BOUNDS ON NON-LINEAR ERRORS FOR VARIANCE COMPUTATION WITH STOCHASTIC ROUNDING

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

BOUNDS ON NON-LINEAR ERRORS FOR VARIANCE COMPUTATION WITH STOCHASTIC ROUNDING

E-M. EL ARAR,*, D. SOHIER†, P. DE OLIVEIRA CASTRO†, AND E. PETIT†

Abstract. The main objective of this work is to investigate non-linear errors and pairwise summation using stochastic rounding (SR) in variance computation algorithms. We estimate the forward error of computations under SR through two methods: the first is based on a bound of the variance and the Bienaymé–Chebyshev inequality, while the second is based on martingales and the Azuma-Hoeffding inequality. The study shows that for pairwise summation, using SR results in a probabilistic bound of the forward error proportional to $\sqrt{\log(n)}u$ rather than the deterministic bound in $O(\log(n)u)$ when using the default rounding mode. We examine two algorithms that compute the variance, called "textbook" and "two-pass", which both exhibit non-linear errors. Using the two methods mentioned above, we show that these algorithms' forward errors have probabilistic bounds under SR in $O(\sqrt{n}u)$ instead of nu for the deterministic bounds. We show that this advantage holds using pairwise summation for both textbook and two-pass, with probabilistic bounds of the forward error proportional to $\sqrt{\log(n)}u$.

Key words. Stochastic rounding, Floating-point arithmetic, Variance computation, Non-linear error, Doob–Meyer decomposition, Pairwise summation.

MSC codes. 65G50, 65C99, 65Y04, 62-08

1. Introduction.

Stochastic Rounding (SR) mode [5] is a probabilistic rounding mode: an inexact computation is rounded to the next smaller or larger floating-point number with probability depending on the distances to those numbers. For several algorithms, such as the inner product [4, 8, 13] and Horner's rule [8, 9], SR is unbiased and provides tighter probabilistic bounds of the forward error compared to the deterministic bounds obtained with round-to-nearest (RN) [1]. In practice, SR shows higher accuracy than RN for some applications and datasets [8], particularly in low-precision formats such as bfloat-16. Additionally, SR avoids numerical stagnation [4] in different applications such as neural networks [10], ODEs, and PDEs [14].

Previous theoretical studies of SR error bounds have only considered algorithms in which the numerical error is a linear function of each operation rounding error. Two main methods have been proposed to bound the forward error of linear error algorithms such as summation or inner product computation. The first, referred to as the BC method in the following, computes the variance of the SR computation and applies Bienaymé–Chebyshev inequality to establish a probabilistic error bound [8]. The second, called AH method in the following, is based on martingales and Azuma-Hoeffding inequality [4]. The two methods are complementary, and each has advantages depending on the size of the problem and the target probabilistic analysis.

Hallman and Ipsen [11] have studied pairwise summation in the context of SR, showing that the forward error for a sum of n values has a probabilistic bound in $O(\sqrt{\log(n)}u)$ instead $O(\log(n)u)$ for RN. In this paper, we propose a more straightforward method that improves Hallman and Ipsen pairwise summation error bound [11].

In 1983, Chan, Golub, and LeVeque proved deterministic error bounds [3] for different algorithms computing the variance of a sample of n data points. These algorithms have non-linear errors due to the presence of squaring in the computation. In

*Université Paris-Saclay, UVSQ, Li-PaRAD, Saint-Quentin en Yvelines, France (el-mehdi.el-arar@uvsq.fr, devan.sohier@uvsq.fr, pablo.oliveira@uvsq.fr).

†Intel Corp (eric.petit@intel.com).

this paper, we prove SR forward error bounds for the "textbook" and "two-pass" algorithms with recursive and pairwise summation studied by Chan, Golub, and LeVeque. To the best of our knowledge, this is the first paper theoretically studying non-linear problems with SR. We extend previous BC and AH methods to the non-linear variance computation by carefully separating the error terms.

We first introduce some floating point background and the stochastic rounding mode SR-nearness in Section 2, and recall its main properties that we will use throughout the rest of the paper.

We analyze the error of pairwise summation under SR-nearness in Section 3, using two methods, AH, and BC. We provide probabilistic bounds for the pairwise summation forward error under SR using two methods, the BC and AH methods. Our AH pairwise bound is simpler and at least as tight as the probabilistic bound proposed in [11].

We then move to the analysis of variance computations, which, unlike summations, present non-linear errors. This, in particular, materializes in the existence of a bias, which we study in Section 4. We prove that both textbook and two-pass algorithms are biased and that their biases are equal at order 1 but of opposite signs.

In Section 5, we show that the deterministic bounds of Chan, Golub and LeVeque [2] extend to SR computations by replacing the n in the bounds by $\sqrt{n}$, and introducing a parameter λ representing the probability that the bound does not hold. We do it with both BC and AH methods, leading to bounds behaving better when $n \to \infty$ or $\lambda \to 0$ respectively, and propose an extension DM of the AH method based on a Doob-Meyer decomposition, which allows to better account for the bias and provides a new tool for SR analysis of non-linear errors.

We then prove that using pairwise summation in variance computations gives bounds in $\sqrt{\log(n)}$ in Section 6. We finally compare the obtained bounds by algorithm (textbook or two-pass) and method (deterministic, BC, AH, DM), and discuss the advantages of each in different situations in Section 7.

2. Notations and definitions.

2.1. Notations. In this paper, for an integer n and a vector $x \in \mathbb{R}^n$, we denote by

- $\|x\|_1 = \sum_{i=1}^n |x_i|$ and $\|x\|_2 = \left(\sum_{i=1}^n |x_i|^2\right)^{\frac{1}{2}}$.
- $s = \sum_{i=1}^n x_i$ and $m = \frac{1}{n}\sum_{i=1}^n x_i = \frac{1}{n}s$.
- $\gamma_n(u) = (1+u)^n - 1$.
- $\log(n)$ the smallest integer greater than $\log_2(n)$.

We adopt the same notations as used in [3]. In the following, the textbook algorithm computes the variance using the formula $y = \sum_{i=1}^n x_i^2 - \frac{1}{n}s^2$, while the two-pass algorithm computes the variance using the formula $z = \sum_{i=1}^n (x_i - m)^2$. We do not study the situation with $y = z = 0$, in which the relative error is undefined. The statistical variance can be obtained by multiplying y and z by $\frac{1}{n-1}$. Computing y and z exactly results in $y = z$. However, rounding errors disturb the numerical computations and the obtained results $\widehat{y}$ and $\widehat{z}$ are not equal.

The condition number using the 2-norm for the variance computation is defined in [3] as $\mathcal{K}_2 = \frac{\|x\|_2}{\sqrt{y}}$. We define the condition number using the 1-norm by $\mathcal{K}_1 = \frac{\|x\|_1}{\sqrt{ny}}$. Using the Cauchy-Schwarz inequality, $\mathcal{K}_1 \leq \mathcal{K}_2$; $\mathcal{K}_1$ can be lower than 1 (for instance, consider $n = 4$ and $x_1 = 1/2$, $x_2 = 1/4$, $x_3 = -x_1$ and $x_4 = -x_2$).

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 $\mathbb{E}[X/Y]$.

LEMMA 2.1. *Let X and Y two random variables, $a, b \in \mathbb{R}_+^*$, and $\lambda, \mu \in]0;1[$ such that: $\mathbb{P}(|X| \leq a) \geq 1-\lambda$ and $\mathbb{P}(|Y| \leq b) \geq 1-\mu$. Then*

- $\mathbb{P}(|XY| \leq ab) \geq 1-(\lambda+\mu)$,
- $\mathbb{P}(|X|+|Y| \leq a+b) \geq 1-(\lambda+\mu)$.

Proof.

$$\begin{aligned}\mathbb{P}(|X|\,|Y| \leq ab) &\geq \mathbb{P}(\{|X| \leq a\} \cap \{|Y| \leq b\}) \\ &= \mathbb{P}(|X| \leq a) + \mathbb{P}(|Y| \leq b) - \mathbb{P}(\{|X| \leq a\} \cup \{|Y| \leq b\}) \\ &\geq 1-\lambda+1-\mu-1 = 1-(\lambda+\mu).\end{aligned}$$

The proof of the second item uses the first point and the following property $\log(ab) = \log(a)+\log(b)$. ∎

2.2. Floating-point background. For a given basis β and a working precision p, a floating-point number is a real x such that $x = m \times \beta^{e-p}$, where e is the exponent and m is an integer (the significand) such that $\beta^{p-1} \leq |m| < \beta^p$. In this paper, we don't take into account special floating-point values such as underflow, overflow, denormals, and NaNs. Detailed information on the floating-point format most generally in use in current computer systems is defined in the IEEE-754 standard [1].

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

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

by definition, $\lfloor x \rfloor \leq x \leq \lceil x \rceil$, with equality 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: $|\delta| \leq \beta^{1-p}$. In the following, we use the same notation as [4, 13] $u = \beta^{1-p}$. IEEE-754 mode RN (round to nearest, ties to even) has the stronger property that $|\delta| \leq \frac{1}{2}\beta^{1-p} = \frac{1}{2}u$. 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 $\mathrm{op} \in \{+, -, *, /\}$. Moreover, for IEEE-754 rounding modes [1] and stochastic rounding [4] the error in one operation is bounded:

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

specifically for RN we have $|\delta| \leq \frac{1}{2}u$.

In this paper, we investigate asymptotic results for a problem of size n and precision u; $nu \ll 1$ means $n \to \infty$, $u \to 0$ and $nu \to 0$.

