambda}\sqrt{nu}$ | $(\mathcal{K}_2^2 + 2\mathcal{K}_1^2)\sqrt{2/\lambda}\sqrt{nu}$ |
| AH | $(\mathcal{K}_2^2 + 2\mathcal{K}_1^2)\sqrt{\ln(4/\lambda)}\sqrt{nu}$ | $(\mathcal{K}_2^2 + \mathcal{K}_1^2\sqrt{u\ln(4/\lambda)})\sqrt{u\ln(4/\lambda)}e^{(2n+1)u}$ |
| DM | $(\mathcal{K}_2^2 + \sqrt{8}\mathcal{K}_1^2)\sqrt{\ln(4/\lambda)}\sqrt{nu}$ | $\left(\sqrt{u\ln(4/\lambda)}(\mathcal{K}_2^2 + \sqrt{2}\mathcal{K}_1^2) + \mathcal{K}_1^2\frac{u}{2}\right)e^{(2n+1)u}$ |

Table: The asymptotic behavior of the textbook forward error bounds for a fixed probability $\lambda$ and over $n$ up to a constant. $\mathcal{K}_1 = \frac{\|x\|_1}{\sqrt{ny}}$ and $\mathcal{K}_2 = \frac{\|x\|_2}{\sqrt{y}}$.

- Extension of the previous results to the pairwise case. $\log_2(n)$ instead of $n$.

