STOCHASTIC ROUNDING VARIANCE AND PROBABILISTIC BOUNDS: A NEW APPROACH

El-Mehdi El Arar, Devan Sohier, Pablo de Oliveira Castro, Eric Petit

Stochastic Rounding Variance and Probabilistic Bounds: A New Approach

El-Mehdi El Arar[1], Devan Sohier[1], Pablo de Oliveira Castro[1] and Eric Petit[2]

[1]Université Paris-Saclay, UVSQ, LI-PaRAD

[2] Intel Corp

{el-mehdi.el-arar, devan.sohier, pablo.oliveira}@uvsq.fr

eric.petit@intel.com

Abstract

Stochastic rounding (SR) offers an alternative to the deterministic IEEE-754 floating-point rounding modes. In some applications such as PDEs, ODEs and neural networks, SR empirically improves the numerical behavior and convergence to accurate solutions while the theoretical background remains partial. Recent works by Ipsen, Zhou, Higham, and Mary have computed SR probabilistic error bounds for basic linear algebra kernels. For example, the inner product SR probabilistic bound of the forward error is proportional to $\sqrt{n}u$ instead of nu for the default rounding mode. To compute the bounds, these works show that the errors accumulated in computation form a martingale.

This paper proposes an alternative framework to characterize SR errors based on the computation of the variance. We pinpoint common error patterns in numerical algorithms and propose a lemma that bounds their variance. For each probability and through Bienaymé–Chebyshev inequality, this bound leads to better probabilistic error bound in several situations. Our method has the advantage of providing a tight probabilistic bound for all algorithms fitting our model. We show how the method can be applied to give SR error bounds for the inner product and Horner polynomial evaluation.

Keywords— Stochastic rounding, Floating-point arithmetic, Concentration inequality, Inner product, Polynomial evaluation, Horner algorithm.

AMS subject classifications— 65G50, 65F05

1 Introduction

Stochastic rounding (SR) is an idea proposed in the 1950s by von Neumann and Goldstine [27]. First, it can be used to estimate empirically the numerical error of computer programs; SR introduces a random noise in each floating-point operation and then a statistical analysis of the set of sampled outputs can be applied to estimate the effect of rounding errors. To make this simulation available, various tools such as Verificarlo [13], Verrou [14] and Cadna [21] have been developed. Second, SR can be used as a replacement for the default deterministic rounding mode in numerical computations. It has been demonstrated that in multiple domains such as neural networks, ODEs, PDEs, and Quantum mechanics [9], SR provides better results compared to the IEEE-754 default rounding mode [3]. Connolly et al. [8] show that SR successfully prevents the phenomenon of stagnation that takes place in various applications such as neural networks, ODEs and PDEs. In particular, Gupta et al show in [15] that deep neural networks are prone to stagnation during the training phase. For PDEs, solved via Runge-Kutta finite difference methods in low precision, SR avoids stagnation in the computations of the heat equation solution as proved in [10].

Hardware units proposing stochastic rounding are still unavailable in most computers. However, it has been introduced in various specialized processors such as Graphcore IPUs [1] which supports SR for 32 bits floating point, *binary32*, and 16 bits floating point, *binary16*, or Intel neuromorphic chip Loihi [12] to improve the accuracy of biological neuron and synapse models. Also, AMD [2], NVIDIA [4], IBM [6, 7], and other computing companies [16, 22, 23] own several related patents. These developments support the idea of hardware implementations using SR becoming more available in the future.

Most current hardware implements the IEEE-754 standard, that defines five rounding modes for floating-point arithmetic which are all deterministic [3]: round to nearest ties to even (default), round to nearest ties away, round to zero, round to $+\infty$ and round to $-\infty$. SR, on the other hand, is a non-deterministic rounding mode: for a number that cannot be represented exactly in the working precision, it randomly chooses the next larger or smaller floating-point number. In the literature, several properties and results of SR have been proven.

Connolly et al show in [8] show that under SR-nearness the expected value coincides with the exact value for a large family of algorithms.

Based on the Azuma-Hoeffding inequality and the martingale theory, recent works on the inner product [20] show that SR probabilistic bound of the forward error is proportional to $\sqrt{n}u$ rather than nu when $nu \ll 1$. Also, the martingale central limit theorem implies that under certain conditions, the error converges in distribution to a normal distribution that is characterized by its mean and variance [11]. This behaviour is often observed in practice. In this case, the number of significant digits can be estimated by $-\log(\frac{\sigma}{|\mu|})$ where σ is the standard deviation (the square root of the variance) and μ is the expected value [26]. However, the results of this paper don't use any of these assumptions.

Variance also allows to use several probabilistic properties such as concentration inequalities that provide a bound on how a random variable deviates from some value (typically, its expected value) [5]. To our knowledge, the variance analysis of a SR computation has not attracted any attention in the literature. The purpose of this paper is to further the probabilistic investigation of SR with the following contributions:

0. We review the works proposed by M. P Connolly, N.J. Higham and T. Mary [8] and I.C.F. Ipsen, H. Zhou [20] that show the forward error for the inner product is proportional respectively to $\sqrt{n\ln(n)}u$ and $\sqrt{n}u$ at any probability $\lambda \leq 1$ rather than to the deterministic bound of nu [20].
1. Under stochastic rounding and without any additional assumption, we propose Lemma 3.1, a general framework applicable to a wide class of algorithms that allows to compute a variance bound. We choose the inner product and Horner algorithms as applications. Our bound is deterministic and depends on the condition number, the problem size n and the unit roundoff of the floating-point arithmetic in use u.
2. We extend the method proposed in [20] to derive a new forward error bound of the Horner algorithm in $O(\sqrt{n}u)$. This illustrates how these tools can be applied (with some work) to any algorithm based on a fixed sequence of sum and products.
3. We introduce a new approach to derive a probabilistic bound in $O(\sqrt{n}u)$ based on the variance calculation and Bienaymé–Chebyshev inequality. This approach gives a tighter forward error bound than the previous bounds [8, 20] for a probability at most 0.758. This bound remains tight from a rank n high with respect to u.

Interestingly, the variance method introduces a tight probabilistic error bound in low precision. In this regard, studying algorithms under stochastic rounding in low precision, especially bfloat-16 is becoming increasingly attractive due to its higher speed and lower energy consumption. Recent works show that in various domains such as PDEs [10], ODEs [19] and neural networks [15], SR provides positive effects compared to the deterministic IEEE-754 [3] default rounding mode in this precision format.

Section 2 presents the background on floating-point arithmetic and more particularly SR-nearness, a stochastic rounding mode introduced in [25, p. 34], that has the important property of being unbiased. It also satisfies the mean independence property, an assumption weaker than independence yet powerful enough to yield important results by martingale theory.

Section 3 is articulated around Lemma 3.1 that bounds the variance of the numerical error for a wide class of algorithms. We apply this result to the inner product and Horner algorithms in Theorem 3.1 and Theorem 3.2, respectively.

Section 4 shows that, under SR-nearness rounding, the numerical error of these two algorithms is probabilistically bounded in $O(\sqrt{n}u)$ instead of the deterministic bound in $O(nu)$. We first prove it with the Azuma–Hoeffding inequality and martingale theory: we analyze techniques used for the inner product in works by Higham and Mary [8, 18] and Ipsen and Zhou [20], point the difference in these two works, and adapt them to compute the relative error of the Horner method for polynomial evaluation. We then use the Bienaymé–Chebyshev inequality which, combined to the previous variance bound, leads to a probabilistic bound in $O(\sqrt{n}u)$.

The probabilistic bounds above depend on three parameters: the precision u, the problem size n, and the probability λ that a SR-nearness computation has an error greater than the bound. In Section 5, we analyze these probabilistic bounds and we show that the one obtained by the Bienaymé-Chebyshev inequality is tighter in many cases; in particular for any given λ and u, there exist a problem size n above which the Bienaymé–Chebyshev bound is tighter.

Numerical experiments in Section 6 illustrate the quality of these bounds on the two aforementioned algorithms and compare them to deterministic rounding.

2 Notations and definitions

2.1 Notation

Throughout this paper, for a random variable X, $E(X)$ denotes its expected value, $V(X)$ denotes its variance and $\sigma(X)$ denotes its standard deviation. The conditional expectation of X given Y is $\mathbb{E}[X/Y]$.

2.2 Floating-point background

A normal floating-point number in such a format is a number x for which there exists a triple (s, m, e) such that $x = \pm m \times \beta^{e-p}$, where β is the basis, e is the exponent, p is the working precision, and m is an integer (the significand) such that $\beta^{p-1} \le m < \beta^{p}$. We only consider normal floating-point numbers; detailed information on the floating-point format most generally in use in current computer systems is defined in the IEEE-754 norm [3].