2.3. Stochastic rounding. Throughout this paper, $\widehat{x} = \mathrm{fl}(x)$ 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:

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

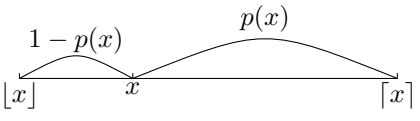

Fig. 1: **SR-nearness**.

where $p(x) = \frac{x - \lfloor x \rfloor}{\lceil x \rceil - \lfloor x \rfloor}$. The rounding SR-nearness mode is unbiased

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

In general, under SR-nearness, the error terms in algorithms appear as a sequence of random variables. The following lemma has been proven in [4, lem 5.2] and shows that this sequence is mean independent when considering operations satisfying the standard model (ie. the result is calculated exactly on the basis of its input, and then rounded according to SR).

LEMMA 2.2. *Consider a sequence of elementary operations* $c_k \leftarrow a_k \,\mathrm{op}_k\, b_k$ *for* $k \geq 1$*, with* op_k *satisfying the standard model and* δ_k *the error of the* k^{th} *operation, that is to say,* $\widehat{c}_k = (\widehat{a}_k \,\mathrm{op}_k\, \widehat{b}_k)(1 + \delta_k)$*. The* δ_k *are random variables with mean zero and* $(\delta_1, \delta_2, \ldots)$ *is mean independent, i.e.,* $\forall k \geq 2, \mathbb{E}[\delta_k \mid \delta_1, \ldots, \delta_{k-1}] = \mathbb{E}(\delta_k)$*.*

3. Pairwise summation. It is known that the accumulator implementation of a sum of n numbers $s = \sum_{i=1}^{n} x_i$ using a binary tree leads to a deterministic error bound in $O(\log(n)u)$. In this section, we investigate the forward error made by the pairwise summation under SR-nearness.

For the AH method, we construct a martingale straight from the tree levels and then use Azuma-Hoeffding inequality. This technique has the advantage of building a martingale from the entire tree. For the BC method, we use [8, lem 3.1] and Bienaymé–Chebyshev inequality. Both methods show $O(\sqrt{\log(n)}u)$ probabilistic bounds on the forward error. These bounds are simpler and more intuitive than the bounds in [11].

Considering h the height of the summation tree, if $2^{h-1} < n < 2^h$, we set the $2^h - n$ absent inputs to zero. Without loss of generality, let us then assume that $n = 2^h$. Denote $S_i^0 = x_i$ and $S_i^k = S_{2i-1}^{k-1} + S_{2i}^{k-1}$ for all $1 \leq i \leq 2^{h-k}$ and $1 \leq k \leq h$. We have

$$S_l^k = \sum_{i=(l-1)2^k+1}^{l2^k} x_i \quad \text{and} \quad S_1^h = \sum_{i=1}^{2^h} x_i = s.$$

Let $\widehat{S}_i^0 = S_i^0$ and $\widehat{S}_i^k = (\widehat{S}_{2i-1}^{k-1} + \widehat{S}_{2i}^{k-1})(1 + \delta_i^k)$ for all $1 \leq i \leq 2^{h-k}$ and $1 \leq k \leq h$. We have $\widehat{S}_l^k = \sum_{i=(l-1)2^k+1}^{l2^k} x_i \prod_{j=1}^{k} (1 + \delta^j_{\lceil \frac{i}{2^j} \rceil})$.

In particular

$$\widehat{S}_1^h = \sum_{i=1}^{2^h} x_i \prod_{j=1}^{h} (1 + \delta^j_{\lceil \frac{i}{2^j} \rceil}) = \sum_{i=1}^{2^h} x_i \psi_i \text{ with } \psi_i = \prod_{j=1}^{h} (1 + \delta^j_{\lceil \frac{i}{2^j} \rceil}). \tag{3.1}$$

As mentioned in Section 2.2, we compare the asymptotic behavior of the forward error bounds. El Arar et al [8] have introduced a new approach based on a bound of

the variance and Bienaymé–Chebyshev inequality to obtain probabilistic bounds of the forward error and applied it to Horner's rule. These bounds have the advantage of being closer to the forward error for a large n and a fixed probability than the ones based on the Azuma-Hoeffding inequality. At the same time, Higham and Mary [4] and Ipsen and Zhou [13] used martingales and the Azuma-Hoeffding inequality to obtain probabilistic bounds of the forward error. BC bounds prove better than AH asymptotically in n, while AH outperforms BC for $\lambda \to 0$. In the following, we present these two methods and show that SR benefits extend to pairwise summation. In particular, our probabilistic bounds are lower than any deterministic ones (at the expense of introducing a probability that they do not hold).

3.1. BC method. Let us recall the lemma that bounds the variance of an error product $\varphi = \prod_{k=1}^{n}(1+\delta_k)$ under SR-nearness. A general expression of this lemma can be found in [8, Lemma 3.1].

LEMMA 3.1. *(from [8, Lemma 3.1]) Under SR-nearness φ satisfies*
1. $E(\varphi) = 1$.
2. $V(\varphi) \leq \gamma_n(u^2)$,

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

This lemma has been used to study the inner product and Horner's algorithm in [8]. For the pairwise summation (provided that $\sum x_i \neq 0$, because we are considering a relative error that cannot be defined if the result is 0), we have

THEOREM 3.2. *For all $0 < \lambda < 1$, the computed $\widehat{S}_1^h$ satisfies under SR-nearness*

$$\frac{\left|\widehat{S}_1^h - S_1^h\right|}{\left|S_1^h\right|} \leq \kappa\sqrt{\gamma_{\log(n)}(u^2)/\lambda}, \tag{3.2}$$

with probability at least $1-\lambda$, where $\kappa = \frac{\sum_{i=1}^{n}|x_i|}{\left|\sum_{i=1}^{n} x_i\right|}$ is the condition number using the 1-norm of the sum of the x_i.

Proof. By expectation linearity, $E(\widehat{S}_1^h) = \sum_{i=1}^{2^h} x_i E(\psi_i)$. Lemma 3.1 shows that for all $1 \leq i \leq 2^h$, $E(\psi_i) = 1$ and $V(\psi_i) \leq \gamma_h(u^2)$. It follows that, $E(\widehat{S}_1^h) = S_1^h$ and $V(\widehat{S}_1^h) \leq \left(\sum_{i=1}^{2^h} |x_i| \sqrt{V(\psi_i)}\right)^2 \leq \|x\|_1^2 \gamma_h(u^2)$. the Bienaymé–Chebyshev inequality implies $\mathbb{P}\left(\left|\widehat{S}_1^h - E(\widehat{S}_1^h)\right| \leq \sqrt{V(\widehat{S}_1^h)/\lambda}\right) \geq 1-\lambda$. Thus, with probability at least $1-\lambda$,

$$\frac{\left|\widehat{S}_1^h - S_1^h\right|}{\left|S_1^h\right|} \leq \frac{1}{\left|S_1^h\right|}\sqrt{V(\widehat{S}_1^h)/\lambda} \leq \frac{\|x\|_1}{\left|S_1^h\right|}\sqrt{\gamma_h(u^2)/\lambda} = \kappa\sqrt{\gamma_h(u^2)/\lambda}.$$

Since $h = \log(n)$, we have with probability at least $1-\lambda$,

$$\frac{\left|\widehat{S}_1^h - S_1^h\right|}{\left|S_1^h\right|} \leq \kappa\sqrt{\gamma_{\log(n)}(u^2)/\lambda}.$$

∎

3.2. AH method. This method uses martingales and then applies the Azuma-Hoeffding inequality for a martingale [2, 12].

Definition 3.3. A sequence of random variables $(M_0, \dots, M_n)$ is a martingale with respect to the sequence $X_0, \dots, X_n$ if, for all k,

- M_k is a function of $X_0, \dots, X_k$,
- $E(|M_k|) < \infty$, and
- $E[M_k/X_0, \dots, X_{k-1}] = M_{k-1}$.

If $E[M_k/X_0, \dots, X_{k-1}] \geq M_{k-1}$, $(M_0, \dots, M_n)$ is called submartingale.

LEMMA 3.4 (Azuma-Hoeffding inequality). *Let $(M_0, \dots, M_n)$ be a martingale with respect to a sequence $X_0, \dots, X_n$. We assume that there exist $a_k < b_k$ such that $a_k \leq M_k - M_{k-1} \leq b_k$ for $k \in \{1, \dots, n\}$. 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 $0 < \lambda < 1$.

THEOREM 3.5. *For all $0 < \lambda < 1$, the computed $\widehat{S}_1^h$ satisfies under SR-nearness*

$$\frac{\left|\widehat{S}_1^h - S_1^h\right|}{\left|S_1^h\right|} \leq \kappa\sqrt{u\gamma_{2\lceil\log(n)\rceil}(u)}\sqrt{\ln(2/\lambda)}, \tag{3.3}$$

with probability at least $1 - \lambda$.