# Numerical experiments

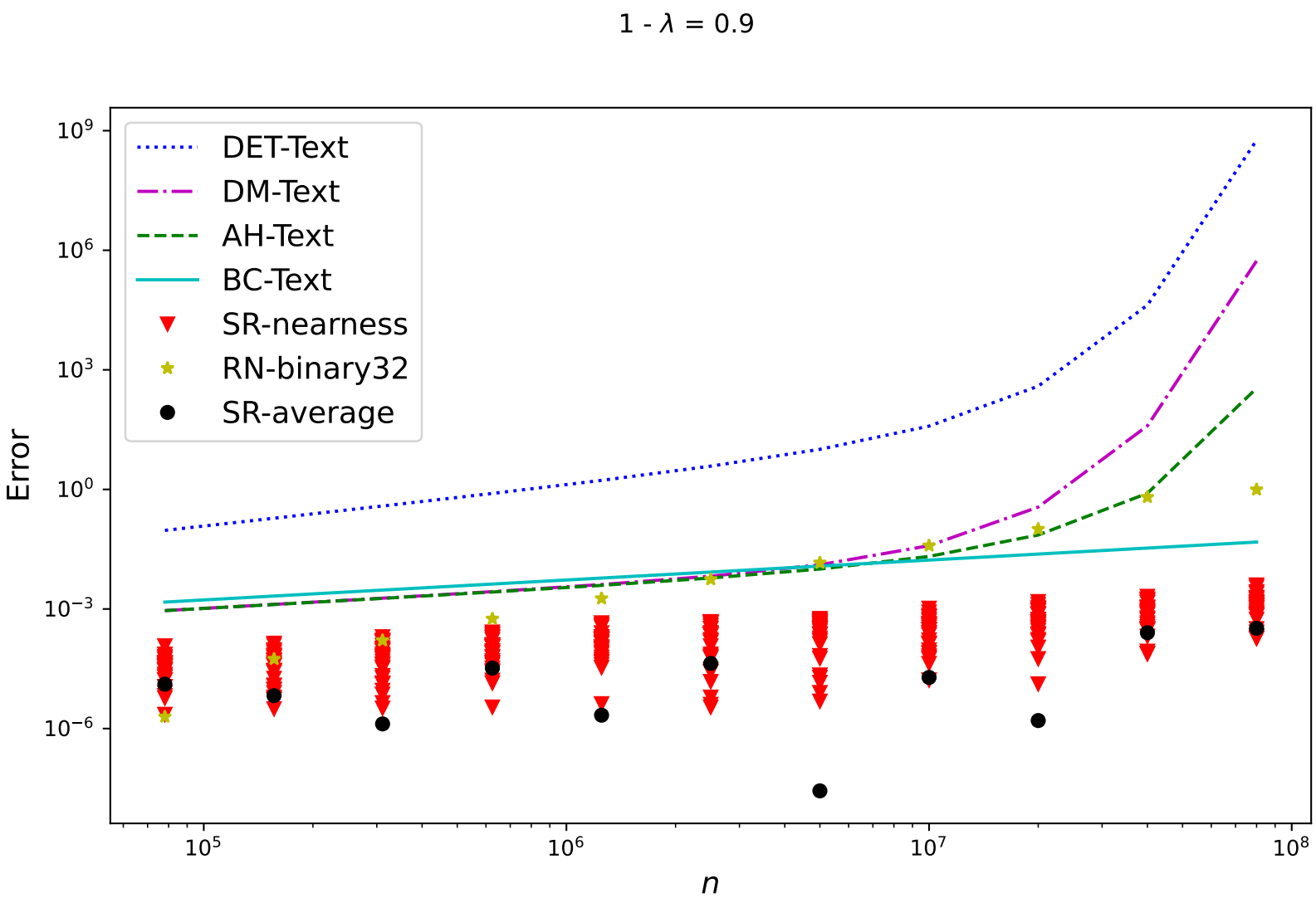



Figure: Probabilistic error bounds over $n$ with probability $1 - \lambda = 0.9$, $u = 2^{-23}$ for the textbook algorithm and for floating-points chosen uniformly at random in $\in [0; 1]$.

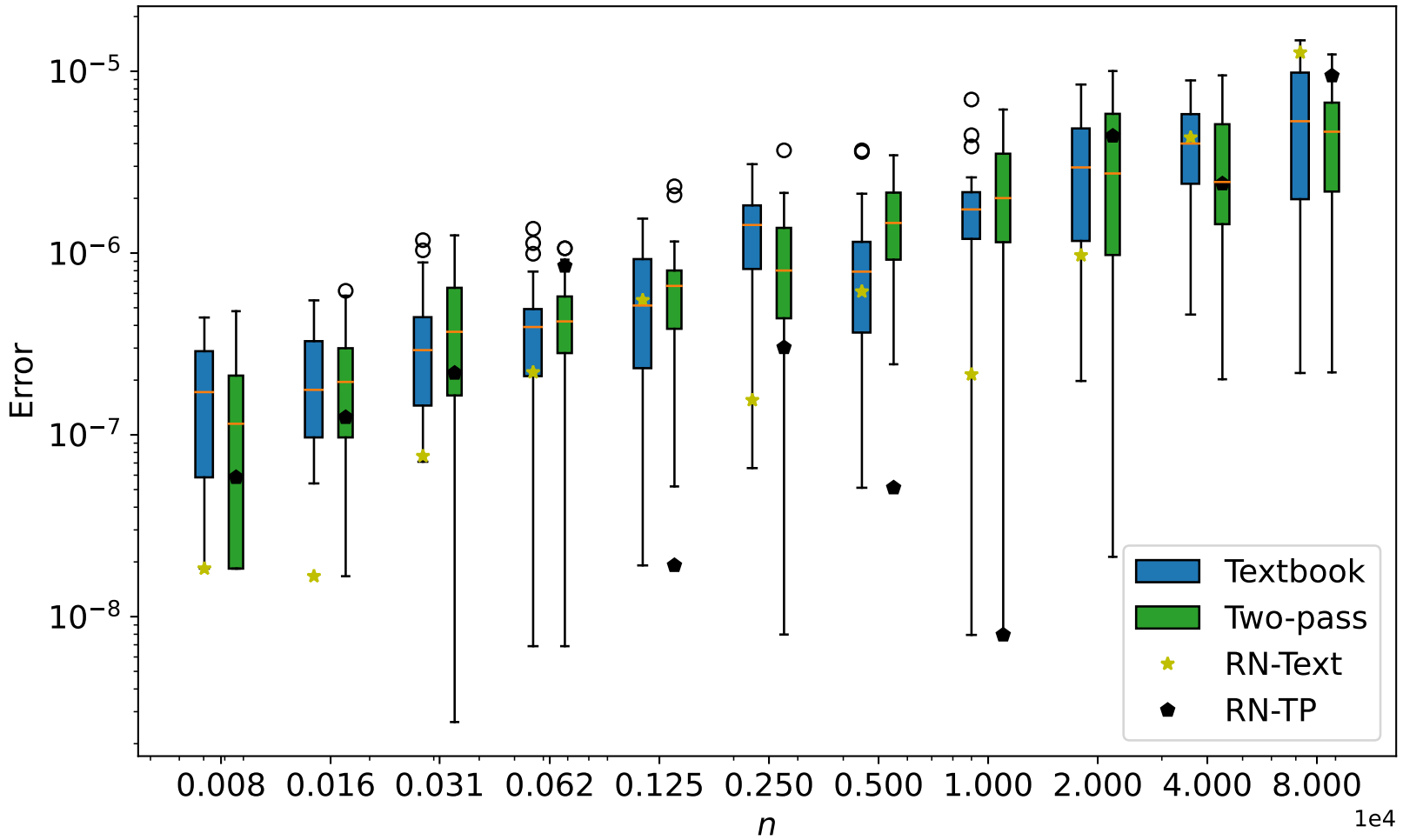



Figure: The forward errors of textbook and two-pass algorithms in binary32 precision for floating-points chosen uniformly at random in $[-1; 1]$. $y = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}s^2$ and $z = \sum_{i=1}^{n} (x_i - m)^2$.

