Numerical Methods

Sharif University of Technology
Department of Computer Engineering
Fatemeh Baharifard
Mail: fateme.baharifard@gmail.com

Course Introduction

Webpage TAs Resources Grade

Join the Course Page on [Quera].

Everything will be posted on Quera page.

Course Introduction

Webpage | **TAs** | Resources | Grade

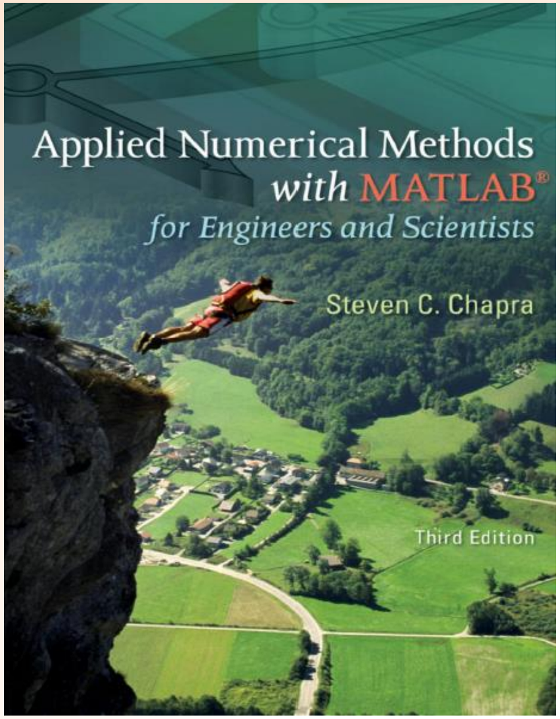
Mohammad Reza Daviran	Head
Roya Ghavami	
Kahbod Aeini	
Zohre Abbasi	
Mahdi Abootorabi	
Bahar Asadi	
Nima Jamali	
Kasra Amani	
Alireza Daghigh	
Mahdiyeh Ebrahimpoor	
Mohammad Pourtaheri	

For any enquiries, please contact Mr. Daviran via
mailto: mohammadreza.dn80@gmail.com

Course Introduction

Webpage TAs **Resources** Grade

- APPLIED NUMERICAL METHODS WITH MATLAB® for Engineers and Scientists -Steven C. Chapra -Third Edition 2012
- NUMERICAL METHODS IN ENGINEERING WITH MATLAB
  Jaan Kiusalaas, 3rd Edition
  Pennsylvania State University
  Cambridge University Press 2016
- محاسبات عددی، مسعود نیکوکار



Course Introduction

Webpage TAs Resources **Grade**

Grade		
Midterm	4	Ch(1-3) – 20 Ordibehesht
Final	7	Ch(1-6) – 4 Tir
Six HWs	6	Theoretical and Practical
Project	3	TBA
Random Quizzes	1	Extra points

Why Numerical Methods?

Some problems cannot be solved analytically or are too long and tedious to calculate.

$\int_0^1 e^{-x^2} \mathrm{dx}$

$\int_0^1 \frac{1}{1+x^3} \mathrm{dx}$

Numerical Methods

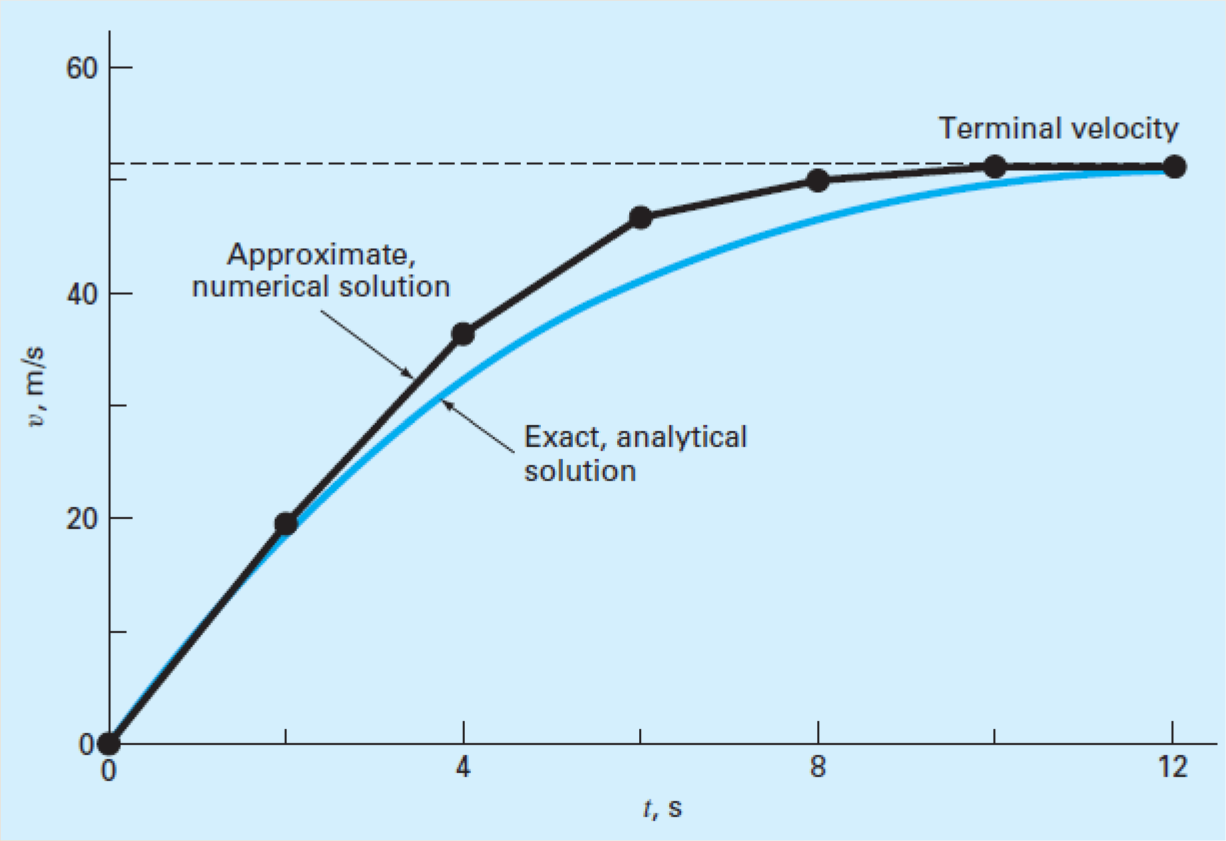
NM Methods:

