MAT 103: Numerical Analysis I

Topic 2: Errors

Dr. Anna Fome

2026-04-30

Table of contents

1 Introduction **2**

2 Absolute and Relative Errors **3**
2.1 Notation . . . 3
2.2 Absolute Error . . . 3
2.2.1 Solved Examples . . . 3
2.3 Relative Error . . . 4
2.3.1 Solved Examples . . . 4

3 Precision and Accuracy **5**
3.1 The Simple Idea . . . 6

4 Significant Figures and Rounding **8**
4.1 Significant Figures . . . 8
4.1.1 Rules for Identifying Significant Figures . . . 8
4.1.2 Solved Examples . . . 9
4.2 Rounding Procedure . . . 9
4.2.1 Standard Rounding Rule . . . 9
4.2.2 Solved Examples . . . 10

5 Fixed-Point and Floating-Point Arithmetic **11**
5.1 Fixed-Point Arithmetic . . . 11
5.2 Floating-Point Arithmetic . . . 12
5.2.1 Solved Examples . . . 12

6 Try the following **13**

7 Sources and Types of Errors **14**
7.1 Type 1: Inherent Errors (Input Data Errors) . . . 14
7.2 Type 2: Rounding Errors . . . 15
7.3 Type 3: Truncation Errors . . . 16

7.4 Summary: Types of Errors 17

8 Amplification of Errors (Stability) 18
8.1 Error Propagation Through a Function 18
8.1.1 Solved Examples 19
8.2 Numerical Stability 19
8.2.1 Solved Example: 20

9 Inherent and Induced Instability 20
9.1 Inherent Instability (Ill-Conditioning) 20
9.1.1 Solved Example 20
9.2 Induced Instability (Numerical Instability) 21
9.2.1 Solved Example 22

10 Tutorial Questions 23
10.1 Section A: Short Answer Questions 23
10.2 Section B: Error Calculations 24
10.3 Section C: Floating-Point and Rounding Errors 25
10.4 Section D: Stability and Ill-Conditioning 25

Lord Kelvin, Physicist

"To measure is to know. If you cannot measure it, you cannot improve it."

---

1 Introduction

In Topic 1, we established that numerical methods produce **approximate** solutions. This immediately raises a very important question:

> *How far is the approximate answer from the true answer — and is it close enough to be useful?*

This question is at the heart of **error analysis** — the systematic study of how errors arise, how they are measured, how they propagate through calculations, and how they can be controlled.

Why Study Errors?

- Every numerical result carries some error. A good numerical analyst does not just compute an answer — they also compute (or estimate) how *trustworthy* that answer is.

- Error estimates give your results meaning. They tell you how reliable your approximation is, how much trust to place in it, and whether further refinement is necessary.
- In practice, they are what separate a guess from a usable solution.

By the end of Topic 2, you should be able to:

- Define and compute absolute and relative errors.
- Distinguish between precision and accuracy.
- Identify significant figures and apply correct rounding procedures.
- Explain fixed-point and floating-point arithmetic.
- Classify sources and types of errors.
- Explain what it means for errors to be amplified (stability).
- Distinguish between inherent and induced instability with examples.

---

2 Absolute and Relative Errors

2.1 Notation

Throughout this topic, we use the following notation:

- x = the **true (exact) value**
- $\tilde{x}$ = the **approximate (computed) value**
- Error = the difference between the true and approximate values

2.2 Absolute Error

The **absolute error** is the magnitude of the difference between the true value and the approximate value:

$$E_a = |x - \tilde{x}|$$

It tells us *by how much* the approximation differs from the truth, in the same units as the quantity being measured.

2.2.1 Solved Examples

Example 2.1

The true value of $\sqrt{5} = 2.2360679\ldots$ A student approximates it as $\tilde{x} = 2.236$.

$$E_a = |2.2360679 - 2.2360000| = 0.0000679$$

The absolute error is 6.79×10^{-5}.

Example 2.2

In an experiment, the true temperature is $T = 98.6°C$ and the thermometer reads $\tilde{T} = 97.9°C$.

$$E_a = |98.6 - 97.9| = 0.7°C$$

The thermometer has an absolute error of $0.7°C$.