Proof. Let us denote for $k > 0$, $M_k = \sum_{i=1}^{2^{h-k}} \widehat{S}_i^k - S_i^k$ and $M_0 = 0$. Then, $M_h = \widehat{S}_1^h - S_1^h$ and $M_k = M_{k-1} + \sum_{i=1}^{2^{h-k}} (\widehat{S}_{2i-1}^{k-1} + \widehat{S}_{2i}^{k-1})\delta_i^k$. The δ_k are mean independent, therefore $M_0, \dots, M_h$ form a martingale with respect to $\{\delta_i^k,\ 1 \leq i \leq 2^{h-k},\ 1 \leq k \leq h-1\}$. Moreover, Equation (3.1) yields

$$\begin{aligned}
|M_k - M_{k-1}| &\leq \sum_{i=1}^{2^{h-k}} \left|(\widehat{S}_{2i-1}^{k-1} + \widehat{S}_{2i}^{k-1})\delta_i^k\right| \leq u\sum_{i=1}^{2^{h-k}} \left|\widehat{S}_{2i-1}^{k-1} + \widehat{S}_{2i}^{k-1}\right| \\
&\leq u(1+u)^{k-1}\sum_{i=1}^{2^{h-k}} \left|\sum_{m=2^{k-1}(2i-2)+1}^{2^{k-1}(2i-1)} x_m + \sum_{m=2^{k-1}(2i-1)+1}^{2^{k-1}(2i)} x_m\right| \\
&\leq u(1+u)^{k-1}\sum_{i=1}^{2^{h-k}} \sum_{m=2^k(i-1)+1}^{2^k i} |x_m| = u(1+u)^{k-1}\sum_{i=1}^{2^h} |x_m| \\
&= u(1+u)^{k-1}\left\|x\right\|_1.
\end{aligned}$$

Denote $C_k = u(1+u)^{k-1}\left\|x\right\|_1$, Azuma-Hoeffding inequality implies that with probability at least $1 - \lambda$, $|M_h| \leq \sqrt{\sum_{k=1}^h C_k^2}\sqrt{2\ln(2/\lambda)}$. Now

$$\sum_{k=1}^h C_k^2 = u^2\left\|x\right\|_1^2 \sum_{k=1}^h (1+u)^{2(k-1)} = u^2\left\|x\right\|_1^2 \frac{(1+u)^{2h} - 1}{(1+u)^2 - 1} = u\left\|x\right\|_1^2 \frac{\gamma_{2h}(u)}{u+2}.$$

Since, $\frac{u}{u+2} \leq \frac{u}{2}$ and $h = \lceil \log(n) \rceil$, we have $|M_h| \leq \|x\|_1 \sqrt{u \frac{\gamma_{2\lceil \log(n) \rceil}(u)}{2}} \sqrt{2\ln(2/\lambda)}$. Finally

$$\frac{\left|\widehat{S}_1^h - S_1^h\right|}{\left|S_1^h\right|} \leq \kappa \sqrt{u \gamma_{2\lceil \log(n) \rceil}(u)} \sqrt{\ln(2/\lambda)},$$

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

Comparison with Hallman and Ipsen's pairwise bound [11]. The probabilistic bound proposed in [11, cor, 2.14] to the pairwise summation forward error is

$$\frac{\left|\widehat{S}_1^h - S_1^h\right|}{\left|S_1^h\right|} \leq \kappa u \sqrt{h} \sqrt{2\ln(2/\delta)}(1 + \phi_{n,h,\eta}), \tag{3.4}$$

with probability at least $1 - (\eta + \delta)$, where h is the height of the computational tree and $\phi_{n,h,\eta} \equiv \lambda_{n,\eta} \sqrt{2h} u \exp\left(\lambda_{n,\eta}^2 h u^2\right)$ with $\lambda_{n,\eta} \equiv \sqrt{2\ln(2n/\eta)}$.

Bound (3.4) uses two parameters δ and η, the sum of which is higher than the probability that the bound does not hold. We could not find any closed form describing the best value for η and δ given that $\delta + \eta = \lambda$, and we doubt that such a form exists. This makes the choice of their values difficult and, to some extent, arbitrary.

Taking $\delta + \eta = \lambda$ and $h = \log(n)$, comparing the two bounds adds up to comparing $\sqrt{\gamma_{2\log(n)}(u)} \sqrt{\ln(2/(\delta + \eta)}$ and $\sqrt{2uh}\sqrt{\ln(2/\delta)}(1 + \phi_{n,h,\eta})$. $\gamma_{2\log(n)}(u) > 2hu$ and $\gamma_{2\log(n)}(u) = 2hu + O((hu)^2)$, giving a short advantage to (3.4) regarding the first factor. However $\ln(2/(\delta + \eta)) < \ln(2/\delta)$, and to close the gap, one needs to take η as small as possible, giving the advantage to (3.2) for the second factor. In the third factor, taking $\eta \to 0$ makes $\phi_{n,h,u}$ grow to ∞. Moreover, the term $\phi_{n,h,u}$ is in $O(hu)$, and thus grows more rapidly than the second order terms in bound (3.2). All in all, the bound established in this paper avoids the use of two parameters with no easy way to choose their values and gives better asymptotic results than the one in (3.4).

4. Bias analysis. The unbiased nature of SR-nearness extends to various algorithms such as the inner product [4] and Horner's rule [9]. Nevertheless, it fails to hold in the general case. In the sequel, we study two algorithms for computing the variance: textbook and two-pass.

4.1. Textbook algorithm. For $x \in \mathbb{R}^n$, let $s = \sum_{i=1}^n x_i$ and $y = \sum_{i=1}^n x_i^2 - \frac{1}{n} s^2$. Using SR-nearness:

- The computed $\widehat{s}$ satisfies $\widehat{s} = \sum_{i=1}^n x_i \prod_{k=\max(2,i)}^n (1+\delta_{k-1}) = \sum_{i=1}^n x_i \phi_i$ with $\phi_i = \prod_{k=\max(2,i)}^n (1 + \delta_{k-1})$ for all $1 \leq i \leq n$.
- The computed $\widehat{y}$ satisfies

$$\widehat{y} = \sum_{i=1}^n x_i^2 \psi_i - \frac{1}{n} \widehat{s}^2 \psi_{n+1}, \tag{4.1}$$

where $\psi_i = (1 + \epsilon_i) \prod_{k=\max(2,i)}^{n+1} (1 + \eta_k)$ and $\psi_{n+1} = (1 + \epsilon_{n+1})(1 + \eta_{n+1})(1 + \theta)$. For all $1 \leq i \leq n+1$, ϵ_i and η_i represent the rounding errors from the products and additions, respectively. θ represent the error of the division of $\widehat{s}^2$ by n.

THEOREM 4.1. *The quantities $\widehat{s}$ and $\widehat{y}$ satisfy under SR-nearness*

- $E(\widehat{s}) = s$,
- $E(\widehat{y}) = y - \frac{1}{n} V(\widehat{s})$.

Proof. The first item can be proved as in the first part of Theorem 3.2 proof. For the second, we have, by expectation linearity, $E(\widehat{y}) = \sum_{i=1}^{n} x_i^2 E(\psi_i) - \frac{1}{n} E(\widehat{s}^2 \psi_{n+1})$. Let $\mathbb{F} = \{\delta_i, \epsilon_j, \eta_k, i \in \{1, \dots, n-1\},\ j \in \{1, \dots, n\},\ k \in \{2, \dots, n\}\}$, the mean independence property implies that $E(\psi_i) = 1$ for all $1 \leq i \leq n$ and $E[\psi_{n+1}/\mathbb{F}] = 1$. Therefore, the law of total expectation $E(X) = E(E[X/Y])$ yields

$$\begin{aligned} E(\widehat{s}^2 \psi_{n+1}) &= E\left(E[\widehat{s}^2 \psi_{n+1}/\mathbb{F}]\right) = E\left(\widehat{s}^2 E[\psi_{n+1}/\mathbb{F}]\right) = E(\widehat{s}^2) \\ &= E(\widehat{s})^2 + V(\widehat{s}) = s^2 + V(\widehat{s}). \end{aligned}$$

It follows that $E(\widehat{y}) = \sum_{i=1}^{n} x_i^2 - \frac{1}{n} s^2 - \frac{1}{n} V(\widehat{s}) = y - \frac{1}{n} V(\widehat{s})$. ∎

Remark 4.2. Lemma 3.1 gives $V(\phi_i) \leq \gamma_{n-1}(u^2)$, so [8, thm 3.2] shows that the bias satisfies

$$\frac{1}{n} V(\widehat{s}) \leq \frac{1}{n} \|x\|_1^2 \gamma_{n-1}(u^2) = y \mathcal{K}_1^2 \gamma_{n-1}(u^2).$$

Thus $E(\widehat{y}) \geq y\left(1 - \mathcal{K}_1^2 \gamma_{n-1}(u^2)\right)$.

4.2. Two-pass algorithm. Let $x_1, x_2, \dots, x_n \in \mathbb{R}$, denote $m = \frac{1}{n} \sum_{i=1}^{n} x_i$ and $z = \sum_{i=1}^{n} (x_i - m)^2$. Using SR-nearness:

- The computed $\widehat{m}$ satisfies $\widehat{m} = \frac{1}{n} \sum_{i=1}^{n} x_i \prod_{k=\max(2,i)}^{n+1} (1 + \delta_{k-1})$ with δ_n is the division error by n.
- The computed $\widehat{z}$ satisfies

$$\widehat{z} = \sum_{i=1}^{n} (x_i - \widehat{m})^2 \psi_i, \tag{4.2}$$

where $\psi_i = (1+\epsilon_i)^2 (1+\eta_i) \prod_{k=\max(2,i)}^{n} (1+\theta_k)$. For all $1 \leq i \leq n$, ϵ_i, η_i and θ_i represent the rounding errors of subtraction, square, and addition, respectively. Let us denote $\varphi_i = (1+\epsilon_i)(1+\eta_i) \prod_{k=\max(2,i)}^{n} (1+\theta_k)$. Then $\psi_i = (1+\epsilon_i)\varphi_i$.

THEOREM 4.3. *The quantities $\widehat{m}$ and $\widehat{z}$ satisfy under SR-nearness*

- $E(\widehat{m}) = m$,
- $E(\widehat{z}) = z + \frac{1}{n} V(\widehat{s}) + O(nu^2)$, *where* $\frac{1}{n} s = m$.