- Iterative
- Direct

Contents

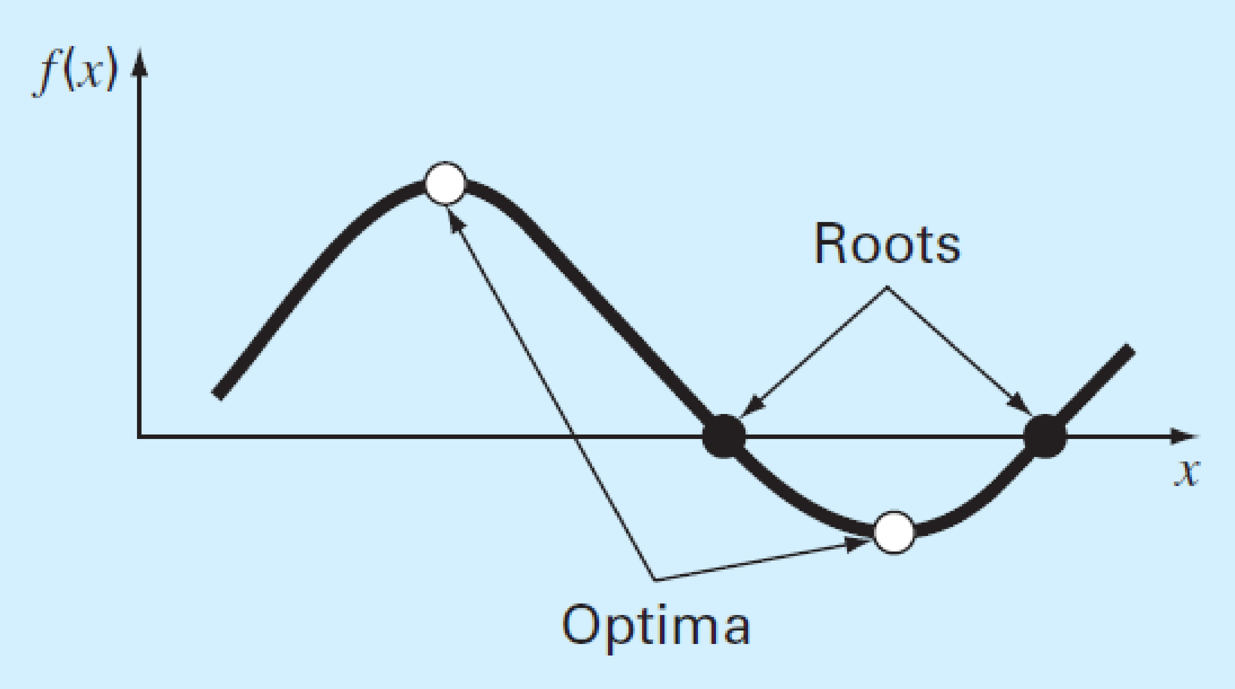
- Six chapters:
  - Errors
  - Numerical methods for solving nonlinear equations
  - Interpolation, extrapolation and curve fitting
  - Numerical Integration and differentiation
  - Ordinary differential equations
  - System of linear equations

Errors



60
40
20
0
v, m/s
Terminal velocity
Approximate,
numerical solution
Exact, analytical
solution
0
4
8
12
t, s