Example 2.3

A bridge has a true length of $L = 500.000$ m. A surveyor measures it as $\tilde{L} = 499.997$ m.

$$E_a = |500.000 - 499.997| = 0.003 \text{ m}$$

2.3 Relative Error

The **absolute error** alone can be misleading. An error of 0.003 m in measuring a bridge of 500 m is negligible. But the same error of 0.003 m in measuring a laboratory specimen of 0.010 m is enormous.

The **relative error** puts the absolute error in context by expressing it as a fraction of the true value:

$$E_r = \frac{|x - \tilde{x}|}{|x|}, \quad x \neq 0$$

It is dimensionless and is often expressed as a **percentage error**:

$$E_r\% = \frac{|x - \tilde{x}|}{|x|} \times 100\%$$

2.3.1 Solved Examples

Example 2.4

Using the bridge example above ($L = 500$ m, $\tilde{L} = 499.997$ m):

$$E_r = \frac{0.003}{500.000} = 0.000006 = 0.0006\%$$

This is an extremely small relative error — the measurement is very good.

Example 2.5

True value: $x = 0.010$ m. Approximation: $\tilde{x} = 0.007$ m.

$$E_a = |0.010 - 0.007| = 0.003 \text{ m}$$

$$E_r = \frac{0.003}{0.010} = 0.3 = 30\%$$

The same absolute error of 0.003 m now corresponds to a **30% relative error** — very poor!

Example 2.6

True value of $\pi = 3.14159265\ldots$ Two approximations are offered:

- Approximation A: $\tilde{\pi}_A = 3.14$
- Approximation B: $\tilde{\pi}_B = 3.1416$

For Approximation A:

$$E_a = |3.14159265 - 3.14000000| = 0.00159265$$

$$E_r\% = \frac{0.00159265}{3.14159265} \times 100\% \approx 0.051\%$$

For Approximation B:

$$E_a = |3.14159265 - 3.14160000| = 0.00000735$$

$$E_r\% = \frac{0.00000735}{3.14159265} \times 100\% \approx 0.00023\%$$

Approximation B is much more accurate, as shown by its smaller relative error.

---

3 Precision and Accuracy

Students often mix up these two words because in everyday speech we use them interchangeably. In mathematics and science, however, they mean **completely different things**. Let us build up the distinction slowly, using simple everyday situations.

3.1 The Simple Idea

Think of it this way:

- **Accuracy** answers the question: *"Is the answer close to the truth?"*
- **Precision** answers the question: *"How many digits did we use to express the answer?"*

That is it. Accuracy is about **closeness to the truth**. Precision is about **the number of digits** used.

> **i** Definitions in Plain Language
>
> **Accurate** — the answer is very close to the true (correct) value. The error $|x - \tilde{x}|$ is small.
> **Precise** — the answer is written with many digits (many significant figures). It looks detailed, regardless of whether those digits are correct or not.

The most important thing to understand is this:

> **A number can be written with many digits (precise) and still be WRONG (inaccurate).** And a number can be written with few digits (not very precise) and still be CORRECT (accurate).

This surprises. The examples below will make it completely clear.

Example 2.7: Measuring a Desk

Suppose the **true length** of a desk is exactly **1.500 m**.

Four students measure it and report:

Student	Reported Value	Number of Digits	Is it close to 1.500?	Verdict
Amara	1.5 m	2	Yes	Low precision, **accurate**
Boni	1.500 m	4	Yes	High precision, **accurate**
Chalo	1.756 m	4	No	High precision, **inaccurate**
Doto	1.8 m	2	No	Low precision, **inaccurate**

Let us compute the errors:

For **Amara** ($\tilde{x} = 1.5$):

$$E_a = |1.500 - 1.5| = 0.000 \text{ m}, \quad E_r\% = 0\%$$

Amara's answer is short (only 2 digits) but perfectly accurate.

For **Boni** ($\tilde{x} = 1.500$):

$$E_a = |1.500 - 1.500| = 0.000 \text{ m}, \quad E_r\% = 0\%$$

Boni's answer is both precise (4 digits) and accurate. This is the ideal.

