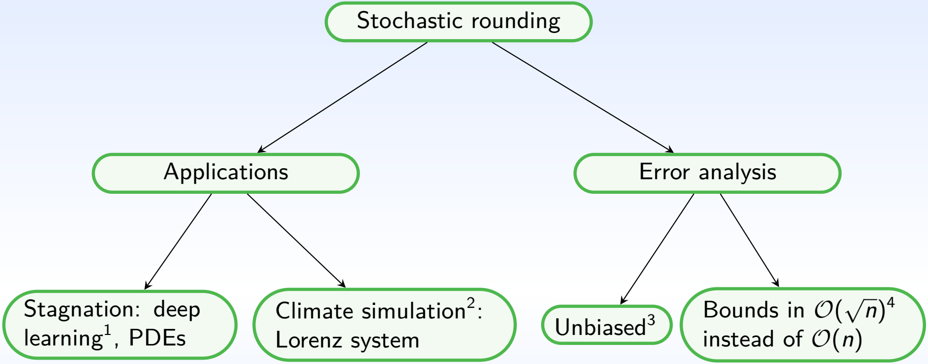
Probabilistic error analysis of limited-precision stochastic rounding

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

PASC25: Massimiliano Fasi, Silviu-Ioan Filip, Mantas Mikaitis

16/06/2025

Motivation



[1] Gupta et al: Deep Learning with Limited Numerical Precision

[2] Paxton et al: Climate Modeling in Low Precision: Effects of Both Deterministic and Stochastic Rounding

[3] Parker: Monte Carlo Arithmetic: Exploiting Randomness in Floating-Point Arithmetic

[4] **El Arar:** Stochastic models for the evaluation of numerical errors

- For $x, y \in \mathbb{R}$ and $\mathrm{op} \in \{+, -, \times, /\}$

$$\mathrm{SR}_p(x) = x(1+\delta), \qquad |\delta| \leqslant u_p$$
$$\mathrm{SR}_p(x \,\mathrm{op}\, y) = (x \,\mathrm{op}\, y)(1+\beta), \qquad |\beta| \leqslant u_p$$

- Upward rounding $\lceil\!\lceil x \rceil\!\rceil$ and downward rounding $\lfloor\!\lfloor x \rfloor\!\rfloor$:

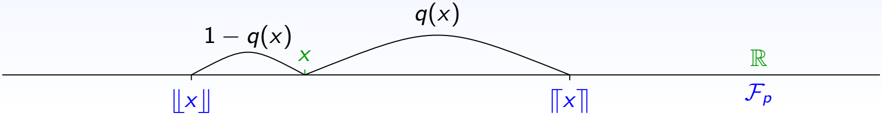


FIGURE. *SR_p with* $q(x) = \frac{x - \lfloor\!\lfloor x \rfloor\!\rfloor}{\lceil\!\lceil x \rceil\!\rceil - \lfloor\!\lfloor x \rfloor\!\rfloor}$

- $\mathbb{E}\left(\mathrm{SR}_p(x)\right) = q(x)\lceil\!\lceil x \rceil\!\rceil + (1 - q(x))\lfloor\!\lfloor x \rfloor\!\rfloor = x$, then $\mathbb{E}(\delta) = \mathbb{E}\left(\frac{\mathrm{SR}_p(x) - x}{x}\right) = 0$

- Using SR_p, for $x_1, x_2, x_3 \in \mathcal{F}_p$ and $\mathrm{op}_1, \mathrm{op}_2 \in \{+, -, *, /\}$

$$c = x_1 \,\mathrm{op}_1\, x_2 \,\mathrm{op}_2\, x_3 \quad \Longrightarrow \quad \mathrm{SR}_p(c) = \left((x_1 \,\mathrm{op}_1\, x_2)(1+\delta_1)\,\mathrm{op}_2\, x_3\right)(1+\delta_2),$$

  with $\mathbb{E}(\delta_1) = \mathbb{E}(\delta_2) = 0$
- Mean independence: $\mathbb{E}[X_k / X_1, ..., X_{k-1}] = \mathbb{E}(X_k)$ for all k
- Independence $\Longrightarrow$ **Mean independence** $\Longrightarrow$ uncorrelatedness

Lemma 1 (M. P. Connolly et al.).