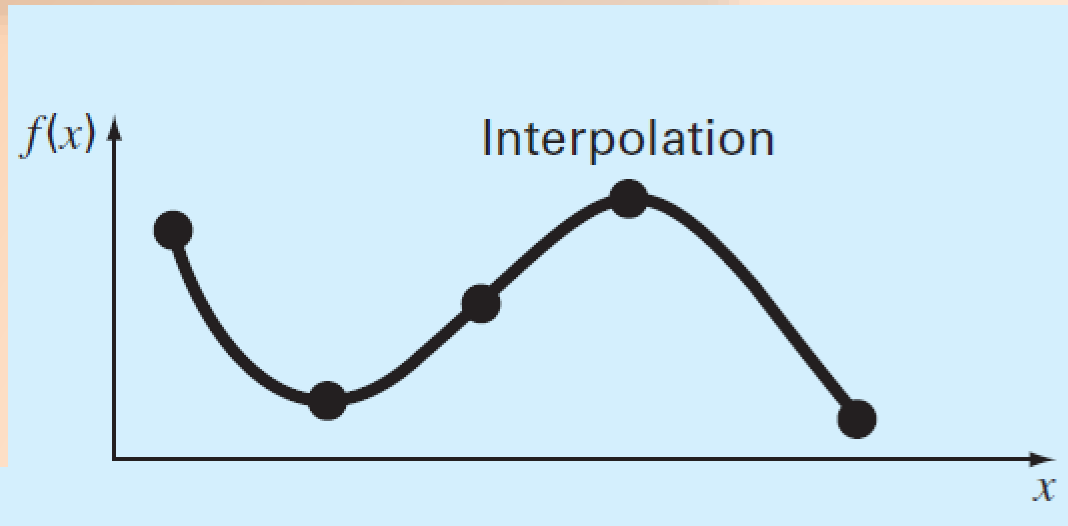
Roots of Nonlinear Equations



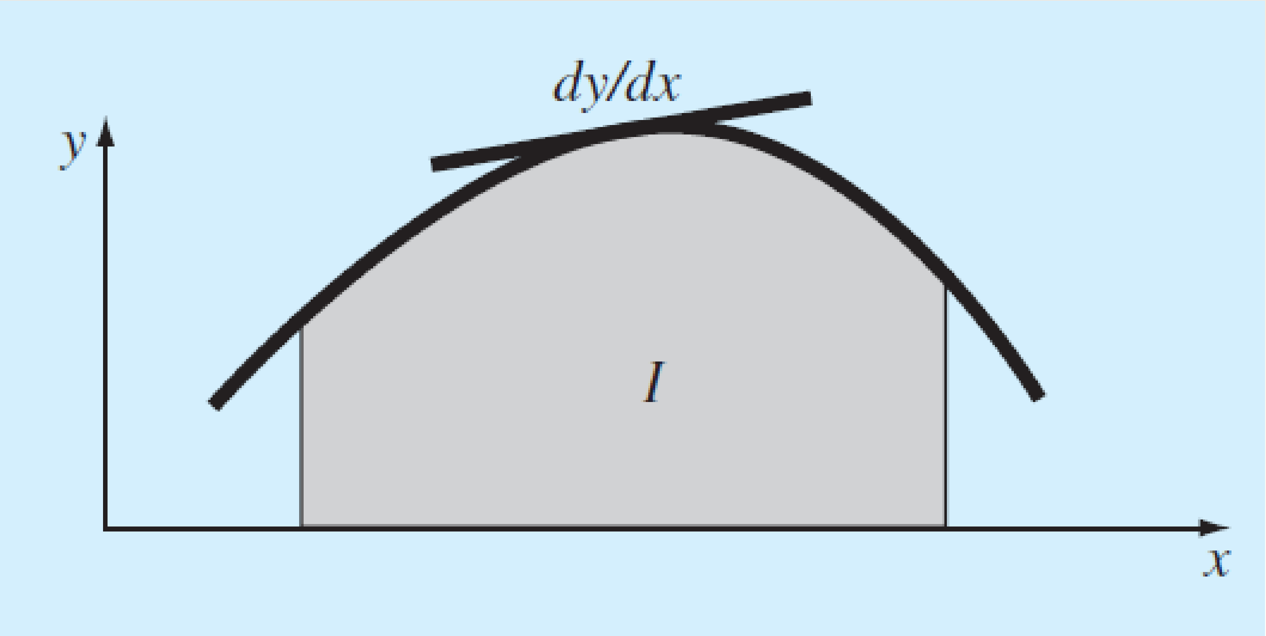
Roots: Solve for x so that $f(x) = 0$

Optimization: Solve for x so that $f'(x) = 0$

Interpolation, Extrapolation, Curve Fitting



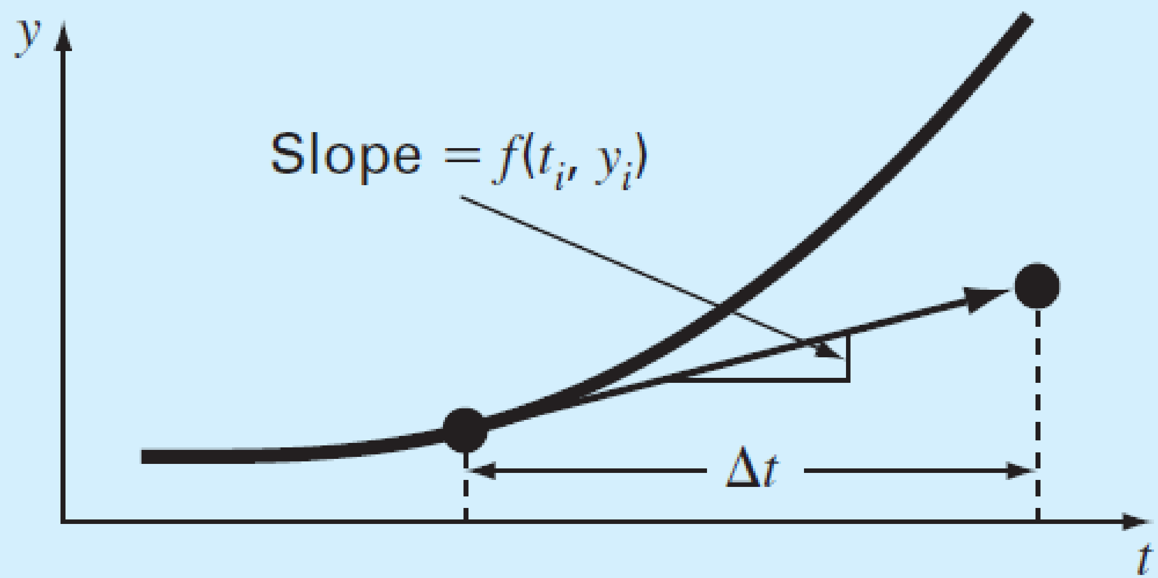
Integration and Differentiation



Integration: Find the area under the curve

Differentiation: Find the slope of the curve

Ordinary Differential Equations



Given

$$\frac{dy}{dt} \approx \frac{\Delta y}{\Delta t} = f(t, y)$$

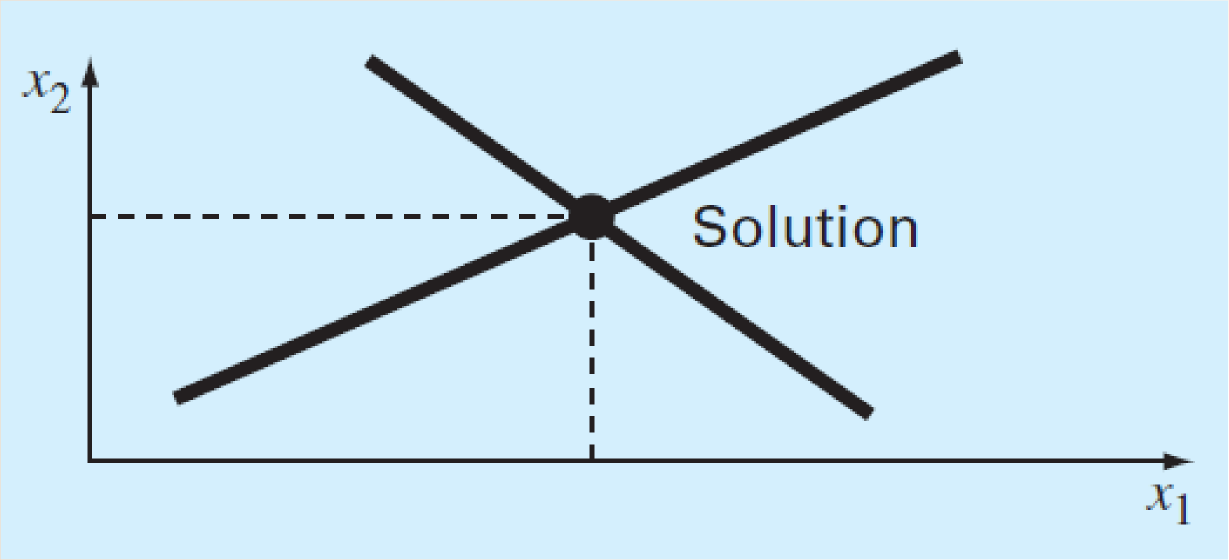
solve for y as a function of t

