Lecture 2

Floating-Point Arithmetic and Error Analysis

Virginia Tech Department of Computer Science
Fall 2026 • Sumeet Khatri

Three questions

01

What does the machine compute?

Define each arithmetic operation through storage, exact arithmetic, and one final rounding.

02

Where do digits disappear?

Recognize cancellation, scaling, and the loss of significant digits.

03

What is the error?

Quantify the propagation of errors under addition and multiplication.

Recall the one-step rounding model

Nearest representable value

A real result is mapped to a nearby machine number.

$$\mathrm{fl}(x) = x(1+\delta), \qquad |\delta| \le u$$

The **unit roundoff** bounds the relative error of one correctly rounded result.

$$u = \frac{1}{2} b^{-p+1}$$

A floating-point operation has three stages

1

Store the inputs

Map real inputs into the chosen floating-point system.

2

Operate exactly

Apply the real arithmetic operation to the stored values.

3

Round the result

Map the exact intermediate back into the system.

$$x \circ_{\mathcal{F}} y = \mathrm{fl}_{\mathcal{F}}(\mathrm{fl}_{\mathcal{F}}(x) \circ \mathrm{fl}_{\mathcal{F}}(y))$$

$\mathscr{F} \equiv$ floating-point system

The four floating-point arithmetic operations

$$
\begin{aligned}
x +_{\mathcal{F}} y &:= \mathrm{fl}_{\mathcal{F}}(\mathrm{fl}_{\mathcal{F}}(x) + \mathrm{fl}_{\mathcal{F}}(y))\,, \\
x -_{\mathcal{F}} y &:= \mathrm{fl}_{\mathcal{F}}(\mathrm{fl}_{\mathcal{F}}(x) - \mathrm{fl}_{\mathcal{F}}(y))\,, \\
x \times_{\mathcal{F}} y &:= \mathrm{fl}_{\mathcal{F}}(\mathrm{fl}_{\mathcal{F}}(x)\,\mathrm{fl}_{\mathcal{F}}(y))\,, \\
x \div_{\mathcal{F}} y &:= \mathrm{fl}_{\mathcal{F}}(\mathrm{fl}_{\mathcal{F}}(x) / \mathrm{fl}_{\mathcal{F}}(y))\,.
\end{aligned}
$$

Every operation may inherit input error and introduce one new rounding error.

Error propagation in addition

$$x +_{\mathscr{F}} y = \mathrm{fl}\Big(\mathrm{fl}(x) + \mathrm{fl}(y)\Big) = \Big(\mathrm{fl}(x) + \mathrm{fl}(y)\Big)(1 + \delta_3) = \Big(x(1 + \delta_1) + y(1 + \delta_2)\Big)(1 + \delta_3)$$

$$\frac{\mathrm{fl}\Big(\mathrm{fl}(x) + \mathrm{fl}(y)\Big) - (x + y)}{x + y} = \delta_3 + \frac{x\delta_1 + y\delta_2}{x + y} + \delta_3\frac{x\delta_1 + y\delta_2}{x + y} \approx \delta_3 + \frac{x}{x + y}\delta_1 + \frac{y}{x + y}\delta_2$$

Then the relative error is: $(|\delta_i| \leq u)$

$$\frac{|x +_{\mathscr{F}} y - (x + y)|}{|x + y|} \leq u + \frac{|x| + |y|}{|x + y|}u$$

Note: when $y \approx -x$, the denominator will be small $\Rightarrow$ catastrophic cancellation!

Choose the toy system and inputs

$$\mathcal{F} = (10, 5, -4, 5)$$

$$x = \frac{5}{7}, \qquad y = \frac{1}{3}$$

Rule

Keep five decimal digits and discard everything after the fifth digit (“chopping”).

Question

How accurate are the four basic operations after both inputs are stored?

First store the operands

$$x = \frac{5}{7} = 0.\overline{714285}, \qquad y = \frac{1}{3} = 0.\overline{3}$$

$$\mathrm{fl}(x) = 7.1428 \times 10^{-1}$$

$$\mathrm{fl}(y) = 3.3333 \times 10^{-1}$$

The arithmetic begins with these stored values—not the original fractions.