Let us denote $\mathcal{F} \subset \mathbb{R}$ the set of normal floating-point numbers and $x \in \mathbb{R}$. Upward rounding $\lceil x \rceil$ and downward rounding $\lfloor x \rfloor$ are defined by:

$$\lceil x \rceil = \min\{y \in \mathcal{F} : y \ge x\}, \quad \lfloor x \rfloor = \max\{y \in \mathcal{F} : y \le x\},$$

by definition, $\lfloor x \rfloor \le x \le \lceil x \rceil$, with equalities if and only if $x \in \mathcal{F}$. The floating-point approximation of a real number $x \neq 0$ is one of $\lfloor x \rfloor$ or $\lceil x \rceil$:

$$\mathrm{fl}(x) = x(1 + \delta), \tag{2.1}$$

where $\delta = \frac{\mathrm{fl}(x) - x}{x}$ is the relative error: $|\delta| \le \beta^{1-p}$. In the following, we use the same notation as [8, 20] $u = \beta^{1-p}$. IEEE-754 mode RN (round to nearest, ties to even) has the stronger property that $|\delta| \le \frac{1}{2}\beta^{1-p} = \frac{1}{2}u$. In many works focusing on IEEE-754 RN, u is chosen instead to be $\frac{1}{2}\beta^{1-p}$.

For $x, y \in \mathcal{F}$, the considered rounding modes verify $\mathrm{fl}(x \,\mathrm{op}\, y) \in \{\lfloor x \,\mathrm{op}\, y \rfloor, \lceil x \,\mathrm{op}\, y \rceil\}$ for op $\in \{+, -, *, /\}$. Moreover, for IEEE-754 rounding modes [3] and stochastic rounding [8] the error in one operation is bounded:

$$\mathrm{fl}(x \,\mathrm{op}\, y) = (x \,\mathrm{op}\, y)(1 + \delta), \ |\delta| \le u, \tag{2.2}$$

specifically for RN we have $|\delta| \le \frac{1}{2}u$. Let us assume that x is a real that is not representable: $x \in \mathbb{R} \setminus \mathcal{F}$. The machine-epsilon or the distance between the two floating-point numbers enclosing x is $\epsilon(x) = \lceil x \rceil - \lfloor x \rfloor = \beta^{e-p}$. Since $\beta^{p-1} \le m < \beta^{p}$, then $\beta^{e-1} \le |x| < \beta^{e}$ and

$$|\epsilon(x)| = \beta^{e-1} u \le |x| u. \tag{2.3}$$

The fraction of $\epsilon(x)$ rounded away, as shown in Figure 1, is $\theta(x) = \frac{x - \lfloor x \rfloor}{\lceil x \rceil - \lfloor x \rfloor}$.

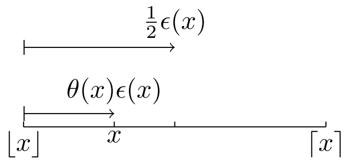

Figure 1: $\theta(x)$ is the fraction of $\epsilon(x)$ to be rounded away.

Let us denote a problem of size n and precision u, in this paper $nu \ll 1$ means $n \to \infty$, $u \to 0$ and $nu \to 0$.

2.3 Stochastic rounding definition

Throughout this paper, $\mathrm{fl}(x) = \widehat{x}$ is the approximation of the real number x under stochastic rounding. For $x \in \mathbb{R} \setminus \mathcal{F}$, we consider the following stochastic rounding mode, called SR-nearness:

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

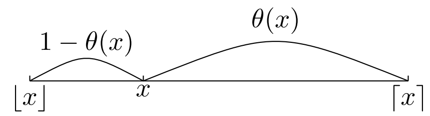

Figure 2: **SR-nearness**.

SR-nearness mode is unbiased [25, p. 34].

$$\begin{aligned} E(\widehat{x}) &= \theta(x)\lceil x \rceil + (1 - \theta(x))\lfloor x \rfloor \\ &= \theta(x)(\lceil x \rceil - \lfloor x \rfloor) + \lfloor x \rfloor = x. \end{aligned}$$

In the following, we focus on this stochastic rounding mode. In general and under SR-nearness, the error terms in algorithms appear as a sequence of random variables such that the independence property does not hold. However, a weaker, and yet fruitful, assumption, called mean independence, does.

Definition 2.1. *A random variable Y is said to be mean independent from random variable X if its conditional mean $\mathbb{E}[Y/X] = \mathbb{E}(Y)$. The random sequence $(X_1, X_2, \ldots)$ is mean independent if $\mathbb{E}[X_k/X_1, ..., X_{k-1}] = \mathbb{E}(X_k)$ for all k.*

Proposition 2.1. *Let X and Y be two real random variables:*

1. *If X and Y are independent then X is mean independent from Y.*
2. *If X is mean independent from Y then X and Y are uncorrelated.*

The reciprocals of these two implications are false.

For $a, b \in \mathcal{F}$, and $c \leftarrow a \operatorname{op} b$ the result of an elementary operation op $\in \{+, -, *, /\}$ obtained from SR-nearness, the relative error δ, such that $\widehat{c} = (a \operatorname{op} b)(1+\delta)$, is a random variable verifying $\mathbb{E}(\delta) = 0$ and $|\delta| \leq u$.

The following lemma has been proven in [8, Lem 5.2] and shows that SR-nearness satisfies the property of mean independence.

Lemma 2.1. *Consider a sequence of elementary operations $c_k \leftarrow a_k \operatorname{op}_k b_k$, with δ_k the error of their kth operation, that is to say, $\widehat{c}_k = (\widehat{a}_k \operatorname{op}_k \widehat{b}_k)(1+\delta_k)$. The δ_k are random variables with mean zero such that $\mathbb{E}[\delta_k/\delta_1, \ldots, \delta_{k-1}] = \mathbb{E}(\delta_k) = 0$.*

3 The variance of the error for stochastic rounding

We now turn to bounding the variance of the error in a computation. If $\widehat{x} = x(1+\delta)$ is the result of an elementary operation rounded with SR-nearness, then $E(\widehat{x}) = x$ and

$$\begin{aligned} V(\widehat{x}) &= E(\widehat{x}^2) - x^2 = \lceil x \rceil^2 \theta(x) + \lfloor x \rfloor^2 (1-\theta(x)) - x^2 \\ &= \theta(x)(\lceil x \rceil^2 - \lfloor x \rfloor^2) - (x^2 - \lfloor x \rfloor^2) \\ &= \theta(x)\epsilon(x)(\lceil x \rceil + \lfloor x \rfloor) - \theta(x)\epsilon(x)(x + \lfloor x \rfloor) \\ &= \theta(x)\epsilon(x)(\lceil x \rceil - x) \\ &= \epsilon(x)^2\theta(x)(1-\theta(x)). \end{aligned}$$

Using (2.3) leads to $V(\widehat{x}) \leq x^2 \frac{u^2}{4}$, in particular $V(\widehat{x}) \leq x^2 u^2$. Lemma 3.1 below allows to estimate the variance of the accumulated errors in a sequence of additions and multiplications. Let K a subset of $\mathbb{N}$ of cardinal n. Assume that $\delta_1, \delta_2, \ldots$ in that order are random errors on elementary operations obtained from SR-nearness. Let us denote

$$\psi_K = \prod_{k \in K} (1+\delta_k).$$

Since $|\delta_k| \leq u$ for all $k \in K$ we have $|\psi_K| \leq (1+u)^n$. Throughout this paper, let $\gamma_n(u) = (1+u)^n - 1$ and $K \triangle K' = (K \cup K') \setminus (K \cap K')$. The following lemma gives some properties of ψ that allows to bound the variance of errors in an algorithm consisting in a fixed sequence of sums and products.

Lemma 3.1. *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 $\forall\, j \in K \triangle K'$, $k \in K \cap K'$, with $j < k$ we have*
$$0 \leq \operatorname{Cov}(\psi_K, \psi_{K'}) \leq \gamma_m(u^2).$$
3. *$V(\psi_K) \leq \gamma_n(u^2)$,*

where $\gamma_n(u^2) = (1+u^2)^n - 1 \approx \exp(nu^2) - 1 = nu^2 + O(u^4)$.

Proof. The first point is an immediate consequence of [8, lem 6.1]. The third point is a particular case of the second with $K = K'$. Let us prove point 2.

$$\operatorname{Cov}(\psi_K, \psi_{K'}) = E(\psi_K \psi_{K'}) - E(\psi_K)E(\psi_{K'}) = E(\psi_K \psi_{K'}) - 1.$$

Assume that $K \cap K' = \{k_1, ..., k_m\}$. Let us denote

$$Q_m := \psi_K \psi_{K'} = \prod_{j \in K \triangle K'} (1+\delta_j) \prod_{l=k_1}^{k_m} (1+\delta_l)^2,$$

such that $j < k_i$ for all $j \in K \triangle K'$ and $i \in \{1, ..., m\}$.