Proof. The first item is similar to the first part of Theorem 3.2 proof. For the second, we have by expectation linearity $E(\widehat{z}) = \sum_{i=1}^{n} E\left((x_i - \widehat{m})^2 \psi_i\right)$. For all $1 \leq i \leq n$, let $\theta_1 = 0$ and

$$\mathbb{F}_i = \{\delta_j, \epsilon_k, \eta_l, \theta_l,\ j \in [1; n],\ k \in [1; i],\ \text{and } l \in [1; i-1]\}.$$

The mean independence property implies that $E[(1+\eta_i) \prod_{k=\max(2,i)}^{n} (1+\theta_k)/\mathbb{F}_i] = 1$. Using the law of total expectation, we have

$$\begin{aligned} E\left((x_i - \widehat{m})^2 \psi_i\right) &= E\left(E\left[(x_i - \widehat{m})^2 (1+\epsilon_i)^2 (1+\eta_i) \prod_{k=\max(2,i)}^{n} (1+\theta_k)/\mathbb{F}_i\right]\right) \\ &= E\left((x_i - \widehat{m})^2 (1+\epsilon_i)^2 E\left[(1+\eta_i) \prod_{k=\max(2,i)}^{n} (1+\theta_k)/\mathbb{F}_i\right]\right) \\ &= E\left((x_i - \widehat{m})^2 (1+\epsilon_i)^2\right) = E\left((x_i - \widehat{m})^2 (1 + 2\epsilon_i + \epsilon_i^2)\right) \\ &= E\left((x_i - \widehat{m})^2 (1+\epsilon_i^2)\right) \qquad \text{by Lemma 2.2} \\ &= E\left((x_i - \widehat{m})^2\right) + E\left((x_i - \widehat{m})^2 \epsilon_i^2\right) \\ &= (x_i - m)^2 + V(\widehat{m}) + E\left((x_i - \widehat{m})^2 \epsilon_i^2\right). \end{aligned}$$

It follows that

$$E(\widehat{z}) = \sum_{i=1}^{n}(x_i - m)^2 + V(\widehat{m}) + E\left((x_i - \widehat{m})^2\epsilon_i^2\right)$$
$$= z + nV(\widehat{m}) + \sum_{i=1}^{n} E\left((x_i - \widehat{m})^2\epsilon_i^2\right).$$

Since $\widehat{m} = \frac{1}{n}(1+\delta_n)\widehat{s} = \frac{1}{n}\widehat{s} + \frac{1}{n}\delta_n\widehat{s}$ and $|\epsilon_i|^2, |\delta_n|^2 \leq u^2$ for all $1 \leq i \leq n$,

$$V(\widehat{m}) = \frac{1}{n^2}V(\widehat{s}) + O(nu^2) \quad \text{and} \quad \sum_{i=1}^{n} E\left((x_i - \widehat{m})^2\epsilon_i^2\right) = O(nu^2).$$

Therefore $E(\widehat{z}) = z + \frac{1}{n}V(\widehat{s}) + O(nu^2)$. ∎

Interestingly, these two algorithms under SR have an opposed bias at the first order over u.

Remark 4.4. Lemma 3.1 implies that $V(\widehat{m}) \leq \frac{1}{n^2}\|x\|_1^2\gamma_n(u^2)$. Then

$$E(\widehat{z}) = z + nV(\widehat{m}) + \sum_{i=1}^{n} E\left((x_i - \widehat{m})^2\epsilon_i^2\right) \leq z + nV(\widehat{m}) + u^2\sum_{i=1}^{n} E\left((x_i - \widehat{m})^2\right)$$
$$= z + nV(\widehat{m}) + u^2(z + nV(\widehat{m}))$$
$$\leq (1+u^2)(z + \frac{1}{n}\|x\|_1^2\gamma_n(u^2)) = z(1+u^2)(1+\mathcal{K}_1^2\gamma_n(u^2)).$$

5. Error analysis for algorithms with non-linear error. This section examines SR for non-linear computations via the previous two algorithms. We use the two methods discussed in the introduction to estimate the forward error. In addition, a new approach based on Doob-Meyer decomposition is proposed for the textbook algorithm.

5.1. BC method. This section uses the BC method proposed in [8] to provide a probabilistic bound on the forward error of both textbook and two-pass algorithms under SR-nearness.

5.1.1. Textbook algorithm. In order to estimate the forward errors of the textbook algorithm, compute

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

Let $\mathcal{B} = \left|(\widehat{s} - s)^2\psi_{n+1}\right| + 2\left|s(\widehat{s} - s)\psi_{n+1}\right| + \left|s^2(\psi_{n+1} - 1)\right|$, the following equation will be used in all proofs of the textbook forward errors

$$|\widehat{y} - y| = \left|\sum_{i=1}^{n} x_i^2(\psi_i - 1)\right| + \frac{1}{n}\mathcal{B}. \tag{5.1}$$

Remark 5.1. To handle the non-linearity of errors, the key idea of this approach is to isolate terms of order 1 in error the errors and then use the previous results on the inner product or summation. This error could be decomposed otherwise. For instance,

$$\frac{1}{n}(\widehat{s}^2\psi_{n+1} - s^2) = \frac{1}{n}(\widehat{s}^2\psi_{n+1} - \widehat{s}s + \widehat{s}s - s^2) = \frac{1}{n}\left(\widehat{s}(\widehat{s}\psi_{n+1} - s) + s(\widehat{s} - s)\right).$$

Then, we can apply the same properties on $(\widehat{s}\psi_{n+1} - s)$ and $(\widehat{s} - s)$. The bounds are different but asymptotically equivalent when $nu \ll 1$.

The rounding errors accumulated in the whole process of this algorithm ϕ_i and ψ_i satisfy for all $1 \leq i \leq n$,

$$|\phi_i| \leq (1+u)^{n+1-\max(2,i)}, \quad |\psi_i| \leq (1+u)^{n+3-\max(2,i)} \text{ and } |\psi_{n+1}| \leq (1+u)^3.$$

Let us compute the deterministic bound of this algorithm. We have

$$\left|\sum_{i=1}^{n} x_i^2(\psi_i - 1)\right| \leq \|x\|_2^2 \, \gamma_{n+1}(u).$$

Since $|s| \leq \|x\|_1$ and $|\widehat{s} - s| = |\sum_{i=1}^n x_i(\phi_i - 1)| \leq \|x\|_1 \, \gamma_{n-1}(u)$,

$$\begin{aligned}
\mathcal{B} &\leq (1+u)^3 \|x\|_1^2 \left(\gamma_{n-1}^2(u) + 2\gamma_{n-1}(u)\right) + \|x\|_1^2 \left((1+u)^3 - 1\right) \\
&= (1+u)^3 \|x\|_1^2 \left(\gamma_{n-1}^2(u) + 2\gamma_{n-1}(u) + 1\right) - \|x\|_1^2 \\
&= (1+u)^3 \|x\|_1^2 \left(\gamma_{n-1}(u) + 1\right)^2 - \|x\|_1^2 \\
&= \|x\|_1^2 (1+u)^{2n+1} - \|x\|_1^2 = \|x\|_1^2 \, \gamma_{2n+1}(u).
\end{aligned}$$

Finally

$$\frac{|\widehat{y} - y|}{|y|} \leq \mathcal{K}_2^2 \gamma_{n+1}(u) + \mathcal{K}_1^2 \gamma_{2n+1}(u). \tag{5.2}$$

The following theorem presents a probabilistic bound of the forward error of this algorithm through the BC method.

THEOREM 5.2. *For all $0 < \lambda < 1$, the computed $\widehat{y}$ in Equation (4.1) satisfies under SR-nearness*

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

with probability at least $1 - \lambda$.

Proof. Equation (5.1) states that $|\widehat{y} - y| \leq \left|\sum_{i=1}^n x_i^2(\psi_i - 1)\right| + \frac{1}{n}\mathcal{B}$. The quantities $\left|\sum_{i=1}^n x_i^2(\psi_i - 1)\right|$ and $|\widehat{s} - s|$ represent the absolute errors of the inner product $\sum_{i=1}^n x_i^2$ of the vector x by itself and the summation $s = \sum_{i=1}^n x_i$, respectively. Then [8, sec 5.1] proves that

$$\left|\sum_{i=1}^{n} x_i^2(\psi_i - 1)\right| \leq \|x\|_2^2 \sqrt{2\gamma_{n+1}(u^2)/\lambda} \qquad \text{with probability at least } 1 - \frac{\lambda}{2},$$

$$|\widehat{s} - s| \leq \|x\|_1 \sqrt{2\gamma_{n-1}(u^2)/\lambda} \qquad \text{with probability at least } 1 - \frac{\lambda}{2}.$$

Since, $|\psi_{n+1}| \leq (1+u)^3$ and $|s| \leq \|x\|_1$, with probability at least $1 - \frac{\lambda}{2}$,

$$\begin{aligned}\mathcal{B} &\leq (1+u)^3 \|x\|_1^2 \left(2\gamma_{n-1}(u^2)/\lambda + 2\sqrt{2\gamma_{n-1}(u^2)/\lambda}\right) + \|x\|_1^2 \left((1+u)^3 - 1\right) \\ &= (1+u)^3 \|x\|_1^2 \left(2\gamma_{n-1}(u^2)/\lambda + 2\sqrt{2\gamma_{n-1}(u^2)/\lambda} + 1\right) - \|x\|_1^2 \\ &= (1+u)^3 \|x\|_1^2 \left(\sqrt{2\gamma_{n-1}(u^2)/\lambda} + 1\right)^2 - \|x\|_1^2 .\end{aligned}$$

Finally, Lemma 2.1 shows that with probability at least $1 - \lambda$,

$$\begin{aligned}\frac{|\widehat{y} - y|}{|y|} &\leq \frac{1}{|y|} \left|\sum_{i=1}^n x_i^2(\psi_i - 1)\right| + \frac{1}{n\,|y|}\mathcal{B} \\ &\leq \mathcal{K}_2^2 \sqrt{2\gamma_{n+1}(u^2)/\lambda} + \mathcal{K}_1^2 \left((1+u)^3 \left(\sqrt{2\gamma_{n-1}(u^2)/\lambda} + 1\right)^2 - 1\right). \qquad \square\end{aligned}$$