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

SR and mean independence

- Using SR_p, for $x_1, x_2, x_3 \in \mathcal{F}_p$ and $\mathrm{op}_1, \mathrm{op}_2 \in \{+, -, *, /\}$

$$c = x_1 \,\mathrm{op}_1\, x_2 \,\mathrm{op}_2\, x_3 \quad \Longrightarrow \quad \mathrm{SR}_p(c) = ((x_1 \,\mathrm{op}_1\, x_2)(1+\delta_1) \,\mathrm{op}_2\, x_3)\,(1+\delta_2),$$

  with $\mathbb{E}(\delta_1) = \mathbb{E}(\delta_2) = 0$
- Mean independence: $\mathbb{E}[X_k / X_1, ..., X_{k-1}] = \mathbb{E}(X_k)$ for all k
- Independence $\Longrightarrow$ **Mean independence** $\Longrightarrow$ uncorrelatedness

Lemma 1 (M. P. Connolly et al.).

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

- Martingale and Azuma-Hoeffding inequality: AH bound
- Bound of the variance and Chebyshev inequality: BC bound
- Inner product, Horner's polynomial, statistical variance...

Example: inner product

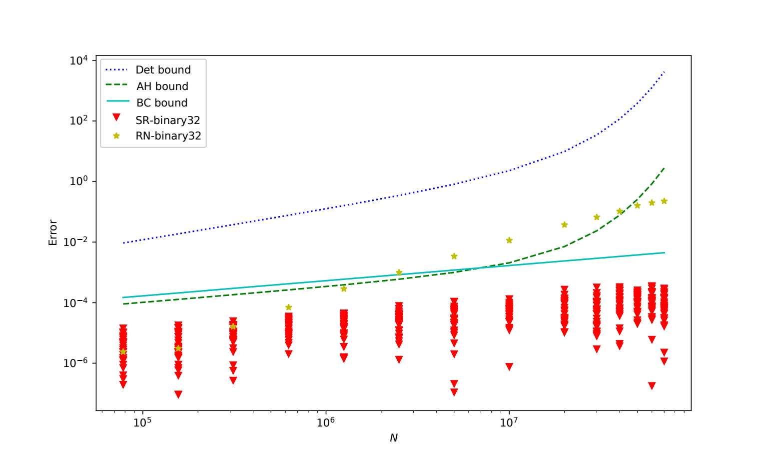


FIGURE. *Probabilistic bounds with probability at least* 0.95 *vs deterministic bound of the computed forward errors of the inner product for floating-point numbers chosen uniformly at random in* $[0;1]$ *with* $u_p = 2^{-23}$

Example: Rosenbrock function

The Rosenbrock function is a non-convex function defined by

$$f(x_1, x_2) = (1 - x_1)^2 + 100(x_2 - x_1^2)^2$$

with a global minimum of 0, occurring at $\mathbf{x}^\star = (1, 1)$.

The gradient descent:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - t_k \nabla f(\mathbf{x}_k)$$

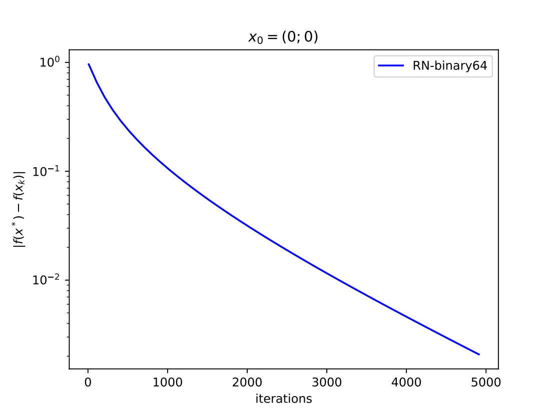


FIGURE. *Convergence profiles for 5000 iterations of gradient descent on the Rosenbrock function, with learning rate $t_k = 0,001$.*