Addition: compute exactly, then chop

$$
\begin{aligned}
x +_{\mathcal{F}} y &= \mathrm{fl}\left(7.1428 \times 10^{-1} + 3.3333 \times 10^{-1}\right) \\
&= \mathrm{fl}\left(1.04761 \times 10^{0}\right) \\
&= 1.0476 \times 10^{0}.
\end{aligned}
$$

One final digit is discarded after the exact addition of the stored inputs.

Measure the addition error

$$x + y = \frac{22}{21}$$

$$\varepsilon_{\text{abs}} = \left|\frac{22}{21} - 1.0476\right| = 0.190 \times 10^{-4}$$

$$\varepsilon_{\text{rel}} = \frac{0.190 \times 10^{-4}}{22/21} = 0.182 \times 10^{-4}$$

Repeat for all four operations

operation	result	exact value	absolute error	relative error
$x +_{\mathcal{F}} y$	1.0476×10^{0}	$22/21$	0.190×10^{-4}	0.182×10^{-4}
$x -_{\mathcal{F}} y$	3.8095×10^{-1}	$8/21$	0.238×10^{-5}	0.625×10^{-5}
$x \times_{\mathcal{F}} y$	2.3809×10^{-1}	$5/21$	0.524×10^{-5}	0.220×10^{-4}
$x \div_{\mathcal{F}} y$	2.1428×10^{0}	$15/7$	0.571×10^{-4}	0.267×10^{-4}

$$u = \frac{1}{2} b^{1-p} = \frac{1}{2} 10^{-4} = 5 \times 10^{-5}$$

$$\max \varepsilon_{\text{rel}} = 0.267 \times 10^{-4} < u \quad \checkmark$$

Now choose inputs that expose specific pitfalls

$$u_0 = 0.714251, \qquad v = 98765.9, \qquad w = 0.111111 \times 10^{-4}$$

Near-equal pair

Subtract to expose cancellation.

Tiny scale

Divide by a small value to magnify absolute error.

Large scale

Multiply or add to test representational imbalance.

$$\boxed{\mathrm{fl}(x) = 7.1428 \times 10^{-1}}$$

$$\boxed{\mathrm{fl}(y) = 3.3333 \times 10^{-1}}$$

Near-equal subtraction

$$\mathrm{fl}(x) = 0.71428, \qquad \mathrm{fl}(u_0) = 0.71425$$

$$x - u_0 = 0.347143 \times 10^{-4}$$

$$x -_{\mathcal{F}} u_0 = 0.30000 \times 10^{-4}$$

$$|(x - u_0) - (x -_{\mathcal{F}} u_0)| = 0.47143 \times 10^{-5}$$

A small absolute error can be a large relative error

$$\varepsilon_{\text{abs}} = 0.47143 \times 10^{-5}$$

$$|x - u_0| = 0.347143 \times 10^{-4}$$

$$\varepsilon_{\text{rel}} = \frac{0.47143 \times 10^{-5}}{0.347143 \times 10^{-4}} \approx 0.136$$

Later scaling preserves relative error but magnifies absolute error

Divide by the tiny value

$$
\begin{aligned}
(x -_{\mathcal{F}} u_0) \div_{\mathcal{F}} w &= 2.7000, \\
\varepsilon_{\text{abs}} &= 0.424, \\
\varepsilon_{\text{rel}} &= 0.136.
\end{aligned}
$$

Multiply by the large value

$$
\begin{aligned}
(x -_{\mathcal{F}} u_0) \times_{\mathcal{F}} v &= 2.9629, \\
\varepsilon_{\text{abs}} &= 0.465, \\
\varepsilon_{\text{rel}} &= 0.136.
\end{aligned}
$$

Adding a tiny value to a large value can erase the tiny contribution

$$u_0 + v = 98766.641251$$

$$u_0 +_{\mathcal{F}} v = 0.98765 \times 10^5$$

$$\varepsilon_{\text{abs}} = 0.161 \times 10^1$$

$$\varepsilon_{\text{rel}} = 0.163 \times 10^{-4}$$

Different operations expose different error symptoms

Cancellation

Near-equal subtraction

Relative error spikes

Scaling

Tiny divisor or large multiplier

Absolute error grows

Absorption

Small addend beside a large one

Small contribution vanishes

Cancellation makes one weak step control the chain

1

Accurate inputs

Each operand has many correct digits.

2

Fragile subtraction

Leading digits cancel and expose rounded tails.

3

Later arithmetic

Scaling or further operations carry the damaged result.

A chain of calculations is no more accurate than its weakest intermediate step.