We prove by induction over m that $1 \leq E(Q_m) \leq (1+u^2)^m$. For $m = 0$, we have $K \cap K' = \emptyset$ and $Q_0 = \prod_{j \in K \triangle K'}(1+\delta_j)$, from the first point $E(Q_0) = 1$. Assume that the inequality holds for Q_{m-1}.

$$Q_m = (1+\delta_{k_m})^2 \prod_{l=k_1}^{k_{m-1}} (1+\delta_l)^2 \prod_{j \in K \triangle K'} (1+\delta_j) = (1+\delta_{k_m})^2 Q_{m-1}.$$

Let us denote $\mathcal{S}_{K \triangle K'} = \{\delta_j,\ j \in K \triangle K')\}$, using the law of total expectation $E(X) = E(E[X/Y])$ and lemma 2.1 we have

$$\begin{aligned} E(Q_m) &= E\big((1+\delta_{k_m})^2 Q_{m-1}\big) = E\big(E[(1+\delta_{k_m})^2 Q_{m-1}/\mathcal{S}_{K \triangle K'}, \delta_{k_1}, ..., \delta_{k_{m-1}}]\big) \\ &= E\big(Q_{m-1} E[(1+\delta_{k_m})^2/\mathcal{S}_{K \triangle K'}, \delta_{k_1}, ..., \delta_{k_{m-1}}]\big) \\ &= E\big(Q_{m-1} E[1+\delta_{k_m}^2/\mathcal{S}_{K \triangle K'}, \delta_{k_1}, ..., \delta_{k_{m-1}}]\big). \end{aligned}$$

Since $|\delta_{k_m}| \leq u$, we have

$$E(Q_{m-1}) \leq E\big(Q_{m-1} E[1+\delta_{k_m}^2/\mathcal{S}_{K \triangle K'}, \delta_{k_1}, ..., \delta_{k_{m-1}}]\big) \leq E\big(Q_{m-1}(1+u^2)\big).$$

Thus, $1 \leq E\big(Q_m\big) \leq (1+u^2)^m$. Finally, by induction, the claim is proven

$$0 \leq E\big(Q_m\big) - 1 = \mathrm{Cov}(\psi_K, \psi_{K'}) \leq \gamma_m(u^2).$$

□

Under SR-nearness, Lemma 3.1 can now be used to derive a variance bounds for many algorithms, such as inner products, matrix-vector and matrix-matrix products, solutions of triangular systems, and the Horner algorithm. In the following, we chose the inner product and Horner algorithms as applications.

3.1 Inner product

Consider the inner product $s_n = y = a_1b_1 + \ldots + a_nb_n$, evaluated from left to right, i.e, $s_i = s_{i-1} + a_ib_i$, starting with $s_1 = a_1b_1$. In this paper, we address the sequential method which has a deterministic bound proportional to nu. However, the accumulator implementation of the inner product using a binary tree leads to a deterministic error bound in $O(\ln(n)u)$.

Let $\delta_0 = 0$, the computed $\widehat{s}_i$ satisfies $\widehat{s}_1 = a_1b_1(1+\delta_1)$ and

$$\widehat{s}_i = (\widehat{s}_{i-1} + a_ib_i(1+\delta_{2i-2}))(1+\delta_{2i-1}), \quad |\delta_{2i-2}|, |\delta_{2i-1}| \leq u,$$

for all $2 \leq i \leq n$, where δ_{2i-2} and δ_{2i-1} represent the rounding errors from the products and additions, respectively. We thus have

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

Theorem 3.1. *Under SR-nearness, the computed* $\widehat{y}$ *satisfies* $E(\widehat{y}) = y$ *and*

$$V(\widehat{y}) \leq y^2 \mathcal{K}_1^2 \gamma_n(u^2), \tag{3.1}$$

where $\mathcal{K}_1 = \frac{\sum_{i=1}^n |a_ib_i|}{|\sum_{i=1}^n a_ib_i|}$ *is the condition number using the 1-norm for the computed* $y = \sum_{i=1}^n a_ib_i$.

Proof. For all $1 \leq i \leq n$, we have

$$\widehat{y} = \sum_{i=1}^{n} a_ib_i(1+\delta_{2i-2}) \prod_{k=i}^{n}(1+\delta_{2k-1}) = \sum_{i=1}^{n} a_ib_i\psi_{K_i},$$

with $K_i = \{2i-2, 2i-1, 2i+1, \ldots, 2n-1\}$. Lemma 3.1 shows that $E(\psi_{K_i}) = 1$ for all $1 \leq i \leq n$, hence

$$E(\widehat{y}) = E\big(\sum_{i=1}^{n} a_ib_i\psi_{K_i}\big) = \sum_{i=1}^{n} a_ib_iE(\psi_{K_i}) = y.$$

For all $1 \le i < j \le n$, $K_j \cap K_i = \{2j-1, 2j+1, ..., 2n-1\}$ and $\mathrm{Card}(K_j \cap K_i) = n - j + 1$.

$$
\begin{aligned}
V(\widehat{y}) &= V\big(\sum_{i=1}^{n} a_i b_i \psi_{K_i}\big) \\
&\le \left(\sum_{i=1}^{n} |a_i b_i| \sqrt{V(\psi_{K_i})}\right)^2 && \text{since } \sigma(X+Y) \le \sigma(X) + \sigma(Y) \\
&\le \left(\sum_{i=1}^{n} |a_i b_i| \sqrt{\gamma_{n-i+1}(u^2)}\right)^2 && \text{by Lemma 3.1} \\
&\le \gamma_n(u^2) \big(\sum_{i=1}^{n} |a_i b_i|\big)^2 && \text{since } \gamma_{n-i+1}(u^2) \le \gamma_n(u^2) \\
&= y^2 \mathcal{K}_1^2 \gamma_n(u^2).
\end{aligned}
$$

□

Remark 3.1. *Because $E(\widehat{y}) = y$, under a normality assumption, the number of significant bits can be lower-bounded by*

$$
-\log_2\left(\frac{\sigma(\widehat{y})}{|E(\widehat{y})|}\right) \ge -\log_2\left(\mathcal{K}_1 \sqrt{\gamma_n(u^2)}\right) \approx -\log_2(\mathcal{K}_1) - \log_2(u) - \frac{1}{2}\log_2(n).
$$

3.2 Horner algorithm

Horner algorithm is an efficient way of evaluating polynomials. When performed in floating-point arithmetic, this algorithm may suffer from catastrophic cancellations and yield a computed value less accurate than expected.

Model 3.1. *Let $P(x) = \sum_{i=0}^{n} a_i x^i$, Horner rule consists in writing this polynomial as*

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

We define by induction the following sequence

Operation	*Floating-point arithmetic*	*Exact computation*
	$\widehat{r}_0 = a_n$	$r_0 = a_n$
$*$	$\widehat{r}_{2k-1} = \widehat{r}_{2k-2} x (1 + \delta_{2k-1})$	$r_{2k-1} = r_{2k-2} x$
$+$	$\widehat{r}_{2k} = (\widehat{r}_{2k-1} + a_{n-k})(1 + \delta_{2k})$	$r_{2k} = r_{2k-1} + a_{n-k}$
Output	$\widehat{r}_{2n} = \widehat{P}(x)$	$r_{2n} = P(x)$

for all $1 \le k \le n$, with δ_{2k-1} and δ_{2k} the rounding errors from the products and the additions respectively. Let $\delta_0 = 0$, we thus have

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

Theorem 3.2. *Using SR-nearness, the computed $\widehat{r}_{2n}$ satisfies $E(\widehat{r}_{2n}) = r_{2n}$ and*

$$
V(\widehat{r}_{2n}) \le r_{2n}^2 \mathcal{K}_1^2 \gamma_{2n}(u^2), \tag{3.2}
$$

where $\mathcal{K}_1 = \frac{\sum_{i=0}^{n} |a_i x^i|}{|\sum_{i=0}^{n} a_i x^i|}$ is the condition number using the 1-norm for the computed $P(x) = \sum_{i=0}^{n} a_i x^i$.

Proof. We have

$$
\widehat{r}_{2n} = \sum_{i=0}^{n} a_i x^i \prod_{k=2(n-i)}^{2n} (1 + \delta_k) = \sum_{i=0}^{n} a_i x^i \psi_{K_i},
$$

with $K_i = \{2(n-i), 2(n-i)+1, ..., 2n\}$ for all $0 \le i \le n$. Lemma 3.1 implies $E(\psi_{K_i}) = 1$, then $E(\widehat{r}_{2n}) =$

$E\big(\sum_{i=0}^{n} a_i x^i \psi_{K_i}\big) = \sum_{i=0}^{n} a_i x^i E(\psi_{K_i}) = r_{2n}$. Therefore, because $\delta_0 = 0$ we have

$$\begin{aligned}
V(\widehat{r}_{2n}) &= V\big(\sum_{i=0}^{n} a_i x^i \psi_{K_i}\big) \\
&\leq \left(\sum_{i=0}^{n} |a_i x^i| \sqrt{V(\psi_{K_i})}\right)^2 && \text{since } \sigma(X+Y) \leq \sigma(X) + \sigma(Y) \\
&\leq \left(\sum_{i=0}^{n} |a_i x^i| \sqrt{\gamma_{2i}(u^2)}\right)^2 && \text{by Lemma 3.1} \\
&\leq \gamma_{2n}(u^2) \big(\sum_{i=0}^{n} |a_i x^i|\big)^2 && \text{since } \gamma_{2i}(u^2) \leq \gamma_{2n}(u^2) \\
&= r_{2n}^2 \mathcal{K}_1^2 \gamma_{2n}(u^2).
\end{aligned}$$

□

Remark 3.2. *Because $E(\widehat{r}_{2n}) = r_{2n}$, under a normality assumption, the number of significant bits can be lower-bounded by*

$$\begin{aligned}
-\log_2\left(\frac{\sigma(\widehat{r}_{2n})}{|E(\widehat{r}_{2n})|}\right) &\geq -\log_2\left(\mathcal{K}_1 \sqrt{\gamma_{2n}(u^2)}\right) \\
&\approx -\log_2(\mathcal{K}_1) - \log_2(u) - \frac{1}{2}\log_2(2n).
\end{aligned}$$

4 Probabilistic bounds of the error for stochastic rounding

Based on the independence assumption, Higham and Mary [18] have shown that for the inner product, a probabilistic bound of the error proportional to $\sqrt{n \ln(n)}u$ can be achieved rather than the deterministic bound in $O(nu)$. With Connelly, they show in [8] that this bound always holds for SR-nearness due to mean independence of errors.