`https://github.com/MehdiElArar/srfloat`

Example: Rosenbrock function

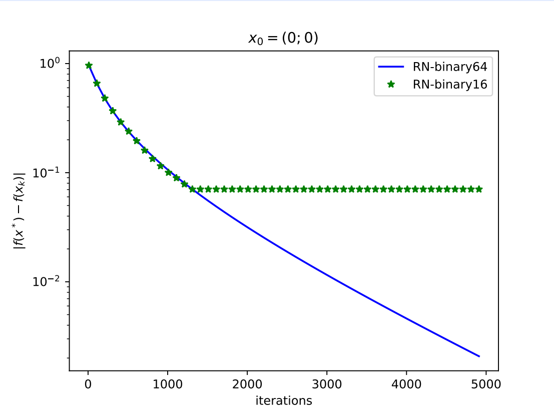


FIGURE. *Convergence profiles for 5000 iterations of gradient descent on the Rosenbrock function, with learning rate* $t_k = 0,001$.

`https://github.com/MehdiElArar/srfloat`

Example: Rosenbrock function

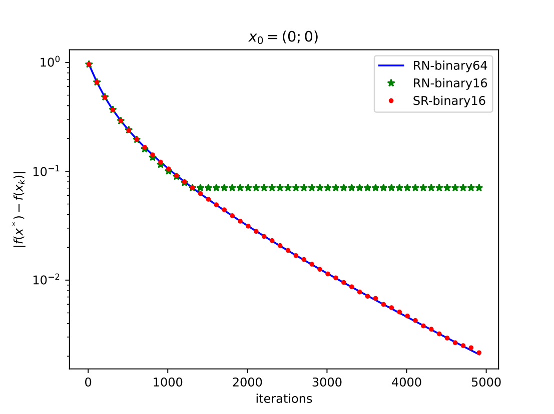


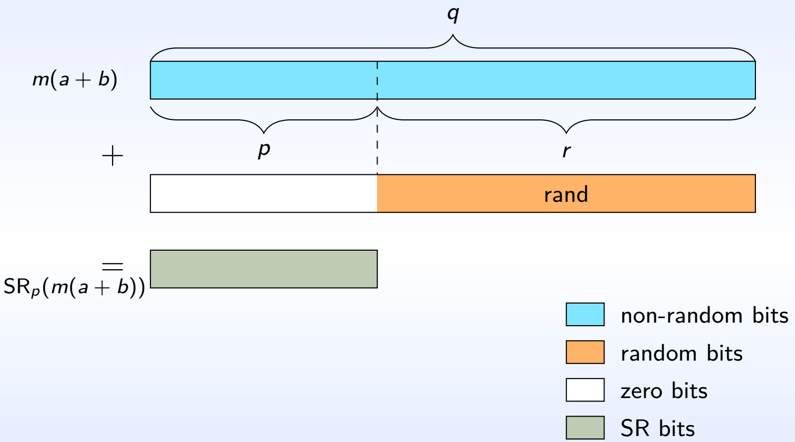
FIGURE. *Convergence profiles for 5000 iterations of gradient descent on the Rosenbrock function, with learning rate $t_k = 0,001$.*

https://github.com/MehdiElArar/srfloat

How can we implement this in hardware?

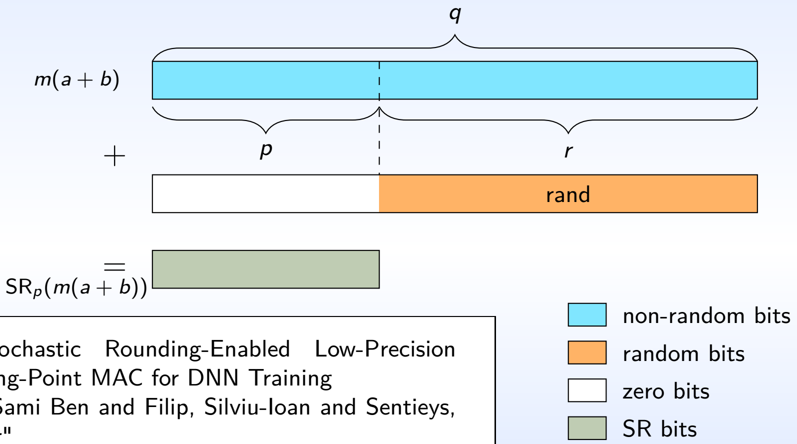
Addition

Let $a, b \in \mathcal{F}_p$, if we compute $a + b$ in $\mathcal{F}_q$ such that $p < q$, we take $r = q - p$.