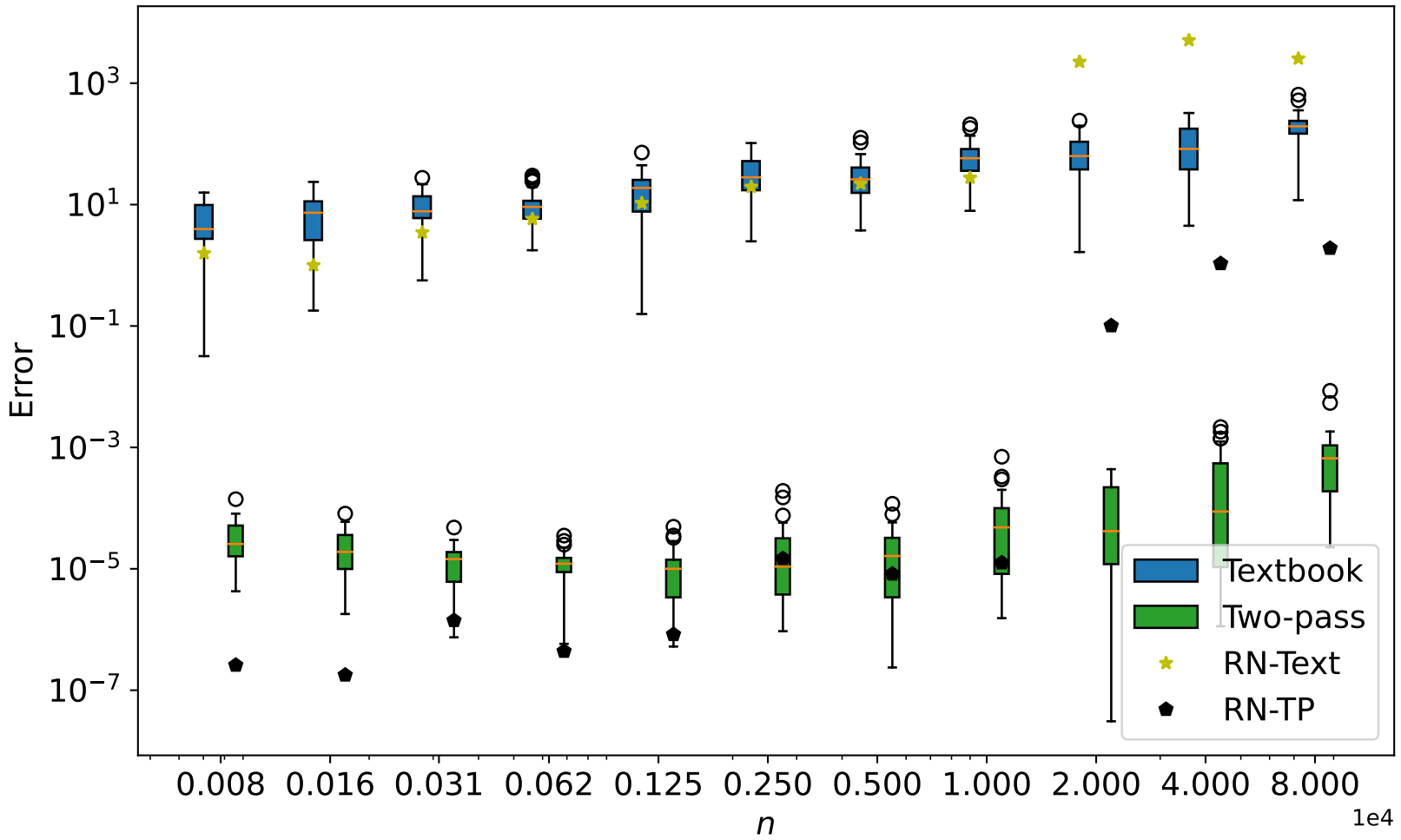



Figure: The forward errors of textbook and two-pass algorithms in binary32 precision for floating-points chosen uniformly at random in $[1024; 1025]$. $y = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}s^2$ and $z = \sum_{i=1}^{n} (x_i - m)^2$.

# Conclusion and Future Works

## Contributions

Using SR-nearness:

- Textbook and two-pass algorithms.
- New approach for algorithms with non-linear errors.

## Future works

- Explore the division case (WIP).
- General framework.

*Thank You For Your Attention.*

# Hardware with SR

- Graphcore IPU supports SR for binary32 and binary16 arithmetic.
- The Intel neuromorphic chip Loihi.
- Patents from IBM, AMD, or NVIDIA.
- **Survey**: Stochastic rounding: Implementation, error analysis, and applications.

