Exercise 6: Rounding Error Accumulation in Matrix-Vector Products

Problem Statement

Analyze and discuss the figure below, which compares the backward error of the matrix-vector product $y = Ax$ computed in fp16 arithmetic, where the coefficients of $A \in \mathbb{R}^{n\times n}$ and x have been randomly generated in $[0, 1]$, for three different rounding modes: round-to-nearest (RTN), round-to-zero (RZ), and stochastic rounding (SR).

Solution

Analysis of the Figure

The figure displays backward error as a function of matrix dimension n on a log-log plot. Three rounding modes are compared against two theoretical growth curves: nu (solid line) and $\sqrt{nu}$ (dashed line), where u is the unit roundoff for fp16 arithmetic ($u \approx 4.88 \times 10^{-4}$).

Theoretical Background

For the matrix-vector product $y = Ax$, each component y_i is computed as:

$$\hat{y}_i = \sum_{j=1}^{n} a_{ij}x_j \tag{1}$$

In finite precision arithmetic, this sum accumulates rounding errors. The backward error measures how much we must perturb the input data (A, x) so that the computed result $\hat{y}$ is the exact result for the perturbed problem.

Expected Error Growth Rates

Linear growth (nu): In the worst case, rounding errors can accumulate coherently (all with the same sign). For a sum of n products, the maximum possible error accumulation is:

$$\text{Backward error} \leq nu + O(u^2) \tag{2}$$

This represents the **pessimistic bound** where all rounding errors conspire in the same direction.

Square root growth ($\sqrt{nu}$): When rounding errors behave as independent random variables with zero mean, the central limit theorem suggests they should accumulate in a root-mean-square fashion:

$$\text{Backward error} \approx \sqrt{n}u + O(u^2) \tag{3}$$

This represents the **probabilistic bound** for random error cancellation.

Observations by Rounding Mode

Round-to-Zero (RZ) - Orange Curve

Behavior: The RZ mode exhibits the worst performance, closely following the linear nu growth rate.

Explanation: Round-to-zero is a *directed rounding mode* that always rounds toward zero. This introduces a systematic bias in the rounding errors:

- For positive numbers, RZ always rounds down (negative error)
- For negative numbers, RZ always rounds up (positive error)
- Since A and x have entries in $[0, 1]$, all products $a_{ij}x_j$ are positive
- All rounding errors have the *same sign* (negative)
- Errors accumulate coherently without cancellation

The backward error grows linearly as $\sim nu$ because there is no randomness to cause cancellation of errors. This is the worst-case scenario predicted by standard rounding error analysis.

For large n (around $n \approx 10^4$), the curve appears to saturate near $10^0 = 1$. This saturation occurs because:

$$nu \approx n \cdot 4.88 \times 10^{-4} \geq 1 \quad \text{when} \quad n \geq 2048 \tag{4}$$

At this point, the backward error becomes $O(1)$, meaning the computed result bears little relation to the true answer even in a backward error sense.

Round-to-Nearest (RTN) - Blue Curve

Behavior: RTN initially follows the $\sqrt{nu}$ curve for small n (up to $n \approx 10^3$), then transitions to linear nu growth for larger n.

Explanation: Round-to-nearest is an *unbiased rounding mode*:

- Each rounding error is approximately uniformly distributed in $[-u/2, u/2]$
- For small n, rounding errors are roughly independent random variables
- Random errors cancel partially, leading to $\sqrt{n}$ growth (probabilistic bound)
- For large n, the accumulation of many rounded sums eventually exhausts the available precision
- The transition to linear growth occurs when accumulated errors become $O(1)$ relative to the computation

The transition point around $n \approx 10^3$ to 10^4 represents where:

$$\sqrt{nu} \sim u \quad \Rightarrow \quad \sqrt{n} \cdot 4.88 \times 10^{-4} \sim 10^{-3} \text{ to } 10^{-2} \tag{5}$$

Beyond this point, the limited precision of fp16 ($u \approx 5 \times 10^{-4}$) causes error cancellation to break down, and errors begin to accumulate more systematically.

Stochastic Rounding (SR) - Yellow Curve

Behavior: SR consistently follows the $\sqrt{nu}$ growth rate throughout the entire range of n, even for very large dimensions.

Explanation: Stochastic rounding is a *randomized rounding mode* where:

- Each value x is rounded to the nearest floating-point number probabilistically
- If x lies between $\mathrm{fl}^-(x)$ and $\mathrm{fl}^+(x)$, then:

$$\mathbb{P}[\mathrm{fl}(x) = \mathrm{fl}^+(x)] = \frac{x - \mathrm{fl}^-(x)}{\mathrm{fl}^+(x) - \mathrm{fl}^-(x)} \tag{6}$$

- This makes $\mathbb{E}[\mathrm{fl}(x)] = x$ (unbiased in expectation)
- Rounding errors are *statistically independent* across operations
- Errors remain random and uncorrelated even for large n

The key advantage is that stochastic rounding maintains the probabilistic error cancellation property regardless of problem size. The errors behave as a random walk, giving:

$$\text{Backward error} \approx \sqrt{nu} \tag{7}$$

This is dramatically better than the nu growth of deterministic rounding modes for large n. For example, at $n = 10^5$:

- SR: backward error $\approx \sqrt{10^5} \cdot 5 \times 10^{-4} \approx 0.15$
- RZ: backward error $\approx 10^5 \cdot 5 \times 10^{-4} \approx 50$ (saturated at 1)

Key Insights

1. **Directed rounding is problematic:** Round-to-zero introduces systematic bias that causes worst-case linear error growth. This is particularly severe in low-precision arithmetic like fp16.
2. **Round-to-nearest provides mixed results:** While unbiased, RTN shows $\sqrt{n}$ growth only for moderate n. As n increases and precision is exhausted, the effective randomness decreases and errors accumulate more systematically.
3. **Stochastic rounding is optimal:** SR maintains true randomness in rounding errors, achieving the probabilistic $\sqrt{n}$ bound even for very large problems. This represents a factor of $\sqrt{n}$ improvement over worst-case bounds.
4. **Practical significance for fp16:** With $u \approx 5 \times 10^{-4}$:
   - For $n = 1000$: RTN and SR both give $\sim 1.5 \times 10^{-2}$ backward error
   - For $n = 100000$: SR gives ~ 0.15 while RTN saturates near 1
   - This makes SR especially valuable for large-scale computations in reduced precision
5. **Backward stability interpretation:** An algorithm is backward stable if backward error $= O(u)$. For matrix-vector multiplication with dimension n:
   - With RZ: backward stability is lost immediately as n increases
   - With RTN: backward stability holds up to $n \sim 1/u^2 \approx 4 \times 10^6$ for fp16
   - With SR: backward stability in a probabilistic sense holds for much larger n

Conclusion

This exercise demonstrates that the choice of rounding mode has profound implications for the accuracy of numerical computations, especially in low-precision arithmetic. Stochastic rounding provides superior error characteristics by maintaining statistical independence of rounding errors, achieving the optimal $\sqrt{n}$ error growth predicted by probabilistic analysis. This makes SR particularly attractive for machine learning and other applications using reduced-precision arithmetic (fp16, bfloat16) where large matrix operations are common.

The figure provides compelling empirical evidence for the theoretical prediction that random error cancellation can significantly reduce accumulated rounding error compared to worst-case deterministic bounds.