$$y_{i+1} = y_i + f(t_i, y_i)\Delta t$$

System of Linear equations

Given the a's and the b's, solve for the x's

$$a_{11}x_1 + a_{12}x_2 = b_1$$
$$a_{21}x_1 + a_{22}x_2 = b_2$$



Chapter 1

Errors

Source of Error	Error Representation	Floating Point Representation	Types of Error	Error Propagation and Process Graph

Introduction

- What is error?
- Where does it come from?
- What types does it have?
- How can we minimize it?

Precision & Accuracy

Accuracy and Precision:

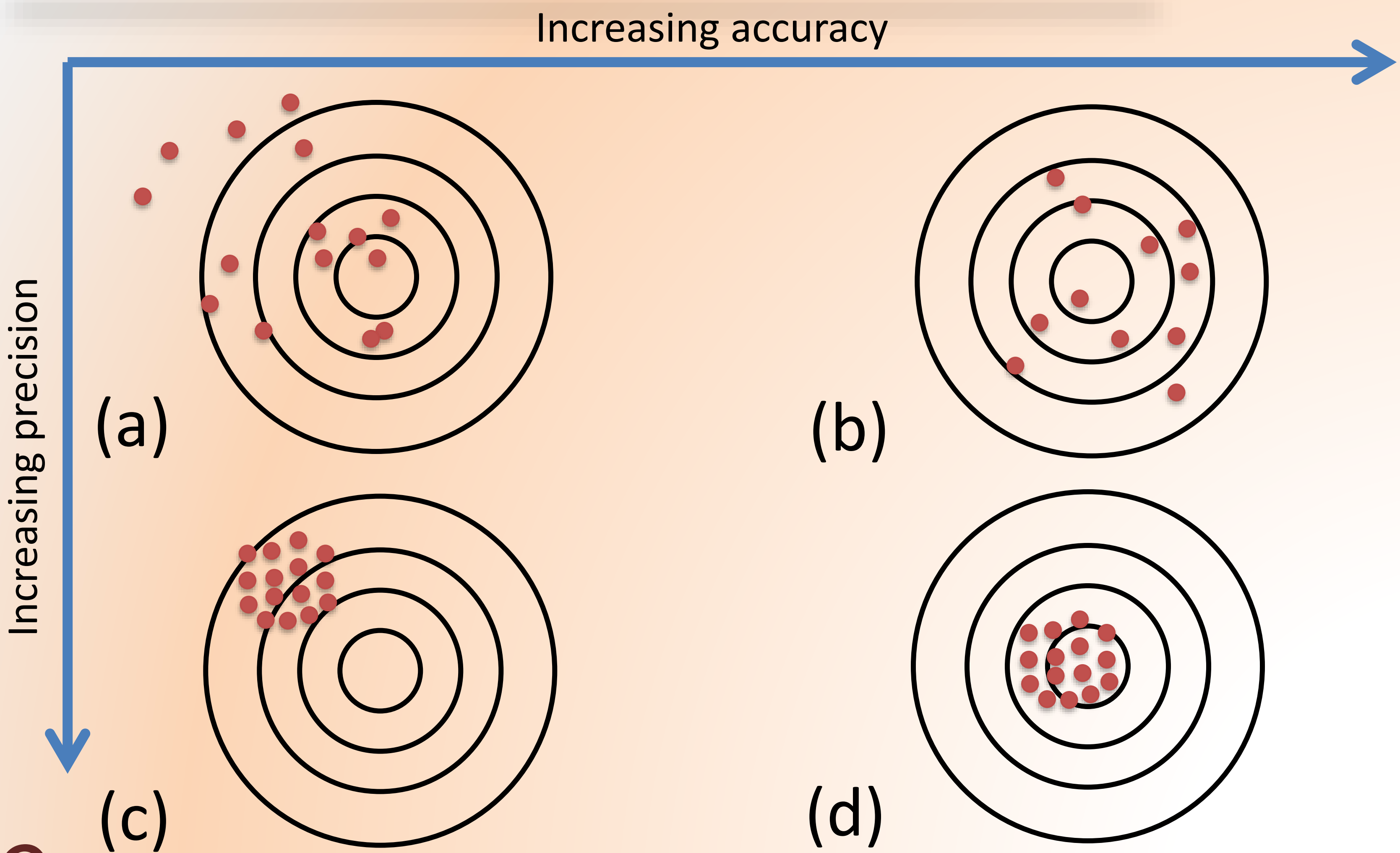
Accuracy refers to how closely a computed or measured value agrees with the true value.