We start with the approach based on the Azuma-Hoeffding inequality and the martingale property (AH1 and AH2 methods in the following). In this context, firstly, we give a rigorous review of the previous results of the inner product forward error by Higham and Mary [8] and Ilse, Ipsen and Zhou [20]. Secondly, we extend these techniques to the Horner algorithm which also give a probabilistic bound proportional to $\sqrt{n}u$.

Then, we present a new approach based on Bienaymé–Chebyshev inequality and the previous variance estimations (BC method in the following), our bound is also in $O(\sqrt{n}u)$ and it is lower than the AH1 and AH2 bounds in several situations for both inner product and Horner algorithms.

4.1 Azuma-Hoeffding method

Let us recall the concept of a martingale and the Azuma-Hoeffding inequality for a martingale [24].

Definition 4.1. *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*
- *$\mathbb{E}[M_k / X_1, ..., X_{k-1}] = M_{k-1}$.*

Lemma 4.1. *(Azuma-Hoeffding inequality). Let $M_0, ..., M_n$ be a martingale with respect to a sequence $X_1, ..., X_n$. We assume that there exist $a_k < b_k$ such that $a_k \leq M_k - M_{k-1} \leq b_k$ for $k = 1 : n$. Then, for any $A > 0$*

$$\mathbb{P}(|M_n - M_0| \geq A) \leq 2\exp\left(-\frac{2A^2}{\sum_{k=1}^{n}(b_k - a_k)^2}\right).$$

In the particular case $a_k = -b_k$ and $\lambda = 2\exp\left(-\frac{A^2}{2\sum_{k=1}^{n} b_k^2}\right)$ we have

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

where $0 < \lambda < 1$.

4.1.1 Inner product

Under SR-nearness, the inner product $y = a^\top b$, where $a, b \in \mathbb{R}^n$ is defined as $\widehat{y} = \widehat{s}_n = \sum_{i=1}^n a_i b_i (1 + \delta_{2(i-1)}) \prod_{k=i}^n (1 + \delta_{2k-1})$. The worst case of the forward error of the computed $\widehat{y}$ is in $O(nu)$. Wilkinson [28, sec 1.33] had the intuition that the roundoff error accumulated in n operations is proportional to $\sqrt{n}u$ rather than nu. Based on the mean independence of errors established in Lemma 2.1, Higham and Mary [8] and Ilse, Ipsen and Zhou [20] have investigated this problem for SR-nearness. Both works build on the mean independence property of SR-nearness. This allows them to form a martingale, and then to apply the Azuma-Hoeffding concentration inequality. The difference between these two works is in the way they form the martingale. In [8, sec 3], the martingale is built using the errors accumulated in the whole process $\psi_{K_i} = (1 + \delta_{2(i-1)}) \prod_{k=i}^n (1 + \delta_{2k-1})$ for all $1 \le i \le n$. Azuma-Hoeffding inequality implies that $|\psi_{K_i}| \le \tilde{\gamma}_n(\lambda)$ with probability at least $1 - 2\exp\frac{-\lambda^2}{2}$, where $\tilde{\gamma}_n(\lambda) = \exp\frac{\lambda\sqrt{n}u + nu^2}{1-u} - 1$. This approach uses the inclusion-exclusion principle to generalize the bound to the summation which results in a pessimistic n in the probability. They prove

$$\frac{|\widehat{y} - y|}{|y|} \le \mathcal{K}_1 \tilde{\gamma}_n(\lambda),$$

with probability at least $1 - 2n\exp\frac{-\lambda^2}{2}$. The factor n in the probability disrupts the $\sqrt{n}u$ property. $\delta = 2n\exp\frac{-\lambda^2}{2}$ implies that $\lambda = \sqrt{2\ln(2n/\delta)}$ and

$$\frac{|\widehat{y} - y|}{|y|} \le \mathcal{K}_1 \tilde{\gamma}_n\left(\sqrt{2\ln(2n/\delta)}\right), \tag{AH1-IP}$$

with probability at least $1 - \delta$. When $nu \ll 1$, we have

$$\begin{aligned}
\tilde{\gamma}_n(\sqrt{2\ln(2n/\delta)}) &= \exp\frac{\sqrt{2n\ln(2n/\delta)}u + nu^2}{1-u} - 1 \\
&= u\sqrt{2n\ln(2n/\delta)} + O(u^2) \\
&= u\sqrt{2n\ln 2n - 2n\ln\delta} + O(u^2) = O(u\sqrt{n\ln n}).
\end{aligned}$$

On the other hand [20, sec 4] forms it by following step-by-step how the error accumulates in the recursive summation of the inner product. In particular, they distinguish between the multiplications and additions computed at each step and carefully monitor their mean independences. This approach leads to the following probabilistic bound

$$\frac{|\widehat{y} - y|}{|y|} \le \mathcal{K}_1 \sqrt{u\gamma_{2n}(u)}\sqrt{\ln(2/\delta)}, \tag{AH2-IP}$$

with probability at least $1 - \delta$. This technique avoids the inclusion-exclusion principle and, when $nu \ll 1$, it leads to

$$\sqrt{u\gamma_{2n}(u)}\sqrt{\ln(2/\delta)} = u\sqrt{2n\ln 2 - 2n\ln\delta} + O(u^2).$$

Note that when $nu \ll 1$, (AH1-IP) and (AH2-IP) differ only in the factor $\sqrt{\ln n}$ that appears in (AH1-IP) due to the use of the martingale property on each partial sum. All in all, (AH2-IP) is proportional to $u\sqrt{n}$, while (AH1-IP) is proportional to $u\sqrt{n\ln n}$. An analysis of the case $nu \gg 1$ will be presented in Section 5.

4.1.2 Horner algorithm

In the following, we derive a probabilistic bound for the computed $\widehat{P}(x)$ based on the previous method applied for the inner product in [20, sec 4].

With the notations defined in Model 3.1, let us denote $Z_i := \widehat{r}_i - r_i$ for all $1 \le i \le 2n$. The total forward error is $|Z_{2n}| = |\widehat{r}_{2n} - r_{2n}| = |\widehat{P}(x) - P(x)|$ and

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

Finally

$$\frac{|\widehat{P}(x) - P(x)|}{|P(x)|} \le \mathcal{K}_1 \gamma_{2n}(u). \tag{4.1}$$

The deterministic bound is proportional to nu. In the following, we prove a probabilistic bound in $O(\sqrt{n}u)$.

The partial sum forward errors satisfy

$$\begin{aligned}
Z_{2k-1} &= \widehat{r}_{2k-1} - r_{2k-1} = \widehat{r}_{2k-2}x(1 + \delta_{2k-1}) - r_{2k-2}x \\
&= xZ_{2k-2} + \widehat{r}_{2k-2}x\delta_{2k-1}, \\
Z_{2k} &= \widehat{r}_{2k} - r_{2k} = (\widehat{r}_{2k-1} + a_{n-k})(1 + \delta_{2k}) - r_{2k-1} - a_{n-k} \\
&= Z_{2k-1} + (\widehat{r}_{2k-1} + a_{n-k})\delta_{2k},
\end{aligned}$$

for all $1 \leq k \leq n$. The sequence $Z_1, ..., Z_{2n}$ does not form a martingale with respect to $\delta_1, ..., \delta_{2n}$ due to the multiplication in odd steps,

$$E[Z_{2k-1}/\delta_1, ..., \delta_{2k-2}] = xZ_{2k-2}.$$

In order to form a martingale and use the Azuma-Hoeffding inequality, we define the following variable change

$$Y_i = \frac{Z_i}{x^{\llbracket (i+1)/2 \rrbracket}},$$

where $\llbracket (i+1)/2 \rrbracket$ is the integer part of $(i+1)/2$, we thus have

$$\begin{cases} Y_{2k-1} &= Y_{2k-2} + \frac{1}{x^{k-1}} \widehat{r}_{2k-2} \delta_{2k-1}, \\ Y_{2k} &= Y_{2k-1} + \frac{1}{x^k} (\widehat{r}_{2k-1} + a_{n-k}) \delta_{2k}, \end{cases} \tag{4.2}$$

for all $1 \leq k \leq n$.

Theorem 4.1. *The sequence of random variables $Y_1, ..., Y_{2n}$ is a martingale with respect to $\delta_1, ..., \delta_{2n}$.*

Proof. We check that the three conditions of Definition 4.1 are satisfied. Throughout the proof, we note the set $\mathbb{F}_k = \{\delta_1, ..., \delta_k\}$.