5.1.2. Two-pass algorithm. As with the previous algorithm, we present a computational scheme for the proofs of the two-pass algorithm errors in this paper. One needs first to separate the errors of order two. Let us recall that $\psi_i = \varphi_i(1+\epsilon_i)$ for all $1 \leq i \leq n$. Therefore

$$\begin{aligned}|\widehat{z} - z| &= \left|\sum_{i=1}^n (x_i - \widehat{m})^2 \psi_i - (x_i - m)^2\right| \\ &= \left|\sum_{i=1}^n (x_i - \widehat{m})^2 \varphi_i - (x_i - m)^2 + \sum_{i=1}^n (x_i - \widehat{m})^2 \epsilon_i \varphi_i\right| \\ &\leq \left|\sum_{i=1}^n (x_i - \widehat{m})^2 \varphi_i - (x_i - m)^2\right| + u \left|\sum_{i=1}^n (x_i - \widehat{m})^2 \varphi_i\right| \\ &\leq \left|\sum_{i=1}^n (x_i - \widehat{m})^2 \varphi_i - (x_i - m)^2\right| + u \left|\sum_{i=1}^n (x_i - \widehat{m})^2 \varphi_i - (x_i - m)^2\right| + u\,|z| \\ &= (1+u) \left|\sum_{i=1}^n (x_i - \widehat{m})^2 \varphi_i - (x_i - m)^2\right| + u\,|z| .\end{aligned}$$

Since $(x_i - \widehat{m}) = (x_i - m) + (m - \widehat{m})$,

$$\begin{aligned}\left|\sum_{i=1}^n (x_i - \widehat{m})^2 \varphi_i - (x_i - m)^2\right| \leq \left|\sum_{i=1}^n (x_i - m)^2(\varphi_i - 1)\right| + \left|(m - \widehat{m})^2 \sum_{i=1}^n \varphi_i\right| \\ + 2\left|(m - \widehat{m}) \sum_{i=1}^n (x_i - m)(\varphi_i - 1)\right|,\end{aligned}$$

because $\sum_{i=1}^n (x_i - m) = 0$. Denote

$$\mathcal{C} = \left|\sum_{i=1}^n (x_i - m)^2(\varphi_i - 1)\right| + 2\left|(m - \widehat{m}) \sum_{i=1}^n (x_i - m)(\varphi_i - 1)\right| + \left|(m - \widehat{m})^2 \sum_{i=1}^n \varphi_i\right|.$$

The following equation will be used in all proofs of the two-pass forward errors

$$|\widehat{z} - z| \leq (1+u)\mathcal{C} + u\,|z| . \tag{5.3}$$

The following theorem presents a probabilistic bound of the forward error of this algorithm through the BC method.

THEOREM 5.3. *For all $0 < \lambda < 1$, the computed $\widehat{z}$ in Equation (4.2) satisfies under SR-nearness*

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

with probability at least $1-\lambda$.

Proof. Equation (5.3) states that $|\widehat{z}-z| \leq (1+u)\mathcal{C} + u\,|z|$, and $|\sum_{i=1}^n \varphi_i| \leq |\sum_{i=1}^n(\varphi_i - 1)| + n$. The following quantities $\left|\sum_{i=1}^n (x_i-m)^2(\varphi_i-1)\right|$, $|\widehat{m}-m|$, $|\sum_{i=1}^n (x_i-m)(\varphi_i-1)|$ and $|\sum_{i=1}^n(\varphi_i-1)|$ represent the absolute errors of the inner product of $x-m$ by itself $\sum_{i=1}^n(x_i-m)^2$, the average $m = \frac{1}{n}\sum_{i=1}^n x_i$, the summations $s = \sum_{i=1}^n (x_i - m)$ and $\sum_{i=1}^n 1$ respectively. Then [8, sec 5.1] proves that

$$\left|\sum_{i=1}^n (x_i-m)^2(\varphi_i-1)\right| \leq |z|\sqrt{\frac{4\gamma_{n+1}(u^2)}{\lambda}} \qquad \text{with probability at least } 1-\frac{\lambda}{4},$$

$$|\widehat{m}-m| \leq \frac{1}{n}\|x\|_1\sqrt{\frac{4\gamma_n(u^2)}{\lambda}} \qquad \text{with probability at least } 1-\frac{\lambda}{4},$$

$$\left|\sum_{i=1}^n (x_i-m)(\varphi_i-1)\right| \leq \sum_{i=1}^n |x_i-m|\sqrt{\frac{4\gamma_{n+1}(u^2)}{\lambda}} \qquad \text{with probability at least } 1-\frac{\lambda}{4},$$

$$\left|\sum_{i=1}^n (\varphi_i-1)\right| \leq n\sqrt{\frac{4\gamma_{n+1}(u^2)}{\lambda}} + n \qquad \text{with probability at least } 1-\frac{\lambda}{4}.$$

Using the Cauchy–Schwarz inequality, we obtain

$$\sum_{i=1}^n |x_i-m| \leq \sqrt{n\sum_{i=1}^n (x_i-m)^2} = \sqrt{nz}.$$

Since $\gamma_n(u^2) \leq \gamma_{n+1}(u^2)$, Lemma 2.1 implies

$$\begin{aligned}\mathcal{C} &\leq |z|\sqrt{\frac{4\gamma_{n+1}(u^2)}{\lambda}} + 2\frac{\|x\|_1}{n}\frac{4\gamma_{n+1}(u^2)}{\lambda}\sqrt{nz} + \frac{\|x\|_1^2}{n}\frac{4\gamma_{n+1}(u^2)}{\lambda}\left(\sqrt{\frac{4\gamma_{n+1}(u^2)}{\lambda}}+1\right)\\ &= |z|\sqrt{\frac{4\gamma_{n+1}(u^2)}{\lambda}} + \frac{4\gamma_{n+1}(u^2)}{\lambda}\left(2\,|z|\,\frac{\|x\|_1}{\sqrt{nz}} + \frac{\|x\|_1^2}{n}\left(\sqrt{\frac{4\gamma_{n+1}(u^2)}{\lambda}}+1\right)\right),\end{aligned}$$

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

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

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

5.2. AH method. This section uses the AH method proposed in [13] for the inner product and Lemma 2.1 to provide a probabilistic bound of the forward error of both textbook and two-pass algorithms under SR-nearness.

5.2.1. Textbook algorithm.

THEOREM 5.4. *For all $0 < \lambda < 1$, the computed $\widehat{y}$ in Equation (4.1) satisfies under SR-nearness*

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

with probability at least $1-\lambda$.

Proof. Equation (5.1) states that $|\widehat{y}-y| \leq \left|\sum_{i=1}^n x_i^2(\psi_i-1)\right| + \frac{1}{n}\mathcal{B}$. Moreover, [13, cor 4.7] shows that

$$\begin{aligned}\left|\sum_{i=1}^n x_i^2(\psi_i-1)\right| &\leq \|x\|_2^2\sqrt{u\gamma_{2(n+1)}(u)}\sqrt{\ln(4/\lambda)} \quad \text{with probability at least } 1-\frac{\lambda}{2}, \\ |\widehat{s}-s| &\leq \|x\|_1\sqrt{u\gamma_{2(n-1)}(u)}\sqrt{\ln(4/\lambda)} \quad \text{with probability at least } 1-\frac{\lambda}{2}.\end{aligned}$$

Since, $|\psi_{n+1}| \leq (1+u)^3$ and $|s| \leq \|x\|_1$,we have with probability at least $1-\frac{\lambda}{2}$,

$$\begin{aligned}\mathcal{B} \leq & (1+u)^3\|x\|_1^2\left(u\gamma_{2(n-1)}(u)\ln(4/\lambda)+2\sqrt{u\gamma_{2(n-1)}(u)}\sqrt{\ln(4/\lambda)}\right) \\ & + \|x\|_1^2\left((1+u)^3-1\right) \\ = & (1+u)^3\|x\|_1^2\left(u\gamma_{2(n-1)}(u)\ln(4/\lambda)+2\sqrt{u\gamma_{2(n-1)}(u)}\sqrt{\ln(4/\lambda)}+1\right)-\|x\|_1^2 \\ = & (1+u)^3\|x\|_1^2\left(\sqrt{u\gamma_{2(n-1)}(u)}\sqrt{\ln(4/\lambda)}+1\right)^2-\|x\|_1^2.\end{aligned}$$

Finally, Lemma 2.1 shows that with probability at least $1-\lambda$,

$$\begin{aligned}\frac{|\widehat{y}-y|}{|y|} \leq & \mathcal{K}_2^2\sqrt{u\gamma_{2(n+1)}(u)}\sqrt{\ln(4/\lambda)} \\ & + \mathcal{K}_1^2\left((1+u)^3\left(\sqrt{u\gamma_{2(n-1)}(u)}\sqrt{\ln(4/\lambda)}+1\right)^2-1\right).\end{aligned}$$

∎

5.2.2. Two-pass algorithm.

THEOREM 5.5. *For all $0 < \lambda < 1$, the computed $\widehat{z}$ in Equation (4.2) satisfies under SR-nearness*

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

with probability at least $1-\lambda$.