For **Chalo** ($\tilde{x} = 1.756$):

$$E_a = |1.500 - 1.756| = 0.256 \text{ m}, \quad E_r\% = \frac{0.256}{1.500} \times 100\% \approx 17\%$$

Chalo used 4 digits — that looks detailed and careful — but the answer is very wrong. **Precise but inaccurate.**

For **Doto** ($\tilde{x} = 1.8$):

$$E_a = |1.500 - 1.800| = 0.300 \text{ m}, \quad E_r\% = 20\%$$

Doto used only 2 digits and the answer is also wrong. **Imprecise and inaccurate.**

> ⚠ Warning
>
> **Key Lesson from Example 2.7**
> Do not be fooled by a long string of digits. **More digits does NOT automatically mean a more correct answer.** Chalo wrote 1.756 — four digits, looks careful — but was 17% wrong. Amara wrote 1.5 — only two digits — and was perfectly correct.

Example 2.8: Estimating the Value of π

The true value of $\pi = 3.14159265358979\ldots$

Five approximations are given below. For each one, we ask: how many digits does it use (precision), and how close is it to the truth (accuracy)?

Approximation	Digits Used	Absolute Error	% Error	Precision	Accuracy
3	1	0.14159	4.51%	Very low	Poor
3.1	2	0.04159	1.32%	Low	Moderate

Approximation	Digits Used	Absolute Error	% Error	Precision	Accuracy
3.14	3	0.00159	0.051%	Moderate	Good
3.14159	6	0.0000027	0.00008%	High	Excellent
3.14200	6	0.00041	0.013%	High	Moderate

Study the last two rows carefully.

Both 3.14159 and 3.14200 use **6 digits** — they have the same precision.

But 3.14159 has an error of only 0.0000027, while 3.14200 has an error of 0.00041 — about **150 times larger**.

So **same precision, very different accuracy**.

This is exactly like two students both writing a 6-digit answer in an exam — one copied the digits correctly, the other made a mistake. The number of digits they wrote tells you nothing about whether they got it right.

Think About It

Your phone calculator displays $\pi = 3.141592653589793$ — that is 16 digits. Is this precise? Yes, very. Is it accurate? Yes, very — it matches the true value to 16 significant figures.
Now imagine a **broken** calculator that displays $\pi = 3.141592600000000$. It still shows 16 digits — equally precise. But it is slightly less accurate (error in the 9th digit onwards).
Same precision. Different accuracy.

4 Significant Figures and Rounding

4.1 Significant Figures

The **significant figures** (or **significant digits**) of a number are the digits that carry meaningful information about its precision. They tell us *how carefully* a number has been measured or computed.

4.1.1 Rules for Identifying Significant Figures

Rule 1: All **non-zero digits** are significant.

4721 has **4** significant figures. 3.85 has **3** significant figures.

Rule 2: Zeros between non-zero digits are significant.

4021 has **4** significant figures. 3.005 has **4** significant figures.

Rule 3: Leading zeros (zeros before the first non-zero digit) are **NOT** significant. They merely indicate the position of the decimal point.

> 0.0025 has **2** significant figures (2 and 5). 0.00307 has **3** significant figures (3, 0, 7).

Rule 4: Trailing zeros after the decimal point are significant — they indicate precision.

> 3.500 has **4** significant figures. 12.00 has **4** significant figures.

Rule 5: Trailing zeros in a whole number without a decimal point are ambiguous.

> 1400 — could be 2, 3, or 4 significant figures. Write 1.400×10^3 for 4 s.f.

4.1.2 Solved Examples

Example 2.8

Identify the number of significant figures in each of the following:

Number	Significant Figures	Reason
7843	4	All non-zero digits
70043	5	Zeros between non-zeros are significant
0.0043	2	Leading zeros are not significant
0.04300	4	Trailing zeros after decimal are significant
1.0080	5	Interior and trailing zeros both significant
50000	Ambiguous	Use standardized exponential scientific notation to clarify
5.000×10^4	4	Trailing zeros shown explicitly

4.2 Rounding Procedure

Rounding is the process of reducing the number of digits in a number while keeping its value as close as possible to the original.

4.2.1 Standard Rounding Rule

To round a number to n significant figures:

1. Identify the n-th significant digit.

2. Look at the **digit immediately after** (the $(n+1)$-th digit):
   - If it is **less than 5**: simply drop it (and all digits after it). The n-th digit stays the same.
   - If it is **5 or greater**: increase the n-th digit by 1, then drop the rest.