- The recursion in Model 3.1 shows that Y_i is a function of $\delta_1, ..., \delta_i$ for all $1 \leq i \leq 2n$.
- $\mathbb{E}(|Y_i|)$ is finite because x and a_k are finite for all $n - i \leq k \leq n$ and $|\delta_j| \leq u$ for all $1 \leq j \leq i$.
- We prove that $\mathbb{E}[Y_i/\mathbb{F}_{i-1}] = Y_{i-1}$ by distinguishing the even and odd cases. Firstly, using the mean independence of $\delta_1, ...\delta_{2k-1}$ and Equation (4.2) we obtain

$$\begin{aligned} \mathbb{E}[Y_{2k-1}/\mathbb{F}_{2k-2}] &= \mathbb{E}[Y_{2k-2}/\mathbb{F}_{2k-2}] + \mathbb{E}[\frac{1}{x^{k-1}} \widehat{r}_{2k-2} \delta_{2k-1}/\mathbb{F}_{2k-2}] \\ &= Y_{2k-2} + \frac{1}{x^{k-1}} \widehat{r}_{2k-2} \mathbb{E}[\delta_{2k-1}/\mathbb{F}_{2k-2}] = Y_{2k-2}. \end{aligned}$$

Secondly, using the mean independence of $\delta_1, ...\delta_{2k}$ and Equation (4.2) we obtain

$$\begin{aligned} \mathbb{E}[Y_{2k}/\mathbb{F}_{2k-1}] &= \mathbb{E}[Y_{2k-1}/\mathbb{F}_{2k-1}] + \mathbb{E}[\frac{1}{x^k} (\widehat{r}_{2k-1} + a_{n-k}) \delta_{2k}/\mathbb{F}_{2k-1}] \\ &= Y_{2k-1} + \frac{1}{x^k} (\widehat{r}_{2k-1} + a_{n-k}) \mathbb{E}[\delta_{2k}/\mathbb{F}_{2k-1}] = Y_{2k-1}. \end{aligned}$$

□

Lemma 4.2. *The above martingale $Y_1, ..., Y_{2n}$ satisfies $|Y_i - Y_{i-1}| \leq C_i u$, for all $1 \leq i \leq 2n$, where*

$$\begin{cases} C_{2k-1} &= |a_n|(1+u)^{2k-2} + \sum_{j=1}^{k-1} |a_{n-j}||x|^{-j}(1+u)^{2(k-j)-1}, \\ C_{2k} &= |a_n|(1+u)^{2k-1} + \sum_{j=1}^{k} |a_{n-j}||x|^{-j}(1+u)^{2(k-j)}, \end{cases}$$

for all $1 \leq k \leq n$.

Proof. Note that $Y_0 = 0$, then $|Y_1 - Y_0| = |Y_1| = |a_n|$ and the equality holds for C_1. Using Equation (4.2)

$$|Y_{2k-1} - Y_{2k-2}| \leq \frac{1}{|x|^{k-1}} |\widehat{r}_{2k-2}| u.$$

Moreover

$$|\widehat{r}_{2k-2}| \leq |\widehat{r}_{2k-3}|(1+u) + |a_{n-k+1}|(1+u) \leq |\widehat{r}_{2k-4}||x|(1+u)^2 + |a_{n-k+1}|(1+u),$$

by induction we obtain

$$|\widehat{r}_{2k-2}| \leq |a_n||x|^k (1+u)^{2k-2} + \sum_{j=1}^{k-1} |a_{n-j}||x|^{k-j} (1+u)^{2(k-j)-1}.$$

This completes the proof for C_{2k-1} for all $1 \leq k \leq n$. A similar approach can be applied to prove the same result for C_{2k} for all $1 \leq k \leq n$. □

We now have all the tools to state and demonstrate the main result of this section:

Theorem 4.2. *Under SR-nearness, for all $0 < \lambda < 1$ and with probability at least $1 - \lambda$*

$$\frac{|\widehat{P}(x) - P(x)|}{|P(x)|} \leq \mathcal{K}_1 \sqrt{u \gamma_{4n}(u)} \sqrt{\ln(2/\lambda)}, \tag{4.3}$$

where $\mathcal{K}_1 = \frac{\sum_{i=0}^{n} |a_i x^i|}{|P(x)|}$ is the condition number of the polynomial evaluation and $\gamma_{4n}(u) = (1+u)^{4n} - 1$.

Proof. Recall that $|\widehat{r}_{2n} - r_{2n}| = |Z_{2n}| = |x^n||Y_{2n}|$. Therefore, $Y_1, ..., Y_{2n}$ is a martingale with respect to $\delta_1, ..., \delta_{2n}$ and Lemma 4.2 implies $|Y_i - Y_{i-1}| \leq C_i u$ for all $1 \leq i \leq 2n$. Using the Azuma-Hoeffding inequality yields

$$\mathbb{P}\left(|Y_{2n}| \leq u\sqrt{\sum_{i=1}^{2n} C_i^2}\sqrt{2\ln(2/\lambda)}\right) \geq 1 - \lambda,$$

it follows that

$$|Z_{2n}| \leq u\sqrt{\sum_{i=1}^{2n}(|x|^n C_i)^2}\sqrt{2\ln(2/\lambda)},$$

with probability at least $1 - \lambda$, where

$$\begin{aligned}
|x|^n C_{2k} &= |a_n||x|^n(1+u)^{2k-1} + \sum_{j=1}^{k}|a_{n-j}x^{n-j}|(1+u)^{2(k-j)} \\
&\leq (1+u)^{2k-1}\sum_{j=0}^{k}|a_{n-j}x^{n-j}| \leq (1+u)^{2k-1}\sum_{j=0}^{n}|a_j x^j|,
\end{aligned}$$

for all $1 \leq k \leq n$. Hence,

$$(|x|^n C_{2k})^2 \leq (1+u)^{2(2k-1)}\big(\sum_{j=0}^{n}|a_j x^j|\big)^2.$$

In a similar way

$$(|x|^n C_{2k-1})^2 \leq (1+u)^{2(2k-2)}\big(\sum_{j=0}^{n}|a_j x^j|\big)^2.$$

Thus

$$\begin{aligned}
\sum_{i=1}^{2n}(|x|^n C_i)^2 &\leq \big(\sum_{j=0}^{n}|a_j x^j|\big)^2 \sum_{i=0}^{2n-1}((1+u)^2)^i \\
&= \big(\sum_{j=0}^{n}|a_j x^j|\big)^2 \frac{((1+u)^2)^{2n} - 1}{(1+u)^2 - 1} = \big(\sum_{j=0}^{n}|a_j x^j|\big)^2 \frac{\gamma_{4n}(u)}{u^2 + 2u}.
\end{aligned}$$

As a result

$$|\widehat{P}(x) - P(x)| = |Z_{2n}| \leq \sum_{j=0}^{n}|a_j x^j|\sqrt{\frac{u\gamma_{4n}(u)}{2+u}}\sqrt{2\ln(2/\lambda)},$$

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

$$\frac{|\widehat{P}(x) - P(x)|}{|P(x)|} \leq \mathcal{K}_1\sqrt{u\gamma_{4n}(u)}\sqrt{\ln(2/\lambda)},$$

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

4.2 Bienaymé–Chebyshev method

Another way to obtain a probabilistic $O(\sqrt{n}u)$ bound is to use Bienaymé–Chebyshev inequality. This method requires only information on the variance. Moreover, we will see in Section 5 that for any probability λ there exist n such that this method introduces a tighter probabilistic bound than the Azuma-Hoeffding method.

Lemma 4.3. *(Bienaymé–Chebyshev inequality) Let X be a random variable with finite expected value and finite non-zero variance. For any real number $\alpha > 0$,*

$$\mathbb{P}\big(|X - E(X)| \leq \alpha\sqrt{V(X)}\big) \geq 1 - \frac{1}{\alpha^2}.$$

Regarding the two algorithms above, the computed $\widehat{y}$ satisfies $E(\widehat{y}) = y$, then

$$\mathbb{P}\left(|\widehat{y} - y| \leq \alpha\sqrt{V(\widehat{y})}\right) \geq 1 - \frac{1}{\alpha^2},$$

taking $\lambda = \frac{1}{\alpha^2}$ yields $|\widehat{y} - y| \leq \sqrt{V(\widehat{y})/\lambda}$ with probability at least $1 - \lambda$.

4.2.1 Inner product

From Theorem 3.1 we have

$$\frac{\sqrt{V(\widehat{y})/\lambda}}{|y|} \leq \mathcal{K}_1\sqrt{\gamma_n(u^2)/\lambda}.$$

Thus,

$$\frac{|\widehat{y}-y|}{|y|} \leq \frac{\sqrt{V(\widehat{y})/\lambda}}{|y|} \leq \mathcal{K}_1\sqrt{\gamma_n(u^2)/\lambda},$$

and

$$\mathbb{P}\left(\frac{|\widehat{y}-y|}{|y|} \leq \mathcal{K}_1\sqrt{\gamma_n(u^2)/\lambda}\right) \geq \mathbb{P}\left(\frac{|\widehat{y}-y|}{|y|} \leq \frac{\sqrt{V(\widehat{y})/\lambda}}{|y|}\right) \geq 1-\lambda. \tag{4.4}$$

4.2.2 Horner algorithm

From Theorem 3.2 we have

$$\frac{V(\widehat{P}(x))}{|P(x)|} \leq \mathcal{K}_1\sqrt{\gamma_{2n}(u^2)}.$$

The previous reasoning from Sub-section 4.2.1 leads to

$$\mathbb{P}\left(\frac{|\widehat{P}(x)-P(x)|}{|P(x)|} \leq \mathcal{K}_1\sqrt{\gamma_{2n}(u^2)/\lambda}\right) \geq 1-\lambda. \tag{4.5}$$

5 Bounds analysis

In the following, we compare the various bounds of the two previous algorithms and analyze which bound is the tightest depending on the precision in use, the target probability and the number of operations.