Addition

Let $a, b \in \mathcal{F}_p$, if we compute $a + b$ in $\mathcal{F}_q$ such that $p < q$, we take $r = q - p$.



A Stochastic Rounding-Enabled Low-Precision Floating-Point MAC for DNN Training
"Ali, Sami Ben and Filip, Silviu-Ioan and Sentieys, Olivier"

Toy example: $a + b$

$$a = 1.111_{(2)} \cdot 2^8 \qquad\qquad a = 1.111000_{(2)} \cdot 2^8$$
$$b = 1.101_{(2)} \cdot 2^5 \xrightarrow{\text{significand alignment}} b = \underline{0.001101_{(2)} \cdot 2^8}$$

$$\text{significand addition} \qquad a + b = 10.000101_{(2)} \cdot 2^8$$

$$\Big\downarrow \text{normalization}$$

$$a + b = 1.0000101_{(2)} \cdot 2^9$$

$$\Big\downarrow \text{add random bits}$$

$$a + b = 1.0000101_{(2)} \cdot 2^9$$
$$\text{rand} = \underline{0.0001101_{(2)} \cdot 2^9} \quad \text{round}$$

$$\mathrm{SR}(a + b) = 1.001_{(2)} \cdot 2^9$$

Toy example: $a + b$

$$a = 1.111_{(2)} \cdot 2^8 \qquad b = 1.101_{(2)} \cdot 2^5$$

$$\xrightarrow{\text{significand alignment}}$$

$$a = 1.111000_{(2)} \cdot 2^8 \qquad b = \underline{0.001101_{(2)} \cdot 2^8}$$

significand addition

$$a + b = 10.000101_{(2)} \cdot 2^8$$

$$\Big\downarrow \text{ normalization}$$

$$a + b = 1.0000101_{(2)} \cdot 2^9$$

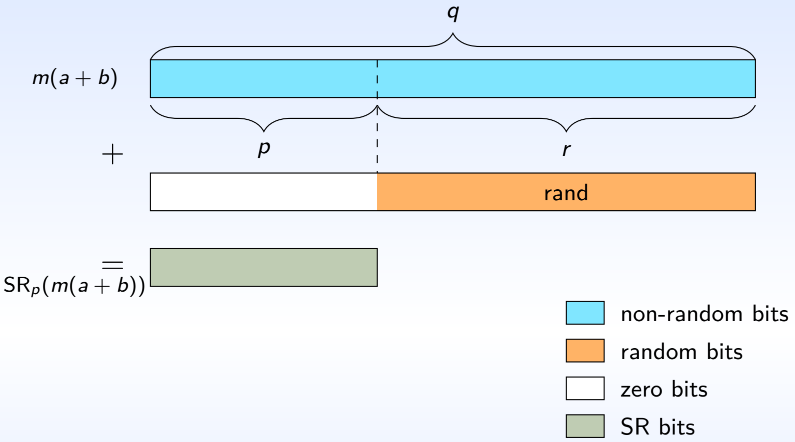
$$\Big\downarrow \text{ add random bits}$$