4.2.2 Solved Examples

Example 2.9

Round 3.14159265 to the indicated number of significant figures:

To n s.f.	Rounded Value	Working
1	3	4th digit is $1 < 5$: round down
2	3.1	3rd digit is $4 < 5$: round down
3	3.14	4th digit is $1 < 5$: round down
4	3.142	5th digit is 5: round up
5	3.1416	6th digit is 9 $\quad$ 5: round up
6	3.14159	7th digit is $2 < 5$: round down

Example 2.10

Round 0.0074651 to 3 significant figures.

Step 1: Identify the significant figures. The leading zeros are not significant. The significant digits are 7, 4, 6, 5, 1.

Step 2: We want 3 significant figures: 7, 4, 6.

Step 3: The next digit is $5 \geq 5$, so we round up: $6 \to 7$.

$$0.0074651 \approx 0.00747 \quad \text{(3 significant figures)}$$

Example 2.11

Round 24950 to 3 significant figures.

Step 1: The significant digits are 2, 4, 9, 5, 0.

Step 2: We want 3 s.f.: 2, 4, 9.

Step 3: The next digit is $5 \geq 5$, so round up: $9 \to 10$. This causes a carry.

$$24950 \approx 25000 \quad \text{(3 significant figures)} = 2.50 \times 10^4$$

Common Mistake

Students sometimes confuse *decimal places* with *significant figures*.

- 0.00314 rounded to **2 decimal places** would give 0.00 (which loses all information!).
- 0.00314 rounded to **2 significant figures** gives 0.0031 — which is meaningful.

Always use **significant figures** in scientific and numerical work.

5 Fixed-Point and Floating-Point Arithmetic

Modern computers cannot store numbers with infinite precision. They use a finite number of digits. Understanding how numbers are stored is crucial for understanding where rounding errors come from.

5.1 Fixed-Point Arithmetic

In **fixed-point** representation, the decimal point is fixed at a predetermined position. A fixed-point number with m digits before and n digits after the decimal point can represent numbers of the form:

$$\underbrace{d_m d_{m-1} \cdots d_1}_{m \text{ digits}} . \underbrace{d_{-1} d_{-2} \cdots d_{-n}}_{n \text{ digits}}$$

Example 2.12

Using 3 digits before and 4 digits after the decimal point (format: *ddd.dddd*): The largest number possible: 999.9999 and the smallest positive nonzero number is: 0.0001

True Value	Fixed-Point Representation	Rounding Error
3.14159	3.1416	0.00001
27.31826	27.3183	0.00006
1000.5	**OVERFLOW**	—
0.000031	0.0000	0.000031 (complete loss!)

Limitations of Fixed-Point:

- Very **limited range**: large numbers overflow, very small numbers underflow to zero.
- Not suitable for scientific computation where values span many orders of magnitude.
- Simple and fast — used in embedded systems and financial applications where the range of values is known in advance.

5.2 Floating-Point Arithmetic

Floating-point representation is the standard in scientific computing. A number is stored in the form:

$$x = \pm m \times \beta^e$$

where:

- $\pm$ is the **sign** (positive or negative)
- m is the **mantissa** (also called the *significand*): $1 \leq m < \beta$
- β is the **base** (usually $\beta = 2$ in computers, $\beta = 10$ for humans)
- e is the **exponent** (an integer, positive, negative, or zero)

Floating-point arithmetic solves the problem of representing extremely large or infinitesimally small numbers using a limited amount of computer memory.

Floating-point works like scientific notation (1.23×10^{15}); it "floats" the decimal point by storing a significand (the digits) and an exponent (the scale). This provides a massive range and consistent precision across different scales, which is essential for scientific engineering.

5.2.1 Solved Examples

Example 2.13

Write the following in normalised floating-point form (base 10):

Number	Floating-Point Form	m	e
3141.59	3.14159×10^3	3.14159	3
0.000271	2.71×10^{-4}	2.71	-4
-56800	-5.68×10^4	5.68	4
1	1.0×10^0	1.0	0

Example 2.14

A computer uses a **4-digit mantissa** (base 10) and stores numbers as $d_1.d_2d_3d_4 \times 10^e$.

Represent $\frac{2}{3} = 0.6666666\ldots$ in this system.