Proof. Equation (5.3) states that $|\widehat{z}-z| \leq (1+u)\mathcal{C}+u|z|$. Note that $|\sum_{i=1}^n \varphi_i| \leq |\sum_{i=1}^n(\varphi_i-1)|+n$ and [13, cor 4.7] shows that each of the following inequalities holds

with probability at least $1 - \frac{\lambda}{4}$:

$$\begin{aligned}
\left|\sum_{i=1}^{n}(x_i - m)^2(\varphi_i - 1)\right| &\leq |z| \sqrt{u\gamma_{2(n+1)}(u)}\sqrt{\ln(8/\lambda)},\\
|\widehat{m} - m| &\leq \frac{1}{n}\left\|x\right\|_1 \sqrt{u\gamma_{2n}(u)}\sqrt{\ln(8/\lambda)},\\
\left|\sum_{i=1}^{n}(x_i - m)(\varphi_i - 1)\right| &\leq \sum_{i=1}^{n}|x_i - m| \sqrt{u\gamma_{2(n+1)}(u)}\sqrt{\ln(8/\lambda)},\\
\left|\sum_{i=1}^{n}(\varphi_i - 1)\right| &\leq n\sqrt{u\gamma_{2(n+1)}(u)}\sqrt{\ln(8/\lambda)}.
\end{aligned}$$

By the Cauchy–Schwarz inequality, $\sum_{i=1}^{n}|x_i - m| \leq \sqrt{n\sum_{i=1}^{n}(x_i - m)^2} = \sqrt{n}z$. Since $\gamma_{2n}(u) \leq \gamma_{2(n+1)}(u)$, Lemma 2.1 implies

$$\begin{aligned}
\mathcal{C} \leq& |z| \sqrt{u\gamma_{2(n+1)}(u)}\sqrt{\ln(8/\lambda)} + 2\frac{\left\|x\right\|_1}{n}u\gamma_{2(n+1)}(u)\ln(8/\lambda)\sqrt{n}z\\
&+ \frac{\left\|x\right\|_1^2}{n^2}u\gamma_{2(n+1)}(u)\ln(8/\lambda)\left(n\sqrt{u\gamma_{2(n+1)}(u)}\sqrt{\ln(8/\lambda)} + n\right)\\
=& |z| \sqrt{u\gamma_{2(n+1)}(u)}\sqrt{\ln(8/\lambda)} + u\gamma_{2(n+1)}(u)\ln(8/\lambda)\Bigg(2|z|\frac{\left\|x\right\|_1}{\sqrt{n}z}\\
&+ \frac{\left\|x\right\|_1^2}{n}\left(\sqrt{u\gamma_{2(n+1)}(u)}\sqrt{\ln(8/\lambda)} + 1\right)\Bigg),
\end{aligned}$$

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

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

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

5.2.3. Textbook algorithm and Doob-Meyer decomposition. This work introduces a new approach based on Doob–Meyer decomposition [6, p 68] to bound the forward error of the textbook algorithm. To apply this method, we study

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

Consider $s_1 = x_1$, $s_k = s_{k-1} + x_k$ and $\widehat{s}_1 = x_1$, $\widehat{s}_k = (\widehat{s}_{k-1} + x_k)(1 + \delta_{k-1})$ for all $2 \leq k \leq n$. Then $s_n = s$ and $\widehat{s}_n = \widehat{s}$. Denote $Z_k = \widehat{s}_k - s_k = Z_{k-1} + (\widehat{s}_{k-1} + x_k)\delta_{k-1}$. Then, $Z_n = \widehat{s}_n - s_n$. By mean independence of δ_k, $Z_1, \ldots, Z_n$ form a martingale with respect to $\delta_1, \ldots, \delta_{n-1}$. Then, $Z_1 + s, \ldots, Z_n + s$ is also a martingale. Denote:

- $\mathbb{F}_k = \{\delta_1, \ldots, \delta_k\}$.
- $Y_{k-1} = Z_k - Z_{k-1} = (\widehat{s}_{k-1} + x_k)\delta_{k-1}$ for all $2 \leq k \leq n$. Then $Z_n = \sum_{k=2}^{n} Y_{k-1}$.

- $\sigma_{k-1}^2 = E[Y_{k-1}^2/\mathbb{F}_{k-2}]$.
- $A_n = \sum_{k=2}^{n} \sigma_{k-1}^2$ with $A_1 = 0$.

On one hand, A_n is predictable:

$$\begin{aligned} E[A_n/\mathbb{F}_{n-1}] &= E\left[\sum_{k=2}^{n} \sigma_{k-1}^2/\mathbb{F}_{n-1}\right] \\ &= E\left[\sum_{k=2}^{n} E\left[Y_{k-1}^2/\mathbb{F}_{k-2}\right]/\mathbb{F}_{n-1}\right] \\ &= \sum_{k=2}^{n} E[E[Y_{k-1}^2/\mathbb{F}_{k-2}]/\mathbb{F}_{n-1}]. \end{aligned}$$

Since $E[Y_{k-1}^2/\mathbb{F}_{k-2}]$ is $\mathbb{F}_{k-2}$-measurable, so it is $\mathbb{F}_{n-1}$-measurable, and for all $2 \le k \le n$, we have $E[E[Y_{k-1}^2/\mathbb{F}_{k-2}]/\mathbb{F}_{n-1}] = E[Y_{k-1}^2/\mathbb{F}_{k-2}]$. Then

$$E[A_n/\mathbb{F}_{n-1}] = \sum_{k=2}^{n} E[Y_{k-1}^2/\mathbb{F}_{k-2}] = A_n.$$

On the other hand, $X_n = (Z_n + s)^2 - A_n - s^2$ is a martingale:

$$\begin{aligned} E[X_n/\mathbb{F}_{n-1}] &= E[(Z_n + s)^2 - A_n - s^2/\mathbb{F}_{n-1}] \\ &= E[(Z_{n-1} + s + Y_{n-1})^2/\mathbb{F}_{n-1}] - A_n - s^2 \\ &= (Z_{n-1} + s)^2 + 2(Z_{n-1} + s)E[Y_{n-1}/\mathbb{F}_{n-1}] + E[Y_{n-1}^2/\mathbb{F}_{n-1}] - A_n - s^2 \\ &= X_{n-1} \quad \text{because } E[Y_{n-1}/\mathbb{F}_{n-1}] = 0. \end{aligned}$$

The expression of $(Z_n + s)^2 = X_n + s^2 + A_n$ is a Doob-Meyer decomposition.

LEMMA 5.6. *The martingale $X_1, \ldots, X_n$ satisfies $|X_k - X_{k-1}| \le uC_k$, for all $2 \le k \le n$, where*

$$C_k = \|x\|_1^2 (1 + u)^{2(k-2)}(2 + u).$$

Proof. Note that $\sigma_{k-1}^2 = E[(\widehat{s}_{k-1} + x_k)^2\delta_{k-1}^2/\mathbb{F}_{k-2}] = (\widehat{s}_{k-1} + x_k)^2 E[\delta_{k-1}^2/\mathbb{F}_{k-2}]$ by definition of $\mathbb{F}_{k-2}$. Then

$$\begin{aligned} X_k - X_{k-1} &= (Z_k + s)^2 - A_k - (Z_{k-1} + s)^2 + A_{k-1} \\ &= (Z_{k-1} + s + (\widehat{s}_{k-1} + x_k)\delta_{k-1})^2 - A_k - (Z_{k-1} + s)^2 + A_{k-1} \\ &= 2(Z_{k-1} + s)(\widehat{s}_{k-1} + x_k)\delta_{k-1} + (\widehat{s}_{k-1} + x_k)^2\delta_{k-1}^2 - \sigma_{k-1}^2 \\ &= 2(Z_{k-1} + s)(\widehat{s}_{k-1} + x_k)\delta_{k-1} + (\widehat{s}_{k-1} + x_k)^2 \left(\delta_{k-1}^2 - E[\delta_{k-1}^2/\mathbb{F}_{k-2}]\right). \end{aligned}$$

Since $|\delta_{k-1}| \le u$, we have $|\widehat{s}_{k-1} + x_k| \le (1 + u)^{k-2} \sum_{i=1}^{k} |x_i| \le (1 + u)^{k-2} \|x\|_1$, $\left|\delta_{k-1}^2 - E[\delta_{k-1}^2/\mathbb{F}_{k-2}]\right| \le u^2$ because $0 \le \delta_{k-1}^2 \le u^2$ and

$$|Z_{k-1} + s| \le ((1 + u)^{k-2} - 1) \sum_{i=1}^{k-1} |x_i| + |s| \le \|x\|_1 (1 + u)^{k-2}.$$

Thus

$$\begin{aligned} |X_k - X_{k-1}| &\le 2u\,|Z_{k-1} + s|\,|\widehat{s}_{k-1} + x_k| + u^2\,|\widehat{s}_{k-1} + x_k|^2 \\ &\le 2u(1 + u)^{2(k-2)} \|x\|_1^2 + u^2(1 + u)^{2(k-2)} \|x\|_1^2 \\ &= u \|x\|_1^2 (1 + u)^{2(k-2)}(2 + u). \end{aligned}$$

∎

THEOREM 5.7. *For $0 < \lambda < 1$, the martingale $X_1, \dots, X_n$ satisfies under SR-nearness*

$$|X_n| \leq \|x\|_1^2 \sqrt{2u\gamma_{4(n-1)}(u)}\sqrt{\ln(2/\lambda)}, \tag{5.4}$$

with probability at least $1-\lambda$.

Proof. Since $X_1 = 0$, Lemma 3.4 and Lemma 5.6 yields

$$|X_n| \leq \sqrt{\sum_{k=2}^{n} u^2 C_k^2}\sqrt{2\ln(2/\lambda)},$$

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

$$\begin{aligned}\sum_{k=2}^{n} u^2 C_k^2 &= u^2 \sum_{k=2}^{n} \|x\|_1^4 (1+u)^{4(k-2)}(2+u)^2 = u^2 \|x\|_1^4 (2+u)^2 \frac{\gamma_{4(n-1)}(u)}{(1+u)^4 - 1} \\ &= u\|x\|_1^4 \frac{4+4u+u^2}{4+6u+4u^2+u^3}\gamma_{4(n-1)}(u) \\ &\leq u\|x\|_1^4 \gamma_{4(n-1)}(u).\end{aligned}$$

Finally, $|X_n| \leq \|x\|_1^2 \sqrt{2u\gamma_{4(n-1)}(u)}\sqrt{\ln(2/\lambda)}$. ∎

We are now in a position to state the main result of this sub-section.