$$a + b = 1.0000101_{(2)} \cdot 2^9$$

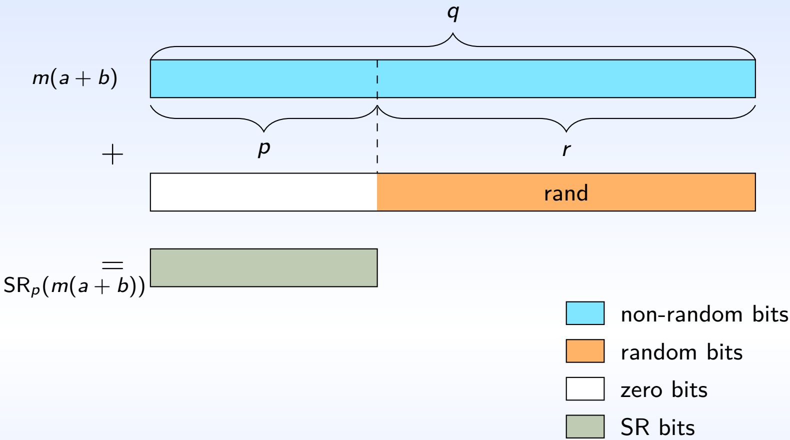
$$\text{rand} = \underline{0.000{\color{red}0110}_{(2)} \cdot 2^9} \quad \text{round}$$

$$\mathrm{SR}(a + b) = 1.00{\color{red}0}_{(2)} \cdot 2^9$$

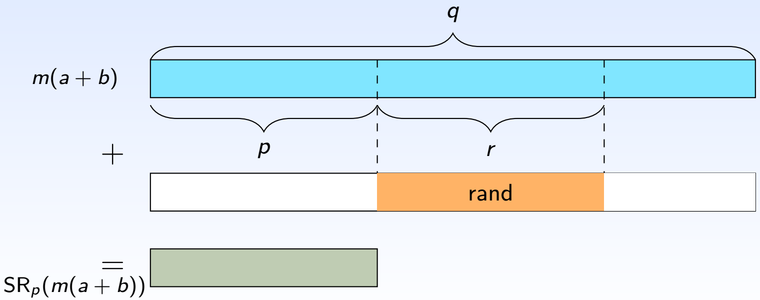
How can we implement this in hardware?

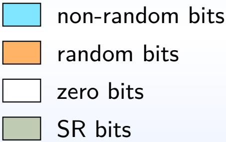


How can we implement this in hardware?



Expensive!!





How does the behavior of SR_p change when r is not sufficiently large?

Expensive!!

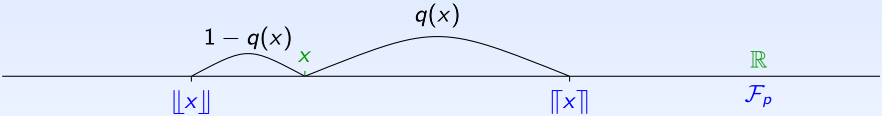


FIGURE. SR_p with $q(x) = \frac{x - \llbracket x \rrbracket}{\llceil x \rrceil - \llbracket x \rrbracket}$

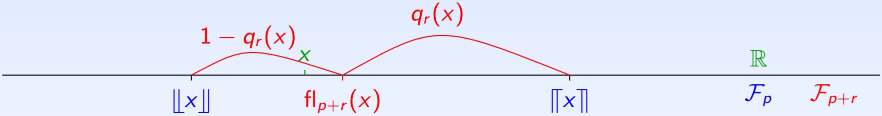


FIGURE. $SR_{p,r}$ *with* $q_r(x) = \frac{\mathrm{fl}_{p+r}(x) - \llbracket x \rrbracket}{\lceil\!\lceil x \rceil\!\rceil - \llbracket x \rrbracket}$

Limited-precision stochastic rounding

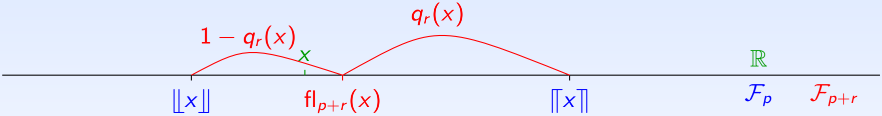


FIGURE. *$SR_{p,r}$ with $q_r(x) = \frac{\mathrm{fl}_{p+r}(x) - \llbracket x \rrbracket}{\lceil\!\lceil x \rceil\!\rceil - \llbracket x \rrbracket}$*

