Stochastic Rounding Variance and Probabilistic Bounds: A New Approach

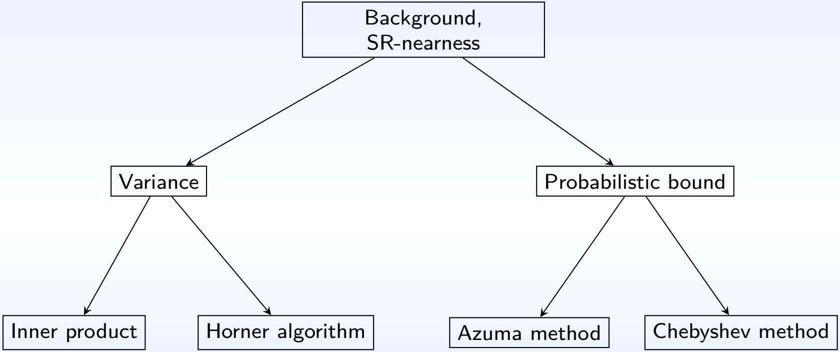
El-Mehdi El ARAR
el-mehdi.el-arar@uvsq.fr

Devan SOHIER, Pablo DE OLIVEIRA CASTRO and Eric PETIT

Plan

Preprint submitted for publication (Available in Arxiv:2207.10321)

Stochastic rounding variance and probabilistic bounds: A new approach.



Background

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

- For $x, y \in \mathcal{F}$ and 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 $\theta(x) = \frac{x - \lfloor x \rfloor}{\lceil x \rceil - \lfloor x \rfloor}$.

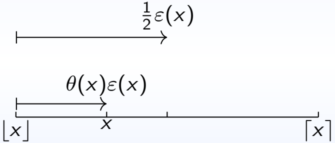


Figure: $\theta(x)$ is the fraction of $\varepsilon(x)$ to be rounded away.

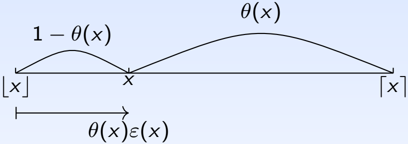


Figure: **SR_nearness**.

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

SR-nearness and mean independence

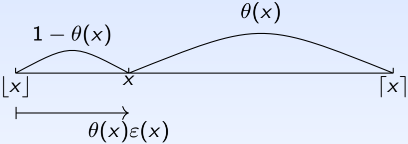


Figure: **SR_nearness**.

- $E(\hat{x}) = \theta(x)\lceil x \rceil + (1-\theta(x))\lfloor x \rfloor = x$.
- 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. δ_1, δ_2 are random variables such that $\mathbb{E}(\delta_1) = \mathbb{E}(\delta_2) = 0$.
- Mean independence: $X_1, X_2, \ldots$ are mean independent if $\mathbb{E}[X_k / X_1, \ldots, X_{k-1}] = \mathbb{E}(X_k)$ 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 $\delta_1, \delta_2, \ldots$, in that order obtained from SR-nearness, the δ_k are random variables with mean zero such that $\mathbb{E}[\delta_k / \delta_1, \ldots, \delta_{k-1}] = \mathbb{E}(\delta_k) = 0$.

The variance of the error for stochastic rounding

Assume that $\delta_1, \delta_2, \dots$ in that order are random errors on elementary operations obtained from SR-nearness. $\phi_j = \prod_{k=j}^{n}(1+\delta_k)$. For $i < j$ we have

$$\phi_i = \prod_{k=i}^{j-1}(1+\delta_k)\prod_{k=j}^{n}(1+\delta_k) = \prod_{k=i}^{j-1}(1+\delta_k)\phi_j.$$

Let K a subset of $\mathbb{N}$ of cardinal n and $\psi_K = \prod_{k\in K}(1+\delta_k)$. Since $|\delta_k| \leqslant u$ for all $k \in K$ we have

$$|\psi_K| \leqslant (1+u)^n.$$

Throughout this presentation, let $\gamma_n(u) = (1+u)^n - 1$ and $K \triangle K' = (K \cup K') \setminus (K \cap K')$.

Lemma 2.

Under SR-nearness ψ_K satisfies

1. $E(\psi_K) = 1$.
2. *Let $K' \subset \mathbb{N}$ such that $|K \cap K'| = m$, under the assumption that $\forall j \in K\triangle K', k \in K \cap K', j < k$ we have*

$$0 \leqslant \mathrm{Cov}(\psi_K, \psi_{K'}) \leqslant \gamma_m(u^2).$$

3. $V(\psi_K) \leqslant \gamma_n(u^2)$,

where $\gamma_n(u^2) = (1+u^2)^n - 1 = nu^2 + O(u^3)$.