Precision refers to how closely individual computed or measured values agree with each other.

Inaccuracy (also known as bias) is the systematic deviation from the truth.

Imprecision (uncertainty) refers to the magnitude of the scatter.

Precision & Accuracy



Examples of Errors

$$\sqrt{5} \overset{!}{=} 2....$$

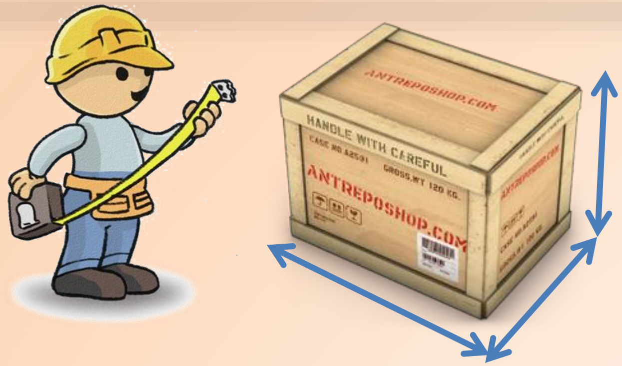
Examples of Errors

$$\sqrt{5} \overset{!}{=} 2.2\ldots3\ldots$$

Examples of Errors

$$\sqrt{5} \overset{!}{=} 2.2360679\ldots$$

Examples of Errors



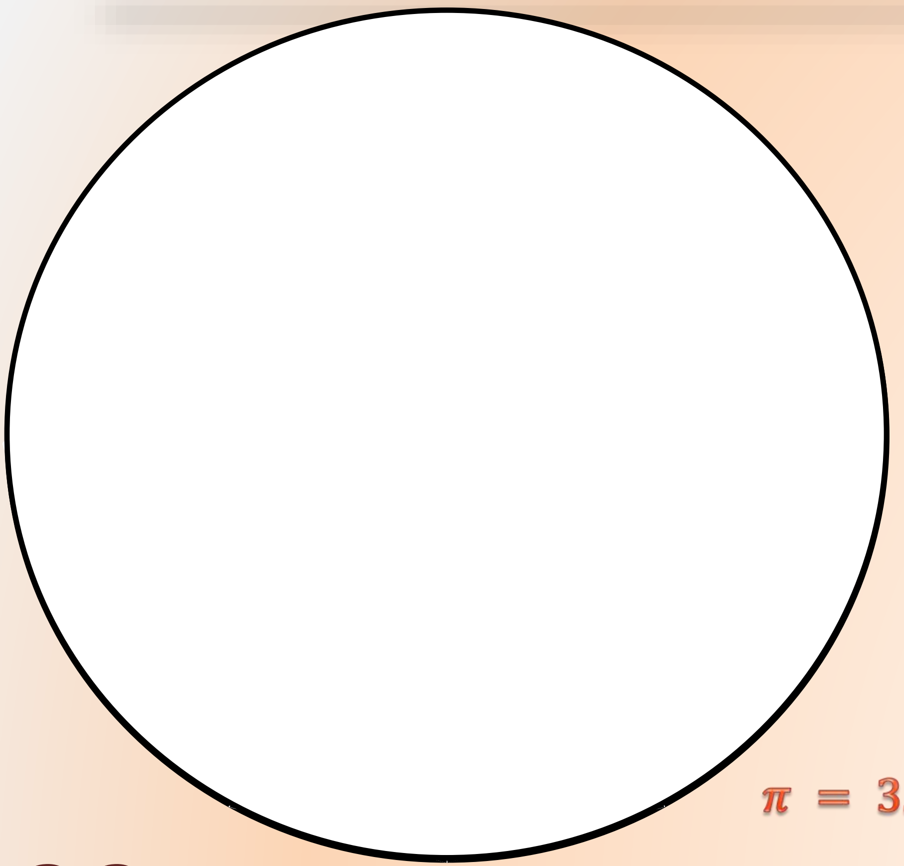
$\Pi = 3.14159265\ldots$

$e = 2.71828182\ldots$

${}^{1}/_{3} = 0.3333333\ldots$

Because of the limitation of showing numbers, we have inherent error!