- $\mathrm{SR}_{p,r}(x) = x(1+\delta),\ |\delta| \leqslant u_p \quad \neq \quad \mathrm{fl}_{p+r}(x) = x(1+\beta),\ |\beta| \leqslant u_{p+r}$
- $\mathbb{E}\left(\mathrm{SR}_{p,r}(x)\right) = q_r(x)\lceil\!\lceil x \rceil\!\rceil + (1 - q_r(x))\llbracket x \rrbracket = \mathrm{fl}_{p+r}(x)$, then

$$\mathbb{E}(\delta) = \mathbb{E}\left(\frac{\mathrm{SR}_{p,r}(x) - x}{x}\right) = \beta$$

- The mean independence is lost

$$\mathbb{E}(\delta_k \mid \delta_1, \ldots, \delta_{k-1}) = \beta_k \neq \mathbb{E}(\delta_k)$$

- β_k is a random variable and $\mathbb{E}(\beta_k) = \mathbb{E}(\delta_k)$

Doob–Meyer decomposition

Lemma 2.

Let $\delta_1, \delta_2, \ldots, \delta_n$ be random errors produced by a sequence of elementary operations using $\mathrm{SR}_{p,r}$, and let $\beta_1, \beta_2, \ldots, \beta_n$ be their corresponding errors incurred by fl_{p+r}. Then, the random variables $\alpha_k = \delta_k - \beta_k$ for $1 \leqslant k \leqslant n$ are mean independent

$$\mathbb{E}(\alpha_k \mid \alpha_1, \ldots, \alpha_{k-1}) = \mathbb{E}(\alpha_k) = 0.$$

Moreover, for all $1 \leqslant i \leqslant n$

$$\prod_{k=i}^{n}(1+\delta_k) = \prod_{k=i}^{n}(1+\alpha_k) + \mathcal{B}_i,$$

with

$$|\mathcal{B}_i| \leqslant \gamma_{n-i+1}(u_p + u_{p+r}) - \gamma_{n-i+1}(u_p)$$

Theorem 3.

For $y = \sum_{i=1}^{n} a_i b_i$ and $0 < \lambda < 1$, the quantity $SR_{p,r}(y)$ satisfies

$$\frac{|SR_{p,r}(y) - y|}{|y|} \leqslant \kappa(a \circ b) \left(\sqrt{u_p \gamma_{2n}(u_p)} \sqrt{\ln(2/\lambda)} + \gamma_n(u_p + u_{p+r}) - \gamma_n(u_p) \right)$$

$$= \kappa(a \circ b) \left(\sqrt{2n} \sqrt{\ln(2/\lambda)} u_p + n u_{p+r} \right) + \mathcal{O}(\|(u_p, u_{p+r})\|_2)$$

with probability at least $1 - \lambda$.

Lemma 4.

Theorem 3 leads to

$$\mathcal{O}(\sqrt{n}u_p + nu_{p+r})$$

we want

$$\sqrt{n}u_p > nu_{p+r}$$

we thus have this good rule of thumb

$$r \geqslant \lceil (\log_2 n)/2 \rceil$$

Numerical experiments

- Rosenbrock function **(srfloat)**
- Parameter update in deep neural network training **(MPTorch)**

- `https://github.com/MehdiElArar/srfloat`
- `https://github.com/mptorch/mptorch`

Rosenbrock function

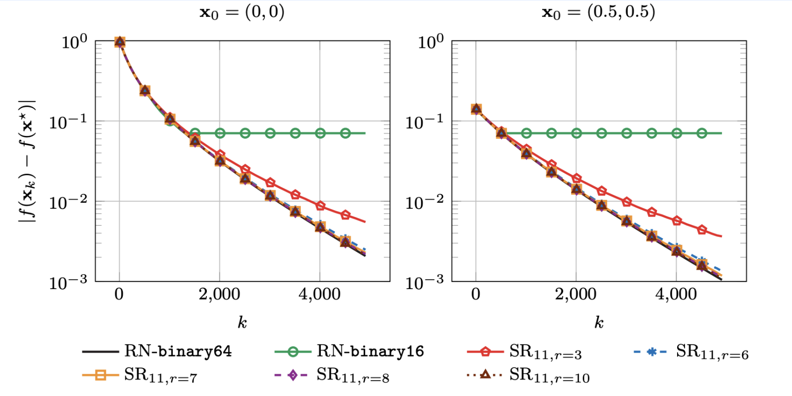


FIGURE. *Convergence profiles for 5,000 iterations of gradient descent on the Rosenbrock function. For both experiments, we average each* $SR_{11,r}$ *error over* 500 *different runs, and the learning rate is* $t_k = 0.001$.