5.1 Inner product

In the beginning, let us recall all bounds obtained for the inner product $y = a^\top b$, where $a, b \in \mathbb{R}^n$

$$\frac{|\widehat{y}-y|}{|y|} \leq \mathcal{K}_1\gamma_n(u), \tag{Det-IP}$$

$$\frac{|\widehat{y}-y|}{|y|} \leq \mathcal{K}_1\tilde{\gamma}_n(\sqrt{2\ln(2n/\lambda)}) \quad \text{with probability at least } 1-\lambda, \tag{AH1-IP}$$

$$\frac{|\widehat{y}-y|}{|y|} \leq \mathcal{K}_1\sqrt{u\gamma_{2n}(u)}\sqrt{\ln(2/\lambda)} \quad \text{with probability at least } 1-\lambda, \tag{AH2-IP}$$

$$\frac{|\widehat{y}-y|}{|y|} \leq \mathcal{K}_1\sqrt{\gamma_n(u^2)}\sqrt{1/\lambda} \quad \text{with probability at least } 1-\lambda, \tag{BC-IP}$$

where $\tilde{\gamma}_n(\sqrt{2\ln(2n/\lambda)}) = \exp\left(\frac{\sqrt{2n\ln(2n/\lambda)}u+nu^2}{1-u}\right) - 1$.

All bounds have the same condition number $\mathcal{K}_1$, but differ in the others factor: $\gamma_n(u)$ for (Det-IP), $\tilde{\gamma}_n(\sqrt{2\ln(2n/\lambda)})$ for (AH1-IP), $\sqrt{u\gamma_{2n}(u)}\sqrt{\ln(2/\lambda)}$ for (AH2-IP), and $\sqrt{\gamma_n(u^2)}\sqrt{1/\lambda}$ for (BC-IP). For a constant λ, we investigate two cases: $nu \ll 1$ and $nu \gg 1$.

For n and u such that $nu \ll 1$ we have

$$\exp\frac{\sqrt{2n\ln(2n/\lambda)}u+nu^2}{1-u} - 1 = u\sqrt{2n\ln(2n/\lambda)} + O(u^2).$$

Moreover, [17, Lemma 3.1] implies

$$\gamma_n(u) \leq \frac{nu}{1-nu},$$

it follows that for $2nu < 1$,

$$\sqrt{u\gamma_{2n}(u)} \leq \sqrt{\frac{2nu^2}{1-2nu}} = u\sqrt{n}\sqrt{\frac{2}{1-2nu}},$$

and for $nu^2 < 1$

$$\sqrt{\gamma_n(u^2)} \leq \sqrt{\frac{nu^2}{1-nu^2}} = u\sqrt{n}\frac{1}{\sqrt{1-nu^2}}.$$

Interestingly, for the inner product, at any fixed probability, when $nu \ll 1$, (AH2-IP) and (BC-IP) bounds are proportional to $\sqrt{n}u$ unlike $\sqrt{n \ln n}u$ for the (AH1-IP) bound. Note that, the deterministic bound is in $O(nu)$.

For n and u such that $nu \gg 1$ and $nu^2 \ll 1$, we have

$$\begin{aligned}\exp \frac{\sqrt{2n\ln{(2n/\lambda)}}u + nu^2}{1-u} - 1 &\approx \exp \frac{\sqrt{2n\ln{(2n/\lambda)}}u + nu^2}{1-u} \\ &\approx \exp{(\sqrt{n\ln{(n)}}u)},\end{aligned}$$

then

$$\tilde{\gamma}_n(\sqrt{2\ln{(2n/\lambda)}}) \approx \exp{(\sqrt{n\ln{(n)}}u)}. \tag{5.1}$$

Furthermore

$$\sqrt{u\gamma_{2n}(u)} \approx \sqrt{u\exp{(2nu)}-1} \approx \sqrt{u}\exp{(nu)} \approx \exp{(nu)}, \tag{5.2}$$

and

$$\sqrt{\gamma_n(u^2)} \approx \sqrt{\exp{(nu^2)}-1} \approx \sqrt{n}u + O(u^2). \tag{5.3}$$

Therefore, (5.1), (5.2) and (5.3) show that (BC-IP) $\le$ (AH1-IP) $\le$ (AH2-IP) when $nu \gg 1$ and $nu^2 \ll 1$.

5.2 Horner algorithm

Let us recall all bounds obtained for the Horner algorithm

$$\frac{|\widehat{P}(x) - P(x)|}{|P(x)|} \le \mathcal{K}_1\gamma_{2n}(u), \tag{Det-H}$$

$$\frac{|\widehat{P}(x) - P(x)|}{|P(x)|} \le \mathcal{K}_1\sqrt{u\gamma_{4n}(u)}\sqrt{\ln\frac{2}{\lambda}} \quad \text{with probability} \ge 1-\lambda, \tag{AH-H}$$

$$\frac{|\widehat{P}(x) - P(x)|}{|P(x)|} \le \mathcal{K}_1\sqrt{\gamma_{2n}(u^2)}\sqrt{\frac{1}{\lambda}} \quad \text{with probability} \ge 1-\lambda. \tag{BC-H}$$

A similar reasoning to Section 5.1 shows that the probabilistic bounds for the Horner algorithm forward error are in $O(\sqrt{n}u)$ versus $O(nu)$ for the deterministic bound. With Horner method the degree of the polynomial, n, is seldom very large in practice.

In conclusion, these probabilistic approaches show that the roundoff error accumulated in n operations is proportional to $\sqrt{n}u$ rather than nu. In the next section, we analyze these two probabilistic methods.

5.3 Bienaymé–Chebyshev vs Azuma-Hoeffding

In the following, we compare the three probabilistic bounds (AH1-IP), (AH2-IP) and (BC-IP) on the inner product forward error (similar reasoning can be applied to the Horner algorithm with the same result). When $nu \ll 1$, at any fixed probability, (AH2-IP) and (BC-IP) are proportional $O(\sqrt{n}u)$. First, we focus on this case. The two probabilistic bounds have the same condition number $\mathcal{K}_1$. Thus, it is enough to compare $\sqrt{\frac{u}{2}\gamma_{2n}(u)}\sqrt{2\ln(2/\lambda)}$ and $\sqrt{\gamma_n(u^2)}\sqrt{1/\lambda}$. These two bounds depend on n and λ. Firstly, using the binomial theorem we have

$$\begin{aligned}\frac{u}{2}\gamma_{2n}(u) - \gamma_n(u^2) &= \frac{u}{2}\sum_{k=1}^{n}\binom{n}{k}(u^2+2u)^k - \sum_{k=1}^{n}\binom{n}{k}(u^2)^k \\ &\ge \sum_{k=1}^{2}\binom{n}{k}\left[\frac{u}{2}(u^2+2u)^k - (u^2)^k\right] \ge n(n-\frac{1}{2})u^3.\end{aligned}$$

We can conclude that

$$\sqrt{\gamma_n(u^2)} \le \sqrt{\frac{u}{2}\gamma_{2n}(u)} \text{ for all } n \ge 1. \tag{5.4}$$

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

Probability	u	Precision format	$n \gtrsim$
$1-\lambda = 0.95$	2^{-7}	bfloat16	110
	2^{-10}	binary16	890
	2^{-23}	binary32	7.3 $e06$
	2^{-52}	binary64	3.9 $e15$
$1-\lambda = 0.99$	2^{-7}	bfloat16	220
	2^{-10}	binary16	1810
	2^{-23}	binary32	1.48 $e07$
	2^{-52}	binary64	7.9 $e15$

Table 1: The smallest n such that BC method gives a tighter probabilistic bound than AH2 method for the inner product.

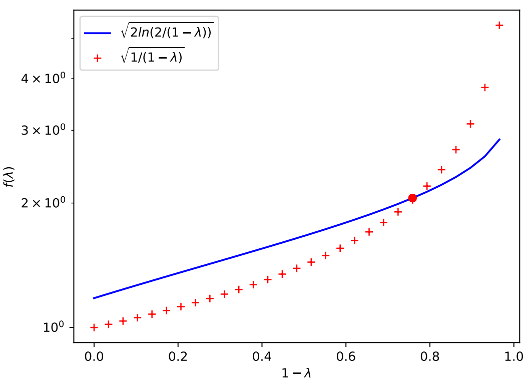

Figure 3: Illustration of $\sqrt{1/\lambda}$ and $\sqrt{2\ln(2/\lambda)}$ behaviour for all $\lambda \in]0;1[$.