Step 1: Write in normalised form: $6.666\overline{6} \times 10^{-1}$

Step 2: Round mantissa to 4 digits: 6.667×10^{-1}

Stored value: 0.6667

Rounding error: $|0.66\overline{6} - 0.6667| = 0.0000\overline{3} \approx 3.3 \times 10^{-5}$

Feature	Fixed-Point	Floating-Point
Decimal Point	Fixed	Moves
Range	Small	Very large
Speed	Faster	Slower
Accuracy	Limited	Better
Best Use	Money	Science

6 Try the following

Q1: Convert to fixed-point (ddd.dddd). State if overflow or underflow. Find the rounding error

a) 5.23891

b) 72.6

c) 0.00348

d) 18.99996

e) 1000.25

f) 999.9999

g) 0.00002

h) 875.1254

Q2: Convert to floating-point presentation. State the Mantissa & exponent

a) 3141.59

b) 0.00045

c) 720000

d) -56.8

Q3: Convert to normal number

a) $3.2 \times 10\hat{}4$

b) $6.5 \times 10\hat{}-3$

c) -8.1 × 10^2

---

7 Sources and Types of Errors

Errors in numerical computation arise from several distinct sources. A careful numerical analyst identifies *where* errors come from in order to control them.

In numerical computation, answers are often approximate, not exact. This happens because of limitations in: the input data, the computer, the numerical method, or repeated calculations. A good numerical analyst always asks: Where did the error come from?

7.1 Type 1: Inherent Errors (Input Data Errors)

Inherent errors are errors that are present in the **data before any computation begins**. They arise from:

- **Measurement uncertainty:** Physical instruments have limited precision.
- **Model errors/Wrong assumptions in the model:** The mathematical model may not perfectly represent reality.
- **Conversion errors:** Continuous quantities converted to digital values introduce error.

These errors are *inherited* by the computation — the algorithm cannot remove them.

Example 2.15: Measuring length

A ruler shows:

$$L = 10.2 \pm 0.1 \text{ cm}$$

Actual value may be:

- 10.1 cm
- 10.2 cm
- 10.3 cm

The uncertainty already exists before calculation.

Example 2.16: Temperature sensor

Sensor says:

$$25.0^{\circ}C$$

But real temperature may be:

$$24.8^{\circ}C$$

This input error affects every later calculation.

Example 2.14

A physicist measures the gravitational acceleration using a pendulum:

$$g = \frac{4\pi^2 L}{T^2}$$

The measured length $L = 1.002 \pm 0.001$ m and period $T = 2.010 \pm 0.005$ s.

Even before computing g, the inputs already carry uncertainty (± 0.001 m and ± 0.005 s). This is an inherent error — it comes from the measurement instrument, not the formula.

Easy Meaning > Bad input gives bad output.

7.2 Type 2: Rounding Errors

Computers store only a limited number of digits. **Rounding errors** arise because computers store numbers with **finite precision**. Any number that cannot be represented exactly is rounded to the nearest representable value.

Rounding errors occur:

- When reading a number into memory (e.g., 1/3 cannot be stored exactly in decimal).
- After each arithmetic operation (the result may need more digits than the system provides).

Example 2.17: Decimal rounding

Exact:

$$2.7182818$$

Stored with 4 digits:

$$2.718$$

Error:

$$0.0002818$$

Example 2.18: Money rounding

Price:

$$19.999$$

Rounded to cents:

$$20.00$$

Example 2.19

A computer with a 5-digit decimal mantissa performs the addition:

$$0.36150 + 0.00025 = 0.36175 \quad \text{(exact)}$$

But if the result is stored in 4 digits:

$$\tilde{x} = 0.3618 \quad \text{(rounded)}$$

Rounding error $= |0.36175 - 0.36180| = 0.00005$.

7.3 Type 3: Truncation Errors

Truncation errors arise when an **infinite mathematical process** (such as an infinite series, an exact derivative, or an exact integral) is replaced by a **finite approximation**.

The part that is *cut off* (truncated) is the truncation error.

Example 2.20: Using π

Use:

$$\pi \approx 3.14$$

True:

$$3.14159265...$$

Difference is truncation/approximation error.

Example 2.21: Taylor Series Truncation

The exponential function has the infinite Taylor series expansion:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots$$

If we approximate using only the first **4 terms**:

$$e^x \approx 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!}$$

The **truncation error** is:

$$E_T = \frac{x^4}{4!} + \frac{x^5}{5!} + \cdots$$

For $x = 1$:

$$e^1 = e = 2.71828182\ldots$$
$$\tilde{e} \approx 1 + 1 + 0.5 + 0.16667 = 2.66667$$

Truncation error $= |2.71828 - 2.66667| = 0.05161$

Using more terms reduces the truncation error:

Terms Used	Approximation	Truncation Error
1	1.00000	1.71828
2	2.00000	0.71828
3	2.50000	0.21828
4	2.66667	0.05161
5	2.70833	0.00995
6	2.71667	0.00162
7	2.71806	0.00022

Each additional term reduces the truncation error substantially.