For $a, b \in \mathbb{R}^n$ such that $y = a^\top b$, let $s_i = s_{i-1} + a_i b_i$. The computed $\widehat{s_i}$ satisfy $\widehat{s_1} = a_1 b_1(1+\delta_1)$ and

$$\widehat{s_i} = (\widehat{s}_{i-1} + a_i b_i(1+\delta_{2(i-1)}))(1+\delta_{2i-1}), \quad |\delta_{2(i-1)}|, |\delta_{2i-1}| \leqslant u,$$

for all $2 \leqslant i \leqslant n$. We thus have

$$\widehat{y} = \widehat{s}_n = \sum_{i=1}^{n} a_i b_i (1+\delta_{2(i-1)}) \prod_{k=i}^{n} (1+\delta_{2k-1}).$$

Theorem 3.

Under SR-nearness, the computed $\widehat{y}$ satisfies $E(\widehat{y}) = y$ and

$$V(\widehat{y}) \leqslant y^2 K_1^2 \gamma_n(u^2),$$

where $K_1 = \frac{\sum_{i=1}^n |a_i b_i|}{|\sum_{i=1}^n a_i b_i|}$ is the condition number for the computed $y = \sum_{i=1}^n a_i b_i$ using the 1-norm and $\gamma_n(u^2) = (1+u^2)^n - 1 = nu^2 + O(u^3)$.

Horner Algorithm

Let $P(x) = \sum_{i=0}^{n} a_i x^i$, Horner rule consists in writing this polynomial as

$$P(x) = (((a_n x + a_{n-1})x + a_{n-2})x ... + a_1)x + a_0.$$

Under SR-nearness

$$\widehat{P}(x) = \sum_{i=0}^{n} a_i x^i \prod_{k=2(n-i)}^{2n} (1 + \delta_k).$$

Theorem 4.

Using SR-nearness, the computed $\widehat{P}(x)$ satisfies $E\left(\widehat{P}(x)\right) = P(x)$ and

$$V(\widehat{P}(x)) \leqslant P(x)^2 cond_1(P,x)^2 \gamma_{2n}(u^2),$$

where $cond_1(P,x) = \frac{\sum_{i=1}^{n} |a_i x^i|}{|\sum_{i=1}^{n} a_i x^i|}$ is the condition number for the computed $P(x) = \sum_{i=1}^{n} a_i x^i$ using the 1-norm.

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 $0 < \lambda < 1$.

How form the martingale?

Inner product
$\widehat{y} = \sum_{i=1}^{n} a_i b_i (1 + \delta_{2(i-1)}) \prod_{k=i}^{n} (1 + \delta_{2k-1})$

Errors accumulated in the whole process (at each $a_i b_i$)
N. J. Higham and T. Mary, "A new approach to probabilistic rounding error analysis"

$M_n = (1 + \delta_{2(i-1)}) \prod_{k=i}^{n} (1 + \delta_{2k-1})$

Probabilistic bound proportional to $u\sqrt{n} \ln n$ at any fixed probability.

Partial sums of the problem
Ilse C. F. Ipsen and Hua Zhou, "Probabilistic Error Analysis for Inner Products,"

$M_n = \widehat{y}$

Probabilistic bound proportional to $u\sqrt{n}$ at any fixed probability.

Horner algorithm bound

Let $P(x) = \sum_{i=0}^{n} a_i x^i$, Horner method consists in writing this polynomial as

$$P(x) = (((a_n x + a_{n-1})x + a_{n-2})x \ldots + a_1)x + a_0.$$

Theorem 5.

Under SR-nearness,

- *The deterministic bound*

$$\frac{|\widehat{P}(x) - P(x)|}{|P(x)|} \leqslant cond_1(P, x)\gamma_{2n}(u),$$

  where $cond_1(P, x) = \frac{\sum_{i=1}^{n} |a_i x^i|}{|P(x)|}$ *is the condition number of the polynomial evaluation and* $\gamma_{2n}(u) = (1+u)^{2n} - 1 = 2nu + O(u^2)$.
- *For all* $0 < \lambda < 1$ *and with probability at least* $1 - \lambda$

$$\frac{|\widehat{P}(x) - P(x)|}{|P(x)|} \leqslant cond_1(P, x)\sqrt{u\gamma_{4n}(u)}\sqrt{\ln(2/\lambda)},$$

  where $\sqrt{u\gamma_{4n}(u)} \approx 2\sqrt{n}u$.