Examples of Errors



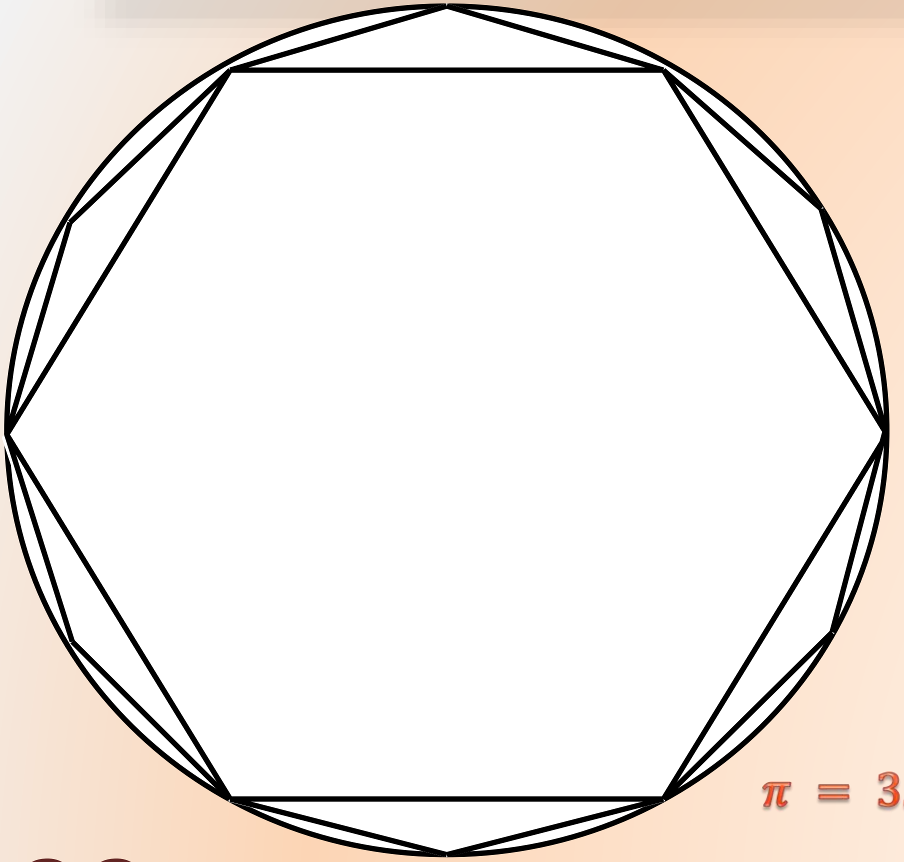
In Euclidean plane geometry, π is defined as the ratio of a circle's circumference (C) to its diameter (d).

$$\pi = \frac{C}{d}$$

As the number of sides of a polygon increases, its area approximates the area of a circle more accurately, showing that the value of π can be estimated with regular polygons.

$\pi = 3.14159265358979323846264338327950 28 \ldots$

Examples of Errors



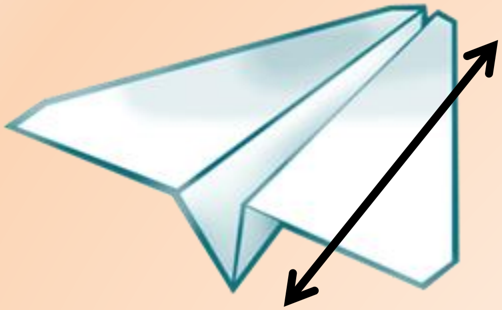
In Euclidean plane geometry, π is defined as the ratio of a circle's circumference (C) to its diameter (d).

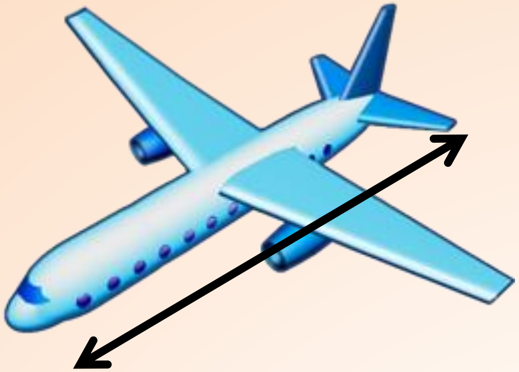
$$\pi = \frac{C}{d}$$

As the number of sides of a polygon increases, its area approximates the area of a circle more accurately, showing that the value of π can be estimated with regular polygons.

$\pi = 3.14159265358979323846264338327950 28 \ldots$

Examples of Errors





Actual Size	17.2 cm	82.49 m
Measured Size	18.2 cm	82.5 m

Sources of Error

- Measurement

  Measurement contains error.
- Mathematical Models

  Some parameters are ignored in mathematical modeling.
- Truncation Errors
- Roundoff Errors
- Operation Errors (propagation)