Two close inputs can create a loss of significant digits

$$x = 0.3721448693, \qquad y = 0.3720214371$$

$$\hat{x} = 0.37214, \qquad \hat{y} = 0.37202$$

$$x - y = 0.0001234322$$

The computed difference contains only two reliable digits

$$\widehat{x} - \widehat{y} = 0.00012$$

$$\frac{|(x-y) - (\widehat{x} - \widehat{y})|}{|x-y|} = \frac{0.0000034322}{0.0001234322} \approx 3 \times 10^{-2}$$

Five-digit arithmetic does not guarantee five-digit results after cancellation.

The quadratic formula has a vulnerable branch

$$x_{\pm} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

$$x^2 + 62.10x + 1 = 0$$

$$x_+ \approx -0.01610723, \qquad x_- \approx -62.08390$$

Round the discriminant and its square root

$$b^2 - 4ac = (62.10)^2 - 4.000 = 3852$$

$$\sqrt{3852} = 62.06 \quad \text{(four-digit rounding)}$$

$$-b + \sqrt{b^2 - 4ac} = -62.10 + 62.06$$

The small root subtracts nearly equal numbers

$$\mathrm{fl}(x_+) = \frac{-62.10 + 62.06}{2.000} = \frac{-0.04000}{2.000} = -0.02000$$

$$x_+ \approx -0.01611$$

$$\varepsilon_{\mathrm{rel}} \approx \frac{|-0.01611 + 0.02000|}{|-0.01611|} \approx 2.4 \times 10^{-1}$$

The large root adds magnitudes and remains accurate

$$\mathrm{fl}(x_{-}) = \frac{-62.10 - 62.06}{2.000} = \frac{-124.2}{2.000} = -62.10$$

$$\varepsilon_{\mathrm{rel}} \approx \frac{|-62.08 + 62.10|}{|-62.08|} \approx 3.2 \times 10^{-4}$$

Rationalize before rounding

$$x_+ = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \cdot \frac{-b - \sqrt{b^2 - 4ac}}{-b - \sqrt{b^2 - 4ac}}$$

$$x_+ = \frac{b^2 - (b^2 - 4ac)}{2a\left(-b - \sqrt{b^2 - 4ac}\right)}$$

$$\boxed{x_+ = \frac{-2c}{b + \sqrt{b^2 - 4ac}}}$$

The reformulated small root is accurate

$$\mathrm{fl}(x_+) = \frac{-2.000}{62.10 + 62.06} = \frac{-2.000}{124.2} = -0.01610$$

$$\varepsilon_{\mathrm{rel}} \approx 6.2 \times 10^{-4}$$

Algebraically equal formulas can have radically different numerical behavior.

Choose the stable branch from the sign of the linear coefficient

When the coefficient is nonnegative

$$x_+ = \frac{-2c}{b + \sqrt{b^2 - 4ac}}$$

Compute the other root with the ordinary branch that adds magnitudes.

When the coefficient is negative

$$x_- = \frac{-2c}{b - \sqrt{b^2 - 4ac}}$$

Reverse the choice so the denominator again adds magnitudes.

Never place a near-zero difference in the denominator.

Trigonometric identities can remove hidden subtraction

$$1 - \cos x = 2\sin^2\!\left(\frac{x}{2}\right)$$

$$\cos x - \cos y = -2\sin\!\left(\frac{x+y}{2}\right)\sin\!\left(\frac{x-y}{2}\right)$$

Symbolic equivalence is only the start—check the reformulation for new hazards.

Error propagation in multiplication

$$\widehat{q} = q \prod_{i=1}^{n} (1 + \delta_i), \qquad |\delta_i| \le u$$

Carrying every local symbol through a long derivation is exact—but cumbersome.

Collect the entire product into one controlled perturbation.

Collect the product into one factor

$$\prod_{i=1}^{n}(1+\delta_i) = 1+\theta_n$$

$$|\theta_n| \le \gamma_n := \frac{nu}{1-nu}, \qquad nu < 1$$

$$|\theta_n| \le (1+u)^n - 1 \le \frac{nu}{1-nu}$$

Summary

01	**Model each rounded operation**	Attach a local relative error to each quantity and follow the errors through the actual evaluation order.
02	**Watch for cancellation**	Near-equal same-sign quantities can amplify small input errors.
03	**Reformulate before computing**	Use identities, stable branches, and nested arithmetic to expose small quantities directly.