THEOREM 5.8. *For all $0 < \lambda < 1$, the computed $\widehat{y}$ in Equation (4.1) satisfies under SR-nearness*

$$\begin{aligned}\frac{|\widehat{y}-y|}{|y|} \leq & \mathcal{K}_2^2\sqrt{u\gamma_{2(n+1)}(u)}\sqrt{\ln(4/\lambda)} + \mathcal{K}_1^2(1+u)^3\big[\sqrt{2u\gamma_{4(n-1)}(u)}\sqrt{\ln(4/\lambda)} \\ & + u\frac{\gamma_{2(n-1)}(u)}{2} + 1\big] - \mathcal{K}_1^2,\end{aligned}$$

with probability at least $1-\lambda$. In the following, this bound will be called DM bound.

Proof. Recall that $Z_n = \widehat{s} - s$ and $(Z_n + s)^2 = X_n + s^2 + A_n$. Therefore, from Sub-section 4.1,

$$\begin{aligned}\widehat{y} - y &= \sum_{i=1}^{n} x_i^2(\psi_i - 1) - \frac{1}{n}\widehat{s}^2\psi_{n+1} + \frac{1}{n}s^2 = \sum_{i=1}^{n} x_i^2(\psi_i - 1) - \frac{1}{n}(Z_n+s)^2\psi_{n+1} + \frac{1}{n}s^2 \\ &= \sum_{i=1}^{n} x_i^2(\psi_i - 1) - \frac{1}{n}\psi_{n+1}(X_n + A_n) - \frac{1}{n}s^2(\psi_{n+1} - 1).\end{aligned}$$

Since $|\psi_{n+1}| \leq (1+u)^3$ and $|s| \leq \|x\|_1$, we deduce that

$$\begin{aligned}|\widehat{y} - y| &\leq \left|\sum_{i=1}^{n} x_i^2(\psi_i - 1)\right| + \frac{1}{n}(1+u)^3(|X_n| + |A_n|) + \frac{1}{n}\|x\|_1^2\gamma_3(u) \\ &= \left|\sum_{i=1}^{n} x_i^2(\psi_i - 1)\right| + \frac{1}{n}(1+u)^3\left(|X_n| + |A_n| + \|x\|_1^2\right) - \frac{1}{n}\|x\|_1^2.\end{aligned}$$

On one hand, Theorem 5.7 states that with probability at least $1 - \frac{\lambda}{2}$,

$$|X_n| \leq \|x\|_1^2 \sqrt{2u\gamma_{4(n-1)}(u)}\sqrt{\ln(4/\lambda)}.$$

On the other hand, $A_n = \sum_{k=2}^{n} E[Y_{k-1}^2/\mathbb{F}_{k-2}] = \sum_{k=2}^{n}(\widehat{s}_{k-1}+x_k)^2 E[\delta_{k-1}^2/\mathbb{F}_{k-2}]$, then

$$\begin{aligned}|A_n| &\le u^2 \sum_{k=2}^{n} \left|\widehat{s}_{k-1}+x_k\right|^2 \le u^2 \sum_{k=2}^{n} \left((1+u)^{k-2}\sum_{i=1}^{k}|x_i|\right)^2 \\ &\le u^2 \left\|x\right\|_1^2 \sum_{k=2}^{n}(1+u)^{2(k-2)} \le u^2 \left\|x\right\|_1^2 \frac{\gamma_{2(n-1)}(u)}{2u+u^2} \le u \left\|x\right\|_1^2 \frac{\gamma_{2(n-1)}(u)}{2}.\end{aligned}$$

Moreover [13, cor 4.7] yields:

$$\left|\sum_{i=1}^{n} x_i^2(\psi_i - 1)\right| \le \left\|x\right\|_2^2 \sqrt{u\gamma_{2(n+1)}(u)}\sqrt{\ln(4/\lambda)} \quad \text{with probability at least } 1-\frac{\lambda}{2}.$$

Finally, Lemma 2.1 implies

$$\begin{aligned}\frac{|\widehat{y}-y|}{|y|} \le& \frac{\left\|x\right\|_2^2}{|y|}\sqrt{u\gamma_{2(n+1)}(u)}\sqrt{\ln(4/\lambda)} + \frac{\left\|x\right\|_1^2}{n\,|y|}(1+u)^3\left(\sqrt{2u\gamma_{4(n-1)}(u)}\sqrt{\ln(4/\lambda)}\right. \\ &\left. + u\frac{\gamma_{2(n-1)}(u)}{2}+1\right) - \frac{\left\|x\right\|_1^2}{n\,|y|} \\ =& \mathcal{K}_2^2\sqrt{u\gamma_{2(n+1)}(u)}\sqrt{\ln(4/\lambda)} + \mathcal{K}_1^2(1+u)^3\left(\sqrt{2u\gamma_{4(n-1)}(u)}\sqrt{\ln(4/\lambda)}\right. \\ &\left. + u\frac{\gamma_{2(n-1)}(u)}{2}+1\right) - \mathcal{K}_1^2,\end{aligned}$$

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

6. Pairwise textbook and pairwise two-pass. In this section, we illustrate the continued applicability of SR results on the forward error of the pairwise summation to the forward error of both pairwise textbook and pairwise two-pass algorithms (ie. the two-pass and textbook algorithms in which sums are computed pairwise). The following theorem derives a probabilistic bound for the pairwise textbook using the BC method.

THEOREM 6.1. *For the pairwise textbook algorithm, for all* $0<\lambda<1$*, the computed* $\widehat{y}$ *in Equation (4.1) satisfies under SR-nearness*

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

with probability at least $1-\lambda$.

Proof. Equation (5.1) states that $|\widehat{y}-y| \le \left|\sum_{i=1}^{n} x_i^2(\psi_i-1)\right| + \frac{1}{n}\mathcal{B}$. Since the sum is pairwise, the term $\prod_{k=\max\{2,i\}}^{n+1}(1+\eta_k)$ in ψ_i can be replaced with a term $\prod_{k=1}^{\log(n)}(1+\eta_k)$ as shown in Section 3. Thus:

$$\left|\sum_{i=1}^{n} x_i^2(\psi_i-1)\right| \le \left\|x\right\|_2^2\sqrt{2\gamma_{\log(n)+1}(u^2)/\lambda} \quad \text{with probability at least } 1-\frac{\lambda}{2},$$

$$|\widehat{s}-s| \le \left\|x\right\|_1\sqrt{2\gamma_{\log(n)}(u^2)/\lambda} \quad \text{with probability at least } 1-\frac{\lambda}{2}.$$

Since, $|\psi_{n+1}| \leq (1+u)^3$ and $|s| \leq \|x\|_1$, we have with probability at least $1 - \frac{\lambda}{2}$

$$\begin{aligned}\mathcal{B} &\leq (1+u)^3 \|x\|_1^2 \left(2\gamma_{\log(n)}(u^2)/\lambda + 2\sqrt{2\gamma_{\log(n)}(u^2)/\lambda}\right) + \|x\|_1^2 \left((1+u)^3 - 1\right) \\ &= (1+u)^3 \|x\|_1^2 \left(2\gamma_{\log(n)}(u^2)/\lambda + 2\sqrt{2\gamma_{\log(n)}(u^2)/\lambda} + 1\right) - \|x\|_1^2 \\ &= (1+u)^3 \|x\|_1^2 \left(\sqrt{2\gamma_{\log(n)}(u^2)/\lambda} + 1\right)^2 - \|x\|_1^2 .\end{aligned}$$

Finally, Lemma 2.1 shows that with probability at least $1 - \lambda$,

$$\begin{aligned}\frac{|\widehat{y} - y|}{|y|} &\leq \frac{1}{|y|}\left|\sum_{i=1}^n x_i^2(\psi_i - 1)\right| + \frac{1}{n|y|}\mathcal{B} \\ &\leq \mathcal{K}_2^2\sqrt{2\gamma_{\log(n)+1}(u^2)/\lambda} + \mathcal{K}_1^2\left((1+u)^3\left(\sqrt{2\gamma_{\log(n)}(u^2)/\lambda} + 1\right)^2 - 1\right). \quad \square\end{aligned}$$

The following theorem shows the probabilistic bound for the pairwise textbook algorithm using the AH method.

THEOREM 6.2. *For the pairwise textbook algorithm, for all $0 < \lambda < 1$, the computed $\widehat{y}$ in Equation (4.1) satisfies under SR-nearness*

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

with probability at least $1 - \lambda$.

Proof. Equation (5.1) states that $|\widehat{y} - y| \leq \left|\sum_{i=1}^n x_i^2(\psi_i - 1)\right| + \frac{1}{n}\mathcal{B}$. Moreover, Section 3 shows

$$\left|\sum_{i=1}^n x_i^2(\psi_i - 1)\right| \leq \|x\|_2^2\sqrt{u\gamma_{2(\log(n)+1)}(u)}\sqrt{\ln(4/\lambda)} \text{ with probability at least } 1 - \frac{\lambda}{2},$$

$$|\widehat{s} - s| \leq \|x\|_1\sqrt{u\gamma_{2\log(n)}(u)}\sqrt{\ln(4/\lambda)} \text{ with probability at least } 1 - \frac{\lambda}{2}.$$

As the previous proof, we can show that with probability at least $1 - \frac{\lambda}{2}$,

$$\mathcal{B} \leq (1+u)^3 \|x\|_1^2 \left(\sqrt{u\gamma_{2\log(n)}(u)}\sqrt{\ln(4/\lambda)} + 1\right)^2 - \|x\|_1^2 .$$

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

$$\begin{aligned}\frac{|\widehat{y} - y|}{|y|} \leq& \mathcal{K}_2^2\sqrt{u\gamma_{2(\log(n)+1)}(u)}\sqrt{\ln(4/\lambda)} \\ &+ \mathcal{K}_1^2\left((1+u)^3\left(\sqrt{u\gamma_{2\log(n)}(u)}\sqrt{\ln(4/\lambda)} + 1\right)^2 - 1\right). \quad \square\end{aligned}$$

Similar bounds are reached for the pairwise two-pass using the same methods.