Error Representation

Absolute Relative Example

True Value → x

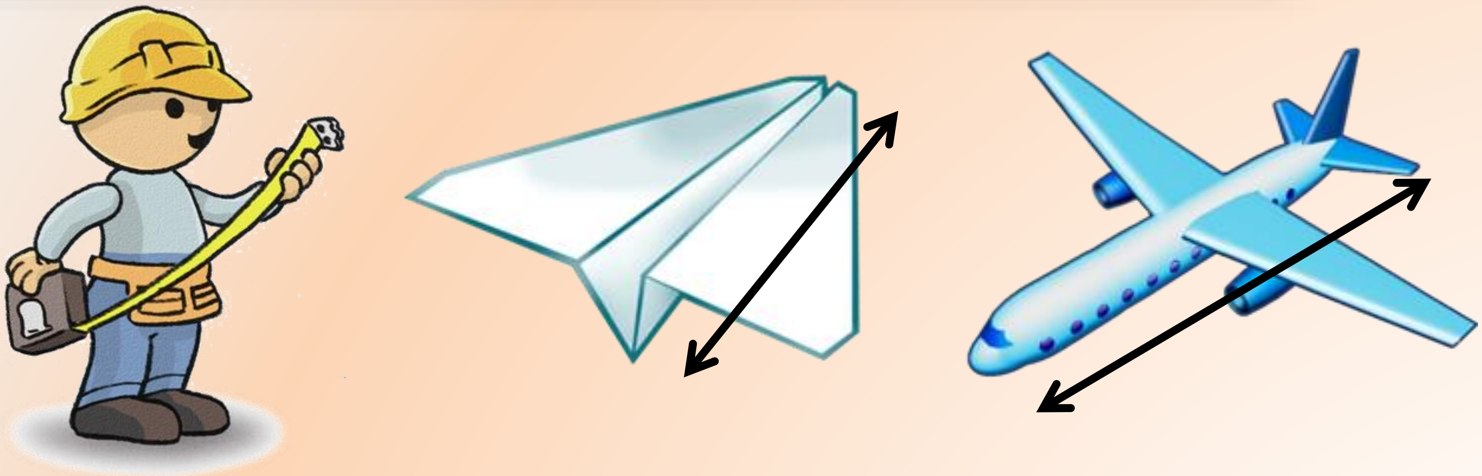
Approximate → $\bar{x}$

Absolute Error → $e(\bar{x}) = |x - \bar{x}|$

$$x \dashrightarrow \bar{x} \dashrightarrow e(\bar{x}) = |x - \bar{x}|$$

The error does not have sign (It is always positive)!

An Example



Actual Size	$17.2\ cm$	$82.49\ m$
Measured Size	$18.2\ cm$	$82.5\ m$
Absolute Error	$1\ cm$	$1\ cm$

Error Representation

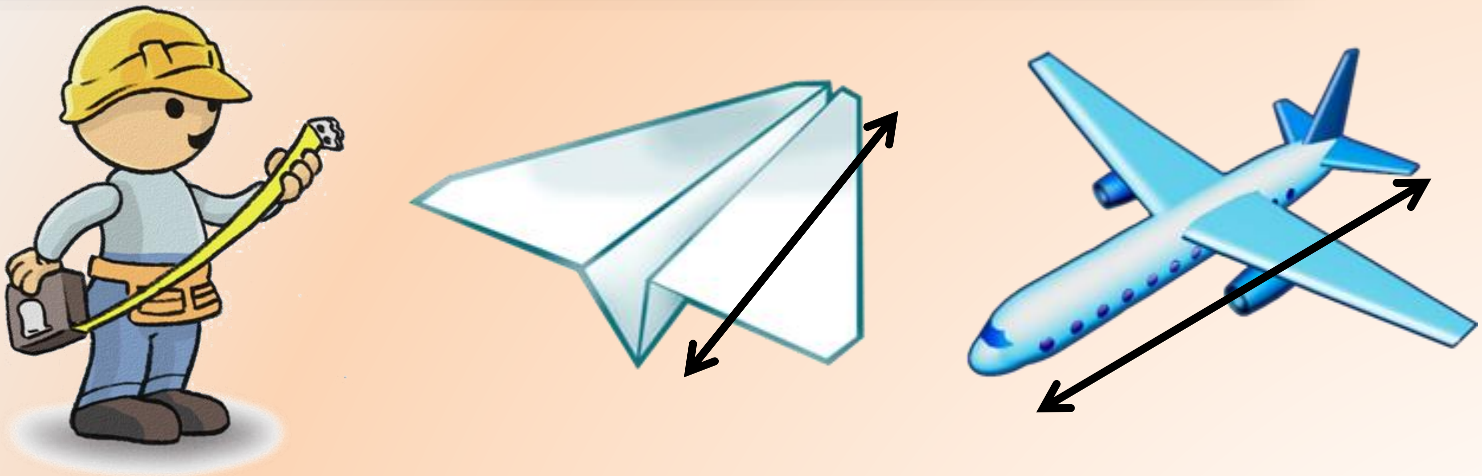
Absolute | **Relative** | Example

True Value | Approximate | Relative Error

$$x \dashrightarrow \bar{x} \dashrightarrow \delta(\bar{x}) = \left| \frac{e(\bar{x})}{x} \right|$$

The error does not have sign (It is always positive)!

An Example



Actual Size	$17.2\ cm$	$82.49\ m$
Measured Size	$18.2\ cm$	$82.5\ m$
Absolute Error	$1\ cm$	$1\ cm$
Relative Error	0.05	0.00012

Error Representation

Absolute Relative Example

$$e(\bar{x}) \leq e_{\bar{x}}$$

Example: $x = \sqrt{2}$

$\bar{x} = 1.41$

$e(\bar{x}) = |x - \bar{x}| = \left|\sqrt{2} - 1.41\right| < 0.005$

$$\boxed{|x - \bar{x}| \leq e_{\bar{x}} \longleftrightarrow x = \bar{x} \pm e_{\bar{x}}}$$

Error Representation

Absolute **Relative** Example

$$\delta(\bar{x}) = \frac{e(\bar{x})}{|x|} \leq \frac{e_{\bar{x}}}{|x|}$$

$$\delta(\bar{x}) \cong \frac{e_{\bar{x}}}{|\bar{x}|}$$

Representation of Floating-Point Numbers

$$23.1 = 2.31 \times 10^{1} = 0.231 \times 10^{2} = 0.0231 \times 10^{3}$$

$$231 \times 10^{-1} = 2310 \times 10^{-2} = 23100 \times 10^{-3}$$

Which form is normalized?