Lemma 6.

Let X be a random variable with finite expected value and finite non-zero variance. For any real number $\alpha > 0$,

$$\mathbb{P}\Big(|X - E(X)| \leqslant \alpha\sqrt{V(X)}\Big) \geqslant 1 - \frac{1}{\alpha^2}.$$

Inner product:

$$\frac{|\widehat{y} - y|}{|y|} \leqslant K_1\sqrt{\gamma_n(u^2)/\lambda},$$

with probability at least $1 - \lambda$.

Horner algorithm:

$$\frac{|\widehat{P}(x) - P(x)|}{|P(x)|} \leqslant cond_1(P, x)\sqrt{\gamma_{2n}(u^2)/\lambda},$$

with probability at least $1 - \lambda$.

Both are proportional to $\sqrt{n}u$.

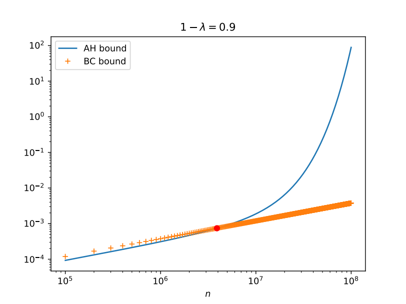


Figure: AH bound vs BC bound with probability 0.9 and $u = 2^{-23}$ for the inner product.

Probability	u	Precision format	$n \gtrsim$
$1 - \lambda = 0.99$	2^{-7}	bfloat16	220
	2^{-10}	fp16	1810
	2^{-23}	fp32	$1.48e07$
	2^{-52}	fp64	$7.9e15$

Figure: The smallest n such that BC method gives a tighter probabilistic bound than AH method for the inner product.

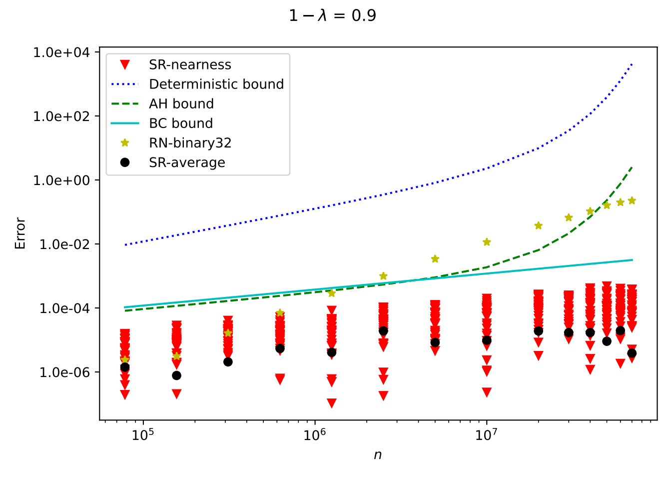


Figure: Probabilistic bounds with probability $1-\lambda = 0.9$ vs deterministic bound of the computed forward errors of the inner product with $u = 2^{-23}$ using verificarlo where $a, b \in [0; 1]$.

Conclusion and Future Works

Contributions

- Lemma allows to give a variance bound for large family of algorithms.
- An extension of the AH method to the Horner algorithm.
- A tight probabilistic bound in $O(\sqrt{n}u)$.

Future works

- Evaluate the applicability of SR to complex HPC codes.
- Show the advantage of SR in complex algorithms.

Thank You For Your Attention.

Algorithms	two-pass	text-book
Deterministic bound	$nu + K_1 n^2 u^2 + 2K_1^2 n^2 u^2$	$nu(K_2^2 + 2K_1^2)$
Probabilistic bound	$\sqrt{n}u + K_1 nu^2 + 2K_1^2 nu^2$ $+\sqrt{n}u^2$	$\sqrt{n}u(K_2^2 + K_1^2)$

Figure: Forward error bounds.

Remark 1.

- *text-book:* $y = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2$.
- *two-pass:* $y = \sum_{i=1}^{n}(x_i - \bar{x})^2$.
- $K_1 = \frac{\|x\|_1}{\sqrt{ny}}$, $K_2 = \frac{\|x\|_2}{\sqrt{y}}$ *with* $K_1 \leqslant K_2$.