7.4 Summary: Types of Errors

Error Type	Source	When It Arises	Controllable?
Inherent	Measurement, model limitations	Before computation	Partially
Rounding	Finite machine precision	During storage and arithmetic	Yes (more digits)
Truncation	Approximating infinite processes	During algorithm design	Yes (more terms)

8 Amplification of Errors (Stability)

8.1 Error Propagation Through a Function

Suppose we want to compute $f(x)$, but we do not know the true(exact) value of x. Instead, we use an approximation value:

$$\tilde{x} = x + \Delta x$$

where:

- x = true value
- $\tilde{x}$ = approxximate value
- Δx = small error in x.

The question is:

If there is a small erro in x, what error appears in $f(x)$?

Approximate Change in Function Value Using the first-order Taylor expansion:

$$f(\tilde{x}) = f(x + \Delta x) \approx f(x) + f'(x) \cdot \Delta x$$

So the error in the function is approximately:

$$\Delta f \approx f'(x)\Delta x$$

This means:

The derivative $f'(x)$ tells us how sensitive the function is to errors

The **absolute error in f**:

$$\Delta f \approx |f'(x)| \cdot |\Delta x|$$

Meaning

- If $|f'(x)|$ is large $\rightarrow$ error become larger

-If $|f'(x)|$ is small $\rightarrow$ error stays small

The **relative error in f**:

$$\frac{|\Delta f|}{|f(x)|} \approx \left|\frac{x \cdot f'(x)}{f(x)}\right| \cdot \frac{|\Delta x|}{|x|}$$

The quantity:

$$\kappa = \left|\frac{x \cdot f'(x)}{f(x)}\right|$$

is called the **condition number** of the problem at x. It tells us whether the problem amplifies errors.

- If κ is large (> 1), a small relative error in x causes a large relative error in $f(x) \rightarrow$ the problem is **ill-conditioned**.
- If κ is small (<1), Small input error gives small output error $\rightarrow$ the problem is **well-conditioned**.

8.1.1 Solved Examples

Example 2.22

Compute the condition number for $f(x) = \sqrt{x}$ at $x = 4$. We know $f'(x) = \frac{1}{2\sqrt{x}}$, so at $x = 4$: $f'(4) = \frac{1}{4}$, $f(4) = 2$.

Now:

$$\kappa = \left| \frac{x \cdot f'(x)}{f(x)} \right| = \left| \frac{4 \cdot \frac{1}{4}}{2} \right| = \frac{1}{2}$$

The condition number is $0.5 < 1$. Errors in x are **damped** when computing $\sqrt{x}$ — this is a well-conditioned operation.

Example 2.23

Compute the absolute error in $f(x) = x^3$ at $x = 5$, given that x has an error of $\Delta x = 0.01$.

Derivative: $f'(x) = 3x^2$, so $f'(5) = 75$.

$$\Delta f \approx |f'(5)| \cdot |\Delta x| = 75 \times 0.01 = 0.75$$

The error in x (0.01) become 0.75 in output **Error amplified** (by a factor of 75).

A relative error of $\frac{0.01}{5} = 0.2\%$ in x leads to a relative error of $\frac{0.75}{125} = 0.6\%$ in f — still small, but amplified 3-fold.

8.2 Numerical Stability

A **numerical method** is **stable** if small perturbations in the data or small rounding errors in intermediate steps do not cause the error to grow uncontrollably as the computation proceeds.

A method is **unstable** if errors grow (often exponentially) as the computation progresses.

8.2.1 Solved Example:

Example 2.24 Stable vs Unstable Iteration

Suppose:

$$e_{n+1} = 2e_n$$

For initial error $e_n = 0.01$. The errors become:

$$0.02, 0.04, 0.08, ...$$

The error double at each step $\rightarrow$ Unstable

Now suppose:

$$e_{n+1} = 0.5e_n$$

The errors become:

$$0.005, 0.0025, 0.00125, ...$$

The error decreases $\rightarrow$ Stable

9 Inherent and Induced Instability

9.1 Inherent Instability (Ill-Conditioning)

Inherent instability (or *ill-conditioning*) is a property of the **mathematical problem itself**, not the method used to solve it.