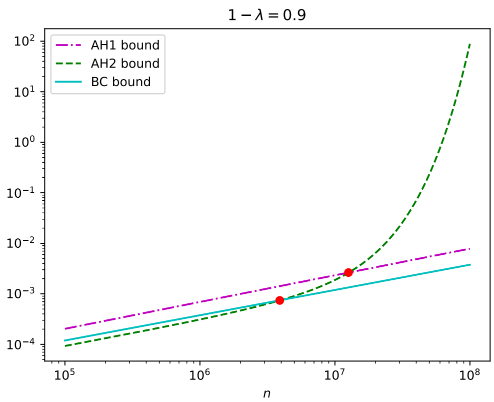

Figure 4: AH1, AH2 and BC bounds with probability 0.9 and $u = 2^{-23}$ for the inner product.

Figure 3 and the inequality (5.4) show that whatever the problem size n and for a probability at most ≈ 0.758, the BC method gives a tighter probabilistic bound than the AH2 method.

Figure 4 confirms the discussion of Section 5.1 and it shows that with a probability 0.9, when $nu \gg 1$, AH2 bound grows rapidly compared to AH1 and BC bounds. Regarding BC bound, the variance calculation and the mean independence allow to bound the error terms $(1+\delta)^2$ by $(1+u^2)$ and avoid all δ terms of degree one because $E(\delta) = 0$. In contrast, the AH1 and AH2 methods requires bounded increments leading to terms $(1+u)^2$. As n increases, the advantage of Azuma-Hoeffding inequality for a probability near 1 becomes negligible.

For all asymptotic comparisons between the bounds in this paper, we have chosen to work with $u \to 0$, $n \to \infty$ and fixed probability λ, which we think adapted to many if not most current practical use cases. A situation with $\lambda \to 0$ and fixed n gives the advantage to the Azuma-Hoeffding bounds over the Bienaymé-Chebyshev one.

The Table 1 illustrates how BC is tighter than AH2 when n grows. The n threshold above which BC is preferable to AH2 bound depends on the format precision. The lower the precision, the lower the threshold becomes. Using SR in low precision is of high interest in the areas of machine learning [15], PDEs [10] and ODEs [19], motivating the use of our improved BC method.

6 Numerical experiments

This section presents numerical experiments that support and complete the theory presented previously. The various bounds are compared on two numerical applications: the inner product and the evaluation of the Chebyshev polynomial.

We show that the probabilistic bounds are tighter than the deterministic bound and faithfully capture the behavior of SR-nearness forward error. For inner product of large vectors, we show that BC bound is smaller than AH1 and AH2 bounds. All SR computations are repeated 30 times with verificarlo [13]; we plot all samples and the forward error of the average of the 30 SR instances.

6.1 Horner algorithm

We use Horner's method to evaluate the polynomial $P(x) = T_N(x) = \sum_{i=0}^{\llbracket \frac{N}{2} \rrbracket} a_i (x^2)^i$ where T_N is the Chebyshev polynomial of even degree $N = 2n$. The previous error bounds, (Det-H), (AH-H), and (BC-H) apply to this computation.

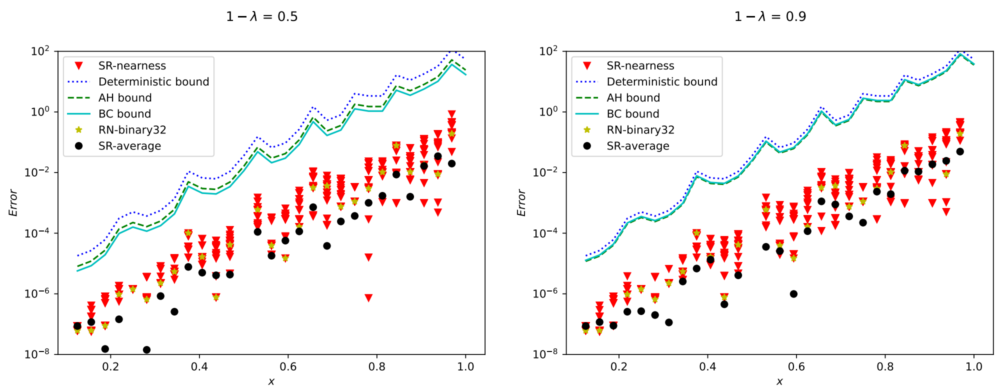

Figure 5: Probabilistic error bounds with probability $1 - \lambda = 0.5$ (left) and $1 - \lambda = 0.9$ (right) vs deterministic bound for the Horner's evaluation of $T_{20}(x)$ and $u = 2^{-23}$. Triangles mark 30 instances of the SR-nearness relative errors
evaluation in binary32 precision, a circle marks the relative errors of the 30 instances average, and a star represents the IEEE RN-binary32 value.

Chebyshev polynomial is ill-conditioned near 1 as shown in Figure 5, which evaluates $T_{20}(x)$ for $x \in [\frac{8}{64}; 1]$. Due to catastrophic cancellations among the polynomial terms, the condition number increases from 10^0 to 10^7 in the chosen x interval, resulting in an increasing numerical error for both RN-binary32 and SR-nearness computations.

The left plot confirms that the Bienaymé–Chebyshev bound (BC-H) is more accurate than the Azuma-Hoeffding bound (AH-H) for probability $1 - \lambda = 0.5$. With a higher probability $1 - \lambda = 0.9$ (right plot), since $N = 20$ and $u = 2^{-23}$ Azuma-Hoeffding bound (AH-H) is tighter, as predicted in Figure 4. Both probabilistic bounds are tighter than the deterministic bound. For $N = 20$, there is no significant difference between SR-nearness and RN-binary32. However, as expected, the average of the SR-nearness computations is more precise than the nearest round evaluation for almost all values of x.

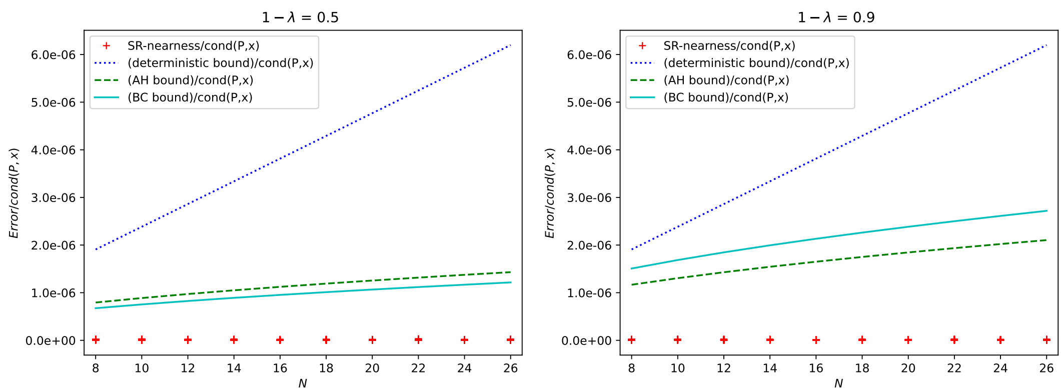

Figure 6: Normalized forward error (error$/cond(P,x)$) with probability $1 - \lambda = 0.5$ (left) and $1 - \lambda = 0.9$ (right) for Horner's evaluation of $T_N(24/26)$.

In Figure 6, the three previous bounds and the forward error are normalized by the condition number $cond(P,x)$. The evaluation in $x = 24/26 \approx 0.923$ is plotted for various polynomial degrees N. As expected, when N increases, the deterministic bound grows faster than the probabilistic bounds. The right plot shows that Azuma-Hoeffding bound is tighter for a high probability and a small n. Overall, Chebyshev polynomial numerical experiment illustrates the advantage of the probabilistic error bounds over the deterministic error bound. However, for most of the evaluations in this experiment, RN-binary32 is more accurate than one instance of

SR-nearness. This result is unsurprising because the degree n is small. To illustrate the behavior of these errors with a large n, we now turn to the inner product.

6.2 Inner product

To showcase the advantage of using BC method for large n, we present a numerical application of the inner product for vectors with positive floating-points chosen uniformly at random between 0 and 1.

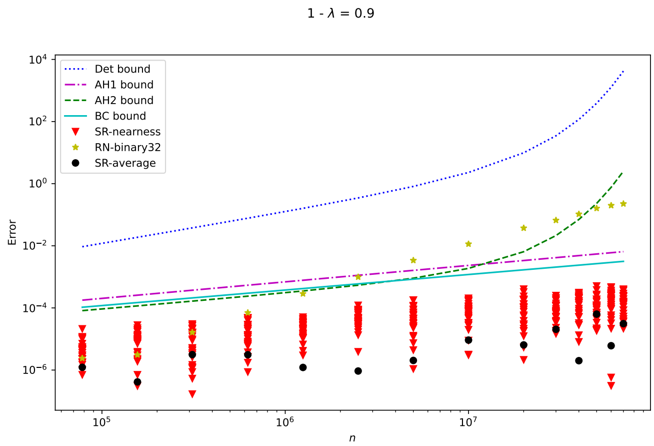

Figure 7: Probabilistic bounds with probability $1 - \lambda = 0.9$ vs deterministic bound of the computed forward errors of the inner product of with $u = 2^{-23}$.

For small n, AH1, AH2, and BC bounds are comparable with a slight advantage for (AH2-IP). However, as shown in Section 5.1, when $nu \gg 1$, the AH2 bound grows exponentially faster than AH1 and BC bounds. Asymptotically, the AH1 and BC bounds are therefore much tighter.