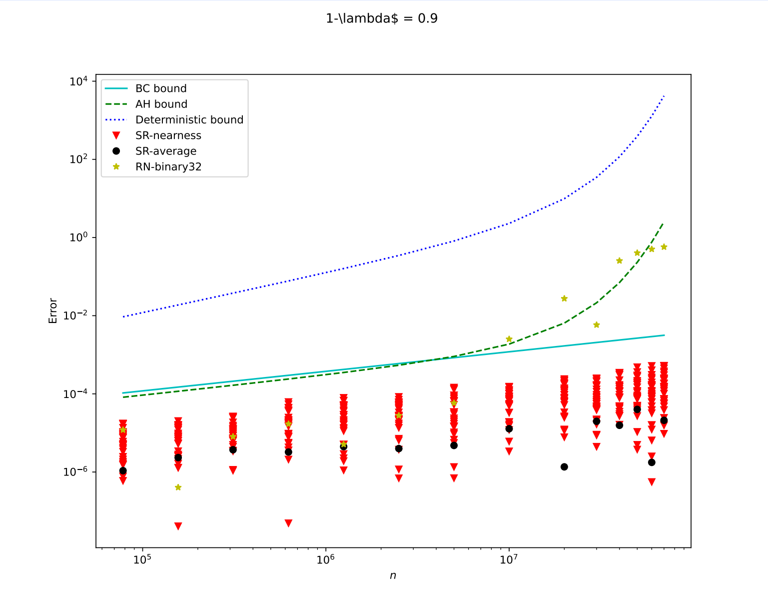


Figure: Probabilistic bounds with probability $1-\lambda = 0.9$ vs deterministic bound of the computed forward errors of the inner product with $u = 2^{-23}$ using verificarlo where $a, b \in [1;2]$.

We focus on the inner product bounds.

AH $=\mathcal{K}_1\sqrt{\frac{u}{2}\gamma_{2n}(u)}\sqrt{2\ln(2/\lambda)}$ and BC$=\mathcal{K}_1\sqrt{\gamma_n(u^2)}\sqrt{1/\lambda}$. Firstly,

$$\sqrt{\gamma_n(u^2)} \leqslant \sqrt{\frac{u}{2}\gamma_{2n}(u)} \text{ for all } n \geqslant 1.$$

let us compare $\sqrt{1/\lambda}$ and $\sqrt{2\ln(2/\lambda)}$ for $\lambda \in]0;1[$,

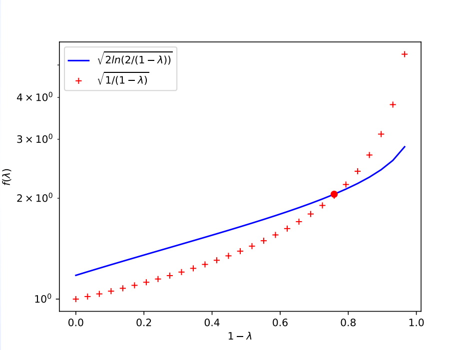
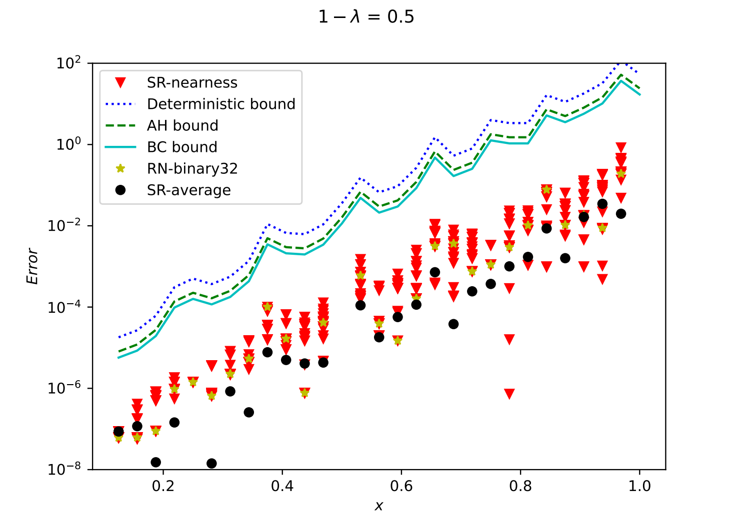


Figure: Illustration of $\sqrt{1/\lambda}$ and $\sqrt{2\ln(2/\lambda)}$ behaviour for all $\lambda \in]0;1[$.

Chebyshev polynomial $P(x) = T_{20}(x) = \sum_{i=0}^{10} a_i (x^2)^i$. For each value of x, we perform the computation 30 times and plot all samples as well as the forward error of the average of the 30 SR instances.