7. Error bound analysis. Table 1 shows the asymptotic forward error bounds for the textbook algorithm. Higher order terms in u have been dropped when $nu \ll 1$ and uniquely for BC when $nu \gg 1$ and $nu^2 \ll 1$, and only dominant terms are

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

Table 1: The asymptotic behavior of the textbook forward error bounds for a fixed probability λ and over n up to a constant.

shown. The results in the table are based on: $\gamma_n(u) \approx nu + O(nu^2)$ and $\sqrt{u\gamma_n(u)} \approx \sqrt{\gamma_n(u^2)} \approx \sqrt{n}u + O(nu^2)$ when $nu \ll 1$. $\gamma_n(u) \approx e^{nu}$, $\sqrt{u\gamma_n(u)} \approx \sqrt{u}e^{\frac{n}{2}u}$ and $\sqrt{\gamma_n(u^2)} \approx \sqrt{n}u + O(nu^2)$ when $nu \gg 1$ and $nu^2 \ll 1$.

This table displays the advantage of the probabilistic bounds of the textbook forward error in terms of $O(\sqrt{n}u)$ compared to the deterministic bounds in $O(nu)$, when $nu \ll 1$. Additionally, the BC method is far better when $nu \gg 1$ and $nu^2 \ll 1$. The previous discussion also holds for the two-pass forward error bounds.

7.1. Numerical experiments. We performed a series of numerical experiments comparing these new probabilistic bounds to the deterministic ones. We show that probabilistic bounds are tighter and accurately reflect the behavior of SR-nearness forward errors. Two types of plots are presented. Firstly, the plots are displayed over n and show that for large values of n, BC bounds provide significant benefits compared to AH or DM bounds for the textbook algorithm. Secondly, the plots are shown over λ, and show that AH bound holds a significant advantage for higher probabilities. All SR computations are repeated 30 times with verificarlo [7]. All samples and the forward error of the average of the 30 SR instances are plotted.

7.1.1. Textbook algorithm. We present a numerical application of the textbook algorithm for floating-points chosen uniformly at random between 0 and 1.

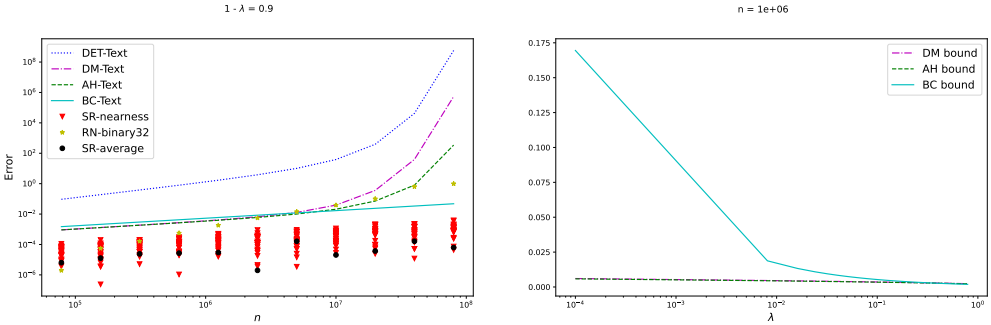

Fig. 2: Probabilistic error bounds over n with probability $1 - \lambda = 0.9$ (left) and over λ with $n = 10^6$ (right) vs deterministic bound for the textbook algorithm.

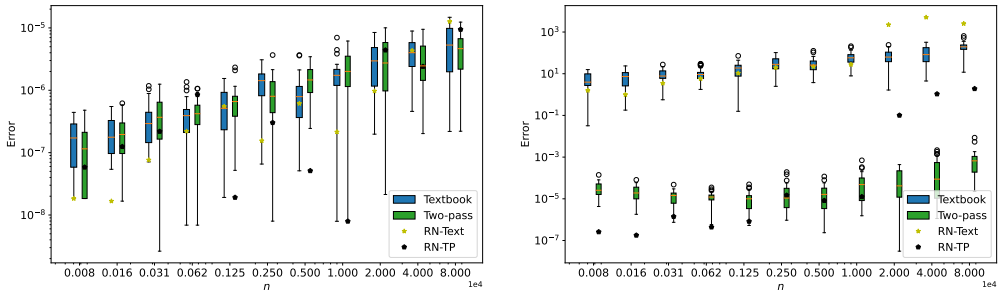

Fig. 3: The forward errors of textbook and two-pass algorithms in binary32 precision for floating-points chosen uniformly at random in $[-1; 1]$ (left) and $[1024; 1025]$ (right).

In Figure 2, triangles represent 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. Interestingly, for small n, the left figure shows that AH, DM, and BC bounds are comparable with a slight advantage for AH-Text and DM. However, as shown in Table 1, when $nu \gg 1$, AH and DM bounds grow exponentially faster than BC bound.

As expected, for a fixed n, the figure on the right shows that the three bounds are close for a probability around 0.9. Nevertheless, AH and DM bounds are more accurate for higher probabilities than BC bound.The result is unsurprising because, generally, Azuma-Hoeffding inequality provides a bound for the deviation of the sum of a sequence of independent and bounded random variables, martingales in this instance, which gives tighter bounds for higher probabilities. In contrast, Bienaymé–Chebyshev inequality is a less restrictive result that provides an upper bound for the probability of deviation between the mean of a distribution and a particular value. The two-pass algorithm exhibits analogous boundary behavior.

7.1.2. Textbook against two-pass. We now compare the forward errors of both algorithms under SR. In figure 3, when the floating-point numbers are randomly chosen with zero mean distribution (left), the absorption errors cancel each other out because both positive and negative errors are uniformly distributed. Therefore, the computed mean is close to zero with low absolute error, and the two-pass algorithm degenerates into the textbook algorithm. Interestingly, this effect is captured by the theoretical bounds because the condition term $\mathcal{K}_2^2 + 2\mathcal{K}_1^2$ becomes smaller for zero-mean distributions. This is confirmed by the experiment in the left figure, which shows a similar forward error for the two algorithms, whether for SR or RN.

As expected, the figure on the right illustrates that when random floating-point numbers are uniformly selected from the interval $[1024, 1025]$, the two-pass algorithm outperforms the textbook algorithm using SR or RN. The mean centering in the two-pass algorithm avoids cancellations and increases its accuracy. While the quantities $\sum_{i=1}^{n} x_i^2$ and $\frac{1}{n}s^2$ are inevitably very large and have the same order of magnitude, their subtraction yields a loss of significant digits in the result, which can compromise the accuracy of the textbook outcome. It is evident from this figure that the use of SR avoids stagnation for $n \geq 10^4$.

8. Conclusion. Many computations are non-linear in various fields such as numerical analysis. In this paper, we have chosen variance computation as an example. In 1983, Chan, Golub, and LeVeque investigated the forward error of variance computation algorithms using RN. To the best of our knowledge, this is the first theoretical study of this problem using stochastic rounding as well as of any algorithm with non-linear errors. In this paper, we have presented probabilistic bounds for two variance computation algorithms that exhibit non-linear errors under SR.

Two methods are used to estimate the forward error of computations: the BC method, which is suitable for large problem sizes n, and the AH method, which is preferable for higher probabilities. The study demonstrates that using SR results in probabilistic bounds on the forward error proportional to $\sqrt{n}u$, which is better than the deterministic bound in $O(nu)$ when using the default rounding mode.

While introducing pairwise algorithm in summation, textbook, and two-pass algorithms, SR leads to probabilistic bounds proportional to $\sqrt{\log(n)}u$, instead of $O(\log(n)u)$ for RN. We also demonstrate that the two-pass algorithm performs better than the textbook algorithm under SR, as it does under RN.

A new approach based on the Doob-Meyer decomposition has been proposed as an alternative method to AH for non-linear SR computations. Our proposed approach contributes to developing new methodologies to bound the algorithms forward error under SR. Though asymptotically in n, this approach is equivalent to the previous two methods, we believe that it can be extended to other algorithms.

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

Acknowledgments. This research was supported by the InterFLOP (ANR-20-CE46-0009) project of the French National Agency for Research (ANR).

REFERENCES

[1] *IEEE standard for floating-point arithmetic*, IEEE Std 754-2019 (Revision of IEEE 754-2008), (2019).

[2] K. Azuma, *Weighted sums of certain dependent random variables*, Tôhoku Mathematical Journal, 19 (1967), p. 357–367, https://doi.org/10.2748/tmj/1178243286.

[3] T. F. Chan, G. H. Golub, and R. J. LeVeque, *Algorithms for computing the sample variance: Analysis and recommendations*, The American Statistician, 37 (1983), pp. 242–247.

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

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

[6] D. Dacunha-Castelle, D. McHale, and M. Duflo, *Probability and Statistics: Volume II*, Springer New York, 2012.

[7] 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.

[8] E.-M. El Arar, D. Sohier, P. d. O. Castro, and E. Petit, *Stochastic rounding variance and probabilistic bounds: A new approach*, arXiv preprint arXiv:2207.10321, (2022).

[9] E.-M. El Arar, D. Sohier, P. de Oliveira Castro, and E. Petit, *The positive effects of stochastic rounding in numerical algorithms*, in 2022 IEEE 29th Symposium on Computer Arithmetic (ARITH), IEEE, 2022, pp. 58–65.

[10] 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.

[11] E. Hallman and I. C. Ipsen, *Precision-aware deterministic and probabilistic error bounds for floating point summation*, arXiv preprint arXiv:2203.15928, (2022).

[12] W. Hoeffding, *Probability inequalities for sums of bounded random variables*, Journal of the American Statistical Association, 58 (1963), p. 13–30, https://doi.org/doi:10.2307/

2282952.

[13] I. C. F. Ipsen and H. Zhou, *Probabilistic error analysis for inner products*, SIAM Journal on Matrix Analysis and Applications, (2020).

[14] E. A. Paxton, M. Chantry, M. Klöwer, L. Saffin, and T. Palmer, *Climate modeling in low precision: Effects of both deterministic and stochastic rounding*, Journal of Climate, (2022).