**Theorem 2 (Textbook algorithm).**

*For all* $0 < \lambda < 1$, *the computed* $\widehat{y}$ *satisfies under SR-nearness*

$$\frac{\left|\widehat{y} - y\right|}{|y|} \leqslant \mathcal{K}_2^2 \sqrt{2\gamma_{n+1}(u^2)/\lambda} + \mathcal{K}_1^2 \left( (1+u)^3 \left( \sqrt{2\gamma_{n-1}(u^2)/\lambda} + 1 \right)^2 - 1 \right),$$

*with probability at least* $1 - \lambda$.

**Theorem 3 (Two-pass algorithm).**

*For all* $0 < \lambda < 1$, *the computed* $\widehat{z}$ *satisfies under SR-nearness*

$$\frac{\left|\widehat{z} - z\right|}{|z|} \leqslant (1+u) \left( \sqrt{\frac{4\gamma_{n+1}(u^2)}{\lambda}} + \frac{4\gamma_{n+1}(u^2)}{\lambda} \left( 2\mathcal{K}_1 + \mathcal{K}_1^2 \left( \sqrt{\frac{4\gamma_{n+1}(u^2)}{\lambda}} + 1 \right) \right) \right) + u,$$

*with probability at least* $1 - \lambda$.

**Theorem 4 (Textbook algorithm).**

*For all* $0 < \lambda < 1$, *the computed* $\widehat{y}$ *satisfies under SR-nearness*

$$\frac{\left|\widehat{y} - y\right|}{|y|} \leqslant \mathcal{K}_2^2 \sqrt{u\gamma_{2(n+1)}(u)}\sqrt{\ln(4/\lambda)} + \mathcal{K}_1^2 \left( (1+u)^3 \left( \sqrt{u\gamma_{2(n-1)}(u)}\sqrt{\ln(4/\lambda)} + 1 \right)^2 - 1 \right),$$

*with probability at least* $1 - \lambda$.

**Theorem 5 (Two-pass algorithm).**

*For all* $0 < \lambda < 1$, *the computed* $\widehat{z}$ *satisfies under SR-nearness*

$$\begin{aligned}\frac{\left|\widehat{z} - z\right|}{|z|} \leqslant (1+u) \Bigg( & \sqrt{u\gamma_{2(n+1)}(u)}\sqrt{\ln(8/\lambda)} \\ & + u\gamma_{2(n+1)}(u)\ln(8/\lambda) \left( 2\mathcal{K}_1 + \mathcal{K}_1^2 \left( \sqrt{u\gamma_{2(n+1)}(u)}\sqrt{\ln(8/\lambda)} + 1 \right) \right) \Bigg) + u,\end{aligned}$$

*with probability at least* $1 - \lambda$.

**Theorem 6 (Textbook algorithm).**

*For all* $0 < \lambda < 1$, *the computed* $\widehat{y}$ *satisfies under SR-nearness*

$$\frac{\left|\widehat{y} - y\right|}{|y|} \leqslant \mathcal{K}_2^2 \sqrt{u\gamma_{2(n+1)}(u)}\sqrt{\ln(4/\lambda)} + \mathcal{K}_1^2(1+u)^3\Big[\sqrt{2u\gamma_{4(n-1)}(u)}\sqrt{\ln(4/\lambda)} + u\frac{\gamma_{2(n-1)}(u)}{2} + 1\Big] - \mathcal{K}_1^2,$$

*with probability at least* $1 - \lambda$.

**Theorem 7 (BC method).**

*For the pairwise textbook, for all* $0 < \lambda < 1$, *the computed* $\widehat{y}$ *satisfies under SR-nearness*

$$\frac{\left|\widehat{y} - y\right|}{|y|} \leqslant \mathcal{K}_2^2 \sqrt{2\gamma_{\log(n)+1}(u^2)/\lambda} + \mathcal{K}_1^2 \left( (1+u)^3 \left( \sqrt{2\gamma_{\log(n)}(u^2)/\lambda} + 1 \right)^2 - 1 \right),$$

*with probability at least* $1 - \lambda$.

**Theorem 8 (AH method).**

*For the pairwise textbook, for all* $0 < \lambda < 1$, *the computed* $\widehat{y}$ *satisfies under SR-nearness*

$$\frac{\left|\widehat{y} - y\right|}{|y|} \leqslant \mathcal{K}_2^2 \sqrt{u\gamma_{2(\log(n)+1)}(u)} \sqrt{\ln(4/\lambda)} + \mathcal{K}_1^2 \left( (1+u)^3 \left( \sqrt{u\gamma_{2\log(n)}(u)} \sqrt{\ln(4/\lambda)} + 1 \right)^2 - 1 \right),$$

*with probability at least* $1 - \lambda$.