A problem is **ill-conditioned** if small changes in the input cause disproportionately large changes in the output. This is measured by the **condition number**: a large condition number means the problem is ill-conditioned.

> **Key Idea**
>
> If the problem is inherently ill-conditioned, **no numerical method** — however carefully designed — can give an accurate answer if the input data is even slightly uncertain. The inaccuracy is a property of the problem, not the algorithm.

9.1.1 Solved Example

Example 2.25: An Ill-Conditioned Linear System

Consider the system of equations:

$$x + y = 2 \quad \cdots (1)$$
$$1.0000x + 1.0001y = 2.0001 \quad \cdots (2)$$

True solution: $x = 1$, $y = 1$.

Now perturb the right-hand side of equation (2) by a tiny amount — change 2.0001 to 2.0002:

$$x + y = 2 \quad \cdots (1)$$
$$1.0000x + 1.0001y = 2.0002 \quad \cdots (2')$$

New solution:

From (1): $x = 2 - y$. Substitute into Equation (2):

$$(2 - y) + 1.0001y = 2.0002$$
$$2 + 0.0001y = 2.0002$$
$$y = \frac{0.0002}{0.0001} = 2$$

Then $x = 2 - 2 = 0$.

New solution: $x = 0$, $y = 2$.

A change of 0.0001 in the right-hand side (a perturbation of 0.005%) caused a change of 1 in x and y — an enormous amplification!

This system is **inherently ill-conditioned**. Any numerical method will struggle with it.

Example 2.26:

$$f(x) = \frac{1}{x}$$

Near $x = 0$, tiny changes in x create huge changes in output. $\rightarrow$ Inherent instability

9.2 Induced Instability (Numerical Instability)

Induced instability is introduced by the **algorithm** itself, even if the underlying mathematical problem is well-conditioned.

It is caused by:

- Poor choice of algorithm
- Accumulated rounding errors in the method
- **Subtractive cancellation** (the most common cause)

9.2.1 Solved Example

Example 2.27: Subtractive Cancellation

Compute $f(x) = \sqrt{x+1} - \sqrt{x}$ for $x = 10000$.

$\sqrt{10001} = 100.004999875\ldots$ and $\sqrt{10000} = 100.000000000$.

On a **6-digit decimal** computer:

$\sqrt{10001} \approx 100.005$ and $\sqrt{10000} = 100.000$.

$$f(10000) \approx 100.005 - 100.000 = 0.005$$

But the true answer is $f(10000) = 100.004999875\ldots - 100.000 = 0.004999875\ldots$

The computed answer 0.005 has only **1 significant figure**, even though we started with 6. We have lost 5 significant figures due to cancellation.

The fix: Rationalise the expression:

$$\sqrt{x+1} - \sqrt{x} = \frac{(\sqrt{x+1} - \sqrt{x})(\sqrt{x+1} + \sqrt{x})}{\sqrt{x+1} + \sqrt{x}} = \frac{1}{\sqrt{x+1} + \sqrt{x}}$$

Using the alternative formula:

$$f(10000) = \frac{1}{\sqrt{10001} + \sqrt{10000}} = \frac{1}{100.005 + 100.000} = \frac{1}{200.005} = 0.0049999\ldots$$

This gives **5 significant figures** — much better! The induced instability is cured by choosing a better computational formula.

Example 2.28: Subtractive Cancellation in Quadratic Roots

Find the roots of $x^2 - 200x + 1 = 0$ using the standard formula.

$$x = \frac{200 \pm \sqrt{40000 - 4}}{2} = \frac{200 \pm \sqrt{39996}}{2}$$

$$\sqrt{39996} \approx 199.990000250$$

$$x_1 = \frac{200 + 199.990000250}{2} = \frac{399.990000250}{2} = 199.995000125$$

$$x_2 = \frac{200 - 199.990000250}{2} = \frac{0.009999750}{2} = 0.004999875$$

Now, on a **6-digit** machine, $\sqrt{39996} \approx 199.990$:

$$x_1 = \frac{200 + 199.990}{2} = 199.995 \quad \text{(5 s.f. — acceptable)}$$
$$x_2 = \frac{200 - 199.990}{2} = \frac{0.010}{2} = 0.005 \quad \text{(1 s.f. — poor!)}$$

The small root x_2 suffers badly from cancellation.