Rosenbrock function

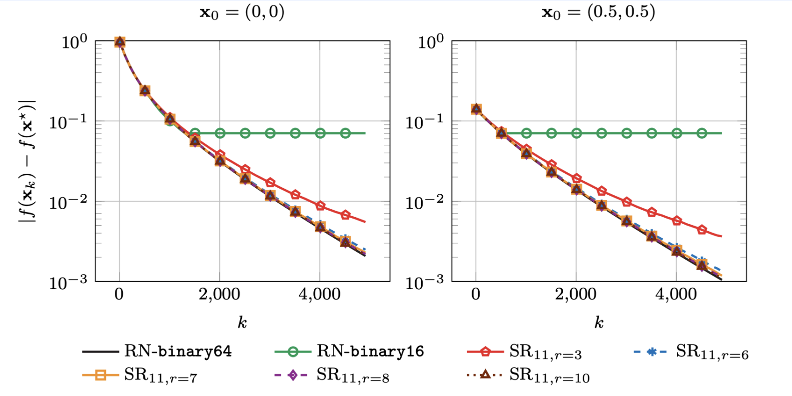


FIGURE. *Convergence profiles for 5,000 iterations of gradient descent on the Rosenbrock function. For both experiments, we average each $SR_{11,r}$ error over* 500 *different runs, and the learning rate is $t_k = 0.001$.*

$$\lceil \log_2(5{,}000)/2 \rceil = 7$$

Parameter update in deep neural network training

Focus:

Training a ResNet32[1] model on the CIFAR-10 dataset

Training Setup:

- **Hyperparameters:**
  - **Batch Size:** 128, **Momentum:** $\mu = 0.9$
  - **Total Training:** 64,000 iterations (200 epochs)
  - **Learning Rate:** $t_k = 0.1$, reduced by 10 at 32,000 and 48,000 iterations

Numerical Precision:

- **Arithmetic:** `bfloat16` ($p = 8$)
- **Update Rule:**

$$
\begin{aligned}
\mathbf{v}_{k+1} &= \circ(\mu \mathbf{v}_k + \mathbf{g}_k), \\
\mathbf{x}_{k+1} &= \circ(\mathbf{x}_k - t_k \mathbf{v}_{k+1})
\end{aligned}
$$

- **Components:**
  - $\mathbf{v}_k$: Velocity vector
  - $\mathbf{g}_k$: Gradient of the loss function

[1] Deep Residual Learning for Image Recognition

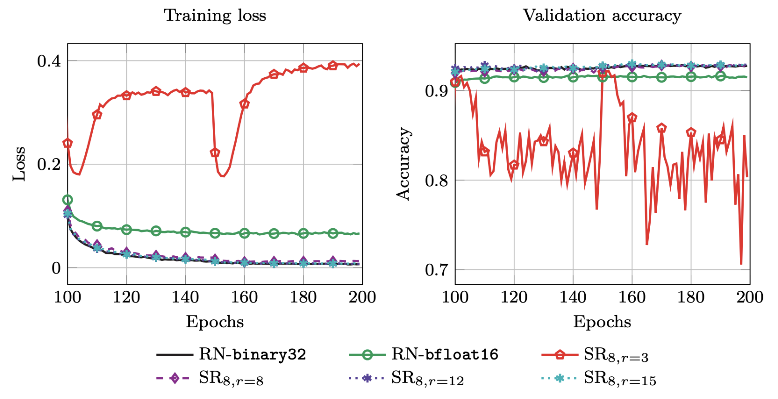


FIGURE. *In the baseline configuration,* `binary32` *arithmetic with RN is used for computing, and the same format is used for storage. For the low-precision configurations, parameters are stored and updated using* `bfloat16` *arithmetic with either RN or* $SR_{p,r}$.