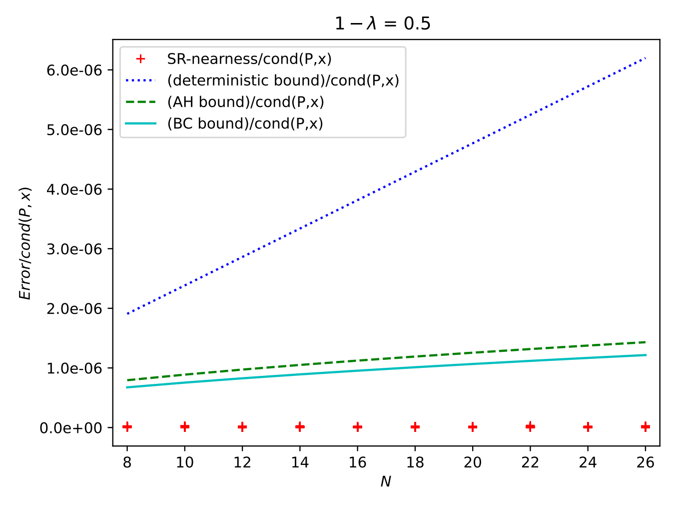


Figure: Forward errors/cond(P,x) of Horner rule for Chebyshev polynomial $T_N(24/26)$. For each value of N, the computation is performed 30 times and all samples of SR-nearness are plotted.

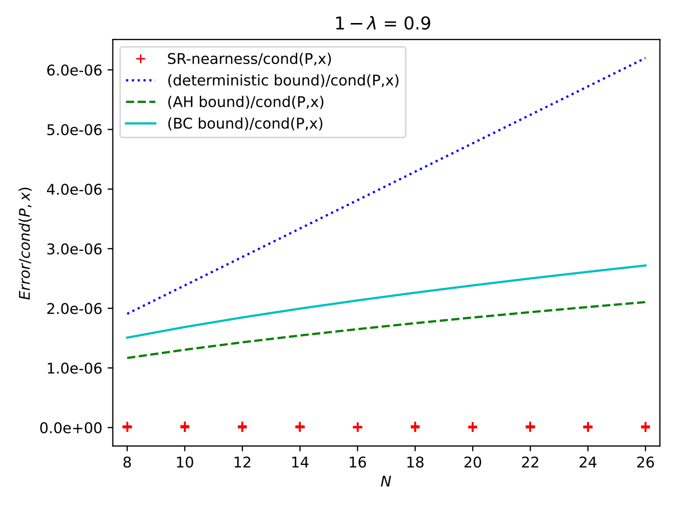


Figure: Forward errors/cond(P,x) of Horner rule for Chebyshev polynomial $T_N(24/26)$. For each value of N, the computation is performed 30 times and all samples of SR-nearness are plotted.

		verificarlo backends		
	original	IEEE	MCA quad	MCA integer
Kahan binary32	1.34s	2.36s ($\times$1.7)	6.28s ($\times$4.7)	7.76s ($\times$5.8)
Kahan binary64	1.34s	2.34s ($\times$1.7)	105s ($\times$78)	64s ($\times$48)
NAS CG A	0.80s	6.41s ($\times$8)	173s ($\times$216)	128s ($\times$160)

Table 6.2: Execution time (and slowdown) for a Kahan sum of 100 millions elements and for the NAS CG A using different verificarlo backends.

Forward error deterministic bound

- Inner product: $y = a^\top b$, where $a, b \in \mathbb{R}^n$

$$\frac{|\widehat{y} - y|}{|y|} \leqslant K\gamma_n,$$

where $K = \frac{\sum_{i=1}^n |a_i b_i|}{|\sum_{i=1}^n a_i b_i|}$ is the condition number and $\gamma_n = (1+u)^n - 1 = nu + O(u^2)$.

- Inner product: $y = a^\top b$, where $a, b \in \mathbb{R}^n$

$$\frac{|\widehat{y} - y|}{|y|} \leqslant K\gamma_n,$$

where $K = \frac{\sum_{i=1}^n |a_i b_i|}{|\sum_{i=1}^n a_i b_i|}$ is the condition number and $\gamma_n = (1+u)^n - 1 = nu + O(u^2)$.

$|\hat{x} - x|$ SR-nearness $\Longrightarrow x = E(\hat{x})$ then $|\hat{x} - E(\hat{x})|$.

Concentration inequality: Markov's inequality, Chebyshev's inequality, Azuma–Hoeffding inequality...