The fix: Compute x_1 using the standard formula, then use:

$$x_2 = \frac{1}{x_1} = \frac{1}{199.995} = 0.0050001\ldots \quad \text{(5 s.f. — much better!)}$$

(This uses the fact that for $x^2 + bx + c = 0$, the product of the roots equals c, so $x_1 x_2 = 1$.)

---

10 Tutorial Questions

Work through these questions carefully.

10.1 Section A: Short Answer Questions

Question 1

Define the following terms clearly and give a formula for each:

(a) Absolute error
(b) Relative error
(c) Percentage error

Question 2

Explain the difference between **accuracy** and **precision**. Give an example of a number that is precise but inaccurate.

Question 3

State the **three types of errors** that arise in numerical computation. For each type, state its source and give a one-sentence example.

Question 4

Explain what it means for a numerical method to be **stable**. Use an analogy to illustrate your explanation. What is the difference between **inherent** and **induced** instability?

Question 5

Identify the number of significant figures in each of the following numbers:

(a) 72048
(b) 0.00503
(c) 3.0600
(d) 1.050×10^{-3}
(e) 500 (discuss the ambiguity)

10.2 Section B: Error Calculations

Question 6

For each of the following, calculate the **absolute error**, **relative error**, and **percentage error**.

(a) True value: $x = 7.389056$; Approximation: $\tilde{x} = 7.389$
(b) True value: $x = 0.001234$; Approximation: $\tilde{x} = 0.00123$
(c) True value: $x = 1000.000$; Approximation: $\tilde{x} = 999.5$
(d) True value: $x = \pi$; Approximation: $\tilde{x} = 22/7$

Question 7

Two students measure the length of a laboratory bench. Student A reports 2.345 m and Student B reports 2.3 m. The true length is 2.350 m.

(a) Compute the absolute and relative error for each student.
(b) Which student is more accurate? Which is more precise?
(c) Is it possible for a less precise measurement to be more accurate? Explain.

Question 8

Round each of the following to **3 significant figures**:

(a) 0.0046271
(b) 135.649
(c) 9.99501
(d) 10049
(e) 0.0009985

10.3 Section C: Floating-Point and Rounding Errors

Question 9

Write each of the following in **normalised floating-point form** (base 10):

(a) 0.00567
(b) -45600
(c) 1234.56
(d) 0.1
(e) 0.000000871

Question 10

A computer uses a **5-digit decimal mantissa**. What value does it store for each of the following, and what is the rounding error in each case?

(a) $1/7 = 0.142857142857\ldots$
(b) $\sqrt{2} = 1.41421356\ldots$
(c) $2/3 = 0.666666\ldots$
(d) $\pi = 3.14159265\ldots$

Question 11

The Taylor series for $\sin(x)$ is:

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots$$

(a) Approximate $\sin(0.5)$ using the first **3 terms** of the series.
(b) The true value is $\sin(0.5) = 0.47942553\ldots$ What is the truncation error?
(c) Add the **4th term** and recalculate the truncation error. By what factor did the error decrease?

10.4 Section D: Stability and Ill-Conditioning

Question 12

Compute $f(x) = \sqrt{x} - \sqrt{x-1}$ for $x = 5000$ using:

(a) The direct formula $\sqrt{5000} - \sqrt{4999}$.
(b) The rationalised formula $\dfrac{1}{\sqrt{x} + \sqrt{x-1}}$.

Use a calculator for both. Are the results different? Which formula is numerically more reliable, and why?

Question 13

Consider the linear system:

$$2x + 6y = 8$$

$$2x + 6.00001y = 8.00001$$

(a) Find the exact solution.
(b) Now change the right-hand side of the second equation to 8.00003. Find the new solution.
(c) How large was the change in the right-hand side? How large was the change in the solution?
(d) What does this tell you about the conditioning of the system?

Question 14

The condition number for $f(x) = x^n$ is $\kappa = n$.

(a) Compute the condition number for $f(x) = x^2$ and $f(x) = x^{10}$.
(b) If x has a relative error of 0.1%, what is the approximate relative error in x^2? In x^{10}?
(c) What does this tell you about the difficulty of computing high powers numerically?

End of Topic 2 Tutorial Questions