Interestingly, when n increases, a single instance of SR-nearness in binary32 precision is more accurate than RN-binary32. This is because the summation terms are chosen uniformly between 0 and 1. The terms closest to zero are absorbed. With RN-binary32 the absorption errors are biased and will add up, while SR avoids stagnation and mitigates absorption errors. If we choose the terms in $[-1; 1]$, SR and RN-binary32 have the same behavior. In this case, the absorption errors for RN-binary32 compensate because positive and negative errors are uniformly distributed. If we choose the terms in $[1/2; 1]$, no absorption occurs for $n < 2^{23}$, and on this domain, SR and RN-binary32 behave similarly.

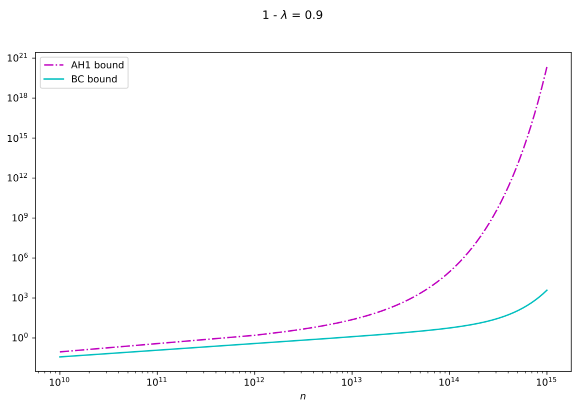

Figure 8: AH1 bound vs BC bound with probability $1 - \lambda = 0.9$ and $u = 2^{-23}$ for the inner product of with.

Figure 8 illustrates the advantage of using (BC-IP) and shows that for a large $n \geq 10^{13}$ and $u = 2^{-23}$, the AH1 bound increases faster than the BC bound.

For large vectors, using stochastic rounding instead of the default round to nearest improves the computation accuracy of the inner product. However, this experiment raises concerns regarding the use of SR as a model to estimate RN rounding errors [13, 26], in particular for a large number of operations. Further studies are required to assess precisely the limits of this model and possibly give criteria to detect them.

7 Conclusion

For a wide field of applications, SR results in a smaller accumulated error, for example by avoiding stagnation effects. Moreover, SR errors satisfy the mean independence property allowing to derive tight probabilistic error bounds from either our variance bound or the martingale property.

For an inner product $y = a^\top b$, Sub-section 4.1 compares the benefits of constructing the martingale from the recursive summation of the inner product [20] versus the construction from the errors accumulated in the whole process at each product $a_i b_i$ [8]. In particular, with a fixed probability, the construction in [20] gives a $O(\sqrt{n}u)$ probabilistic bound, tighter than the $O\big(u\sqrt{n \ln(n)}\big)$ bound in [8] when $nu \ll 1$. Nevertheless, when $nu \gg 1$, Figure 7 shows that the (AH2-IP) bound increases faster than (AH1-IP) bound.

An extension of the method in [20] to the Horner algorithm is presented. Unlike the inner product, Horner algorithm does not explicitly satisfy the martingale property on the partial sums requiring a change of variable before once can use the Azuma-Hoeffding inequality.

Lemma 3.1 is a variance bound for the family of algorithms whose error can be written as a product of error terms of the form $1 + \delta$. Based on the Bienaymé–Chebyshev inequality, a new method is proposed to obtain probabilistic error bounds. This method allows to get tighter probabilistic error bound in various situations, such as computations with a large n. We demonstrate the strength of this new approach on two algorithms: the inner product which has been previously studied, and Horner polynomial evaluation, for which no SR results were known beforehand.

The scripts for reproducing the numerical experiments in this paper are published in the repository `https://github.com/verificarlo/sr-variance-bounds/`.

Acknowledgment

This research was supported by the French National Agency for Research (ANR) via the InterFLOP project (No. ANR-20-CE46-0009).

References

[1] *Graphcore limited. 2021c ai-float™- mixed precision arithmetic for AI: A hardware perspective. version latest: Aug 25, 2021*, https://docs.graphcore.ai/projects/ai-float-white-paper/en/latest/index.html.

[2] *Loh GH. 2019 Stochastic rounding logic. Patent 20 190 294 412A1. Status: Active.*

[3] *IEEE standard for floating-point arithmetic*, IEEE Std 754-2019 (Revision of IEEE 754-2008), (2019), pp. 1–84, https://doi.org/10.1109/IEEESTD.2019.8766229.

[4] J. M. Alben, P. Micikevicius, H. Wu, and M. Y. Siu, *Stochastic rounding of numerical values*, June 16 2020. US Patent 10,684,824.

[5] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*, OUP Oxford, 2013, https://books.google.fr/books?id=koNqWRluhP0C.

[6] J. D. Bradbury, S. R. Carlough, B. R. Prasky, and E. M. Schwarz, *Reproducible stochastic rounding for out of order processors*, Sept. 25 2018. US Patent 10,083,008.

[7] J. D. Bradbury, S. R. Carlough, B. R. Prasky, and E. M. Schwarz, *Stochastic rounding floating-point multiply instruction using entropy from a register*, Oct. 15 2019. US Patent 10,445,066.

[8] M. P. Connolly, N. J. Higham, and T. Mary, *Stochastic rounding and its probabilistic backward error analysis*, SIAM Journal on Scientific Computing, 43 (2021), pp. A566–A585.

[9] M. Croci, M. Fasi, N. J. Higham, T. Mary, and M. Mikaitis, *Stochastic rounding: Implementation, error analysis, and applications*, (2021).

[10] M. Croci and M. Giles, *Effects of round-to-nearest and stochastic rounding in the numerical solution of the heat equation in low precision*, IMA Journal of Numerical Analysis, (2022).

[11] D. Dacunha-Castelle, D. McHale, and M. Duflo, *Probability and Statistics: Volume II*, no. v. 2, Springer New York, 2012, https://books.google.fr/books?id=KNXTBwAAQBAJ.

[12] M. Davies, N. Srinivasa, T.-H. Lin, G. Chinya, Y. Cao, S. H. Choday, G. Dimou, P. Joshi, N. Imam, S. Jain, et al., *Loihi: A neuromorphic manycore processor with on-chip learning*, IEEE Micro, 38 (2018), pp. 82–99.

[13] C. Denis, P. de Oliveira Castro, and E. Petit, *Verificarlo: Checking floating point accuracy through Monte Carlo arithmetic*, in 23nd IEEE Symposium on Computer Arithmetic, ARITH 2016, Silicon Valley, CA, USA, July 10-13, 2016, 2016, pp. 55–62, https://doi.org/10.1109/ARITH.2016.31, http://dx.doi.org/10.1109/ARITH.2016.31.

[14] F. Févotte and B. Lathuiliere, *Verrou: a cestac evaluation without recompilation*, SCAN 2016, (2016), p. 47.

[15] S. Gupta, A. Agrawal, K. Gopalakrishnan, and P. Narayanan, *Deep learning with limited numerical precision*, in International conference on machine learning, PMLR, 2015, pp. 1737–1746.

[16] G. G. Henry and D. R. Reed, *Processor with memory array operable as either cache memory or neural network unit memory*, May 26 2020. US Patent 10,664,751.

[17] N. J. Higham, *Accuracy and stability of numerical algorithms*, SIAM, 2002.

[18] N. J. Higham and T. Mary, *A new approach to probabilistic rounding error analysis*, SIAM Journal on Scientific Computing, 41 (2019), pp. A2815–A2835, `https://doi.org/10.1137/18M1226312`.

[19] M. Hopkins, M. Mikaitis, D. R. Lester, and S. Furber, *Stochastic rounding and reduced-precision fixed-point arithmetic for solving neural ordinary differential equations*, Philosophical Transactions of the Royal Society A, 378 (2020), p. 20190052.

[20] I. C. F. Ipsen and H. Zhou, *Probabilistic error analysis for inner products*, SIAM Journal on Matrix Analysis and Applications, 41 (2020), pp. 1726–1741, `https://doi.org/10.1137/19M1270434`.

[21] F. Jézéquel and J.-M. Chesneaux, *Cadna: a library for estimating round-off error propagation*, Computer Physics Communications, 178 (2008), pp. 933–955.

[22] O. A. Kanter and I. Bar, *Apparatus and methods for hardware-efficient unbiased rounding*, Mar. 3 2015. US Patent 8,972,472.

[23] S. Lifsches, *In-memory stochastic rounder*, Oct. 13 2020. US Patent 10,803,141.

[24] M. Mitzenmacher and E. Upfal, *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*, Cambridge University Press, 2005, `https://doi.org/10.1017/CBO9780511813603`.

[25] D. S. Parker, *Monte Carlo Arithmetic: exploiting randomness in floating-point arithmetic*, Technical Report CSD-970002, University of California (Los Angeles). Computer Science Department, 1997.

[26] D. Sohier, P. D. O. Castro, F. Févotte, B. Lathuilière, E. Petit, and O. Jamond, *Confidence intervals for stochastic arithmetic*, ACM Transactions on Mathematical Software (TOMS), 47 (2021), pp. 1–33.

[27] J. von Neumann and H. H. Goldstine, *Numerical inverting of matrices of high order*, Bulletin of the American Mathematical Society, 53 (1947), pp. 1021–1099.

[28] J. H. Wilkinson, *Rounding errors in algebraic processes*, Courier Corporation, 1994.