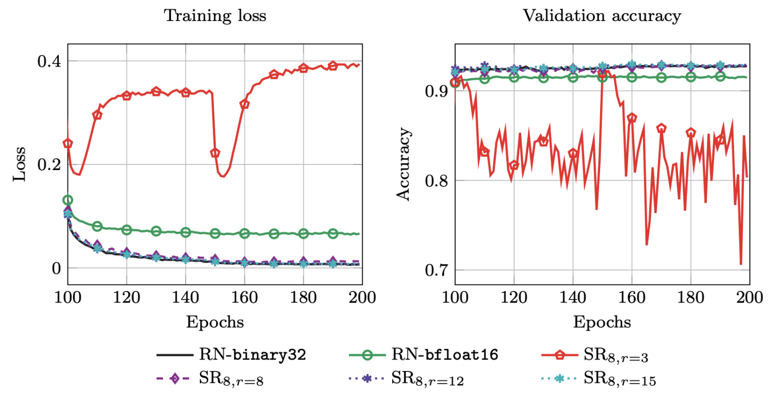


FIGURE. *In the baseline configuration,* `binary32` *arithmetic with RN is used for computing, and the same format is used for storage. For the low-precision configurations, parameters are stored and updated using* `bfloat16` *arithmetic with either RN or* $SR_{p,r}$.

$$\lceil \log_2(64{,}000)/2 \rceil = 8$$

Summary

	SR_p	$\mathrm{SR}_{p,r}$
Unbiased	✓	✗
Mean independence	✓	✗
Probabilistic bound	$\mathcal{O}(\sqrt{n}u_p)$	$\mathcal{O}(\sqrt{n}u_p + nu_{p+r})$
Rule of thumb		$r \geqslant \lceil (\log_2 n)/2 \rceil$

TABLE. *Classic stochastic rounding versus limited-precision stochastic rounding*

To appear in SIAM Journal on Scientific Computing:
https://arxiv.org/abs/2408.03069

Probabilistic error analysis of limited-precision stochastic rounding

El-Mehdi El Arar, Massimiliano Fasi, Silviu-Ioan Filip, Mantas Mikaitis

Hardware with SR

- Graphcore IPU supports SR for binary32 and binary16 arithmetic
- AWS Trainium chips supports SR binary16 or bfloat16 arithmetic (at least additions)
- AMD MI300 GPUs supports SR down-conversion of binary32 values to 8-bit
- Tesla D1 supports SR down-conversion of binary32 values to 8-bit
- Blackwell architecture, NVIDIA GPUs support SR down-conversion from binary32 to 16, 8, 6, and 4-bit values

Example: no stagnation

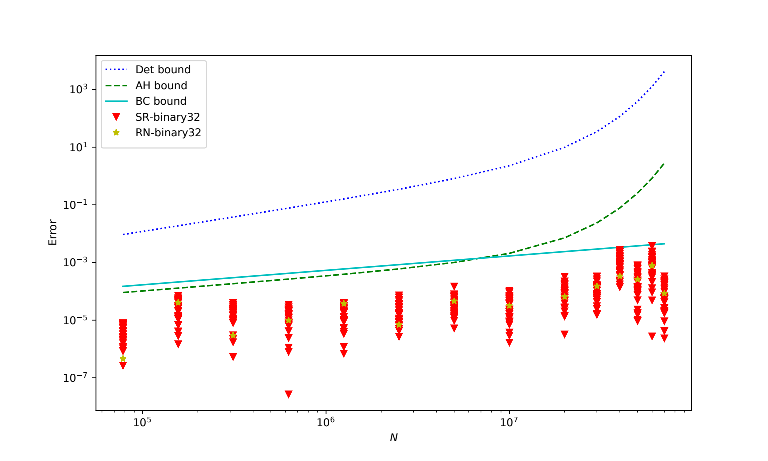


FIGURE. *Probabilistic bounds with probability* $1 - \lambda = 0.95$ *vs deterministic bound of the computed forward errors of the inner product for floating-point numbers chosen uniformly at random in* $[-1; 1]$ *with* $u_p = 2^{-23}$