Representation of Floating-Point Numbers

$$23.1 = 2.31 \times 10^{1} = 0.231 \times 10^{2} = 0.0231 \times 10^{3}$$

$$231 \times 10^{-1} = 2310 \times 10^{-2} = 23100 \times 10^{-3}$$

Which form is normalized?

Normalized Representation

$$z = \sigma \times (a_0.a_1a_2a_3\ldots)_\beta \times \beta^e = \sigma \times m \times \beta^e$$

σ is the sign (+ or -).

β is the base, e is the exponent.

binary : $\beta = 2$

decimal : $\beta = 10$

$\boldsymbol{m}$ is mantissa ($significant$):

$1 \leq \boldsymbol{m} < \beta \quad , \quad (a_0 \neq 0 \;\; and \;\; 0 \leq a_i \leq \beta - 1)$

binary: $1 \leq m < 2$

decimal: $1 \leq m < 10$

Normalized Representation

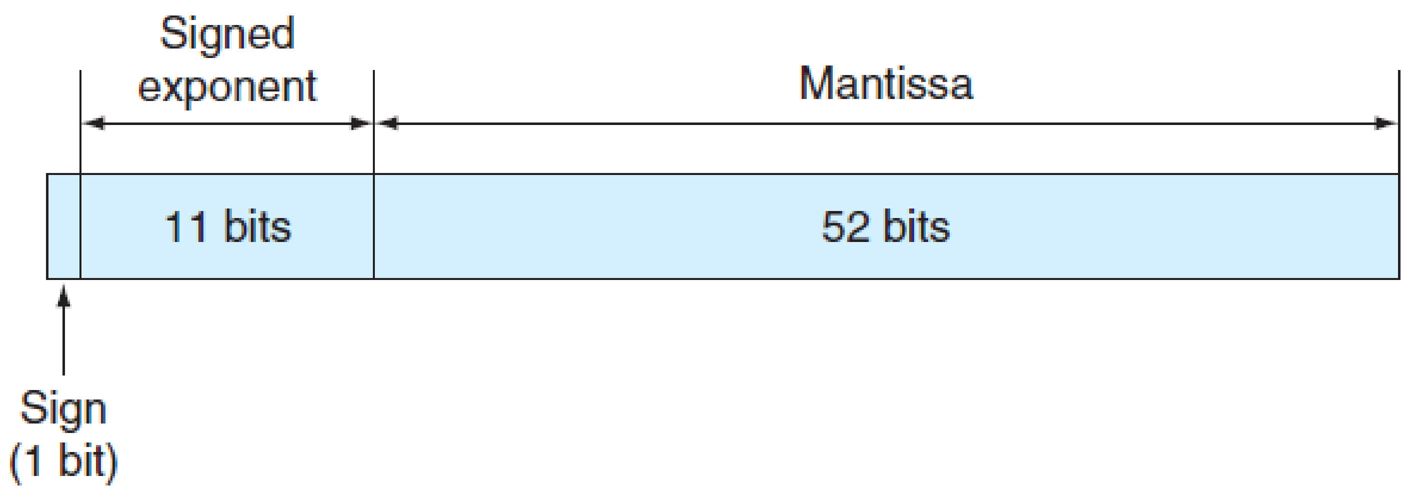
Example: z = 0.005678

decimal normalized : z = $5.678 * 10^{-3}$

$x = 3.5$ *(2 significant digits)* ➡ $3.45 \le x < 3.55$

IEEE-754 Double-Precision Format

The manner in which a floating-point number is stored in an 8-byte word in IEEE double precision format:



IEEE-754 Double-Precision Format

Since binary numbers consist exclusively of 0s and 1s, a bonus occurs when they are normalized:
The bit to the left of the binary point will always be one and does not have to be stored.

$$\pm(1 + f) \times 2^e$$

f **= the *mantissa* (i.e., the fractional part of the significand).**

Example:

Normalized the binary number 1101.1:

1.1011×2^{-3} or $(1+0.1011)\times2^{-3}$

only have to store the four fractional bits instance of five significant bits.

IEEE-754 Double-Precision Format

Range. In a fashion similar to the way in which integers are stored, the 11 bits used for the exponent translates into a range from -1022 to 1023. The largest positive number can be represented in binary as

$$\text{Largest value} = +1.1111\ldots1111 \times 2^{+1023}$$

where the 52 bits in the mantissa are all 1. Since the significand is approximately 2 (it is actually $2 - 2^{-52}$), the largest value is therefore $2^{1024} = 1.7977 \times 10^{308}$. In a similar fashion, the smallest positive number can be represented as

$$\text{Smallest value} = +1.0000\ldots0000 \times 2^{-1022}$$

This value can be translated into a base-10 value of $2^{-1022} = 2.2251 \times 10^{-308}$.

IEEE-754 Double-Precision Format

Precision. The 52 bits used for the mantissa correspond to about 15 to 16 base-10 digits. Thus, π would be expressed as

```
>> format long
>> pi

ans =
   3.14159265358979
```

Note that the machine epsilon is $2^{-52} = 2.2204 \times 10^{-16}$.

Types of Errors

Inherent   Round Off   Truncation

Length = 23.47

$$\Rightarrow \quad 23.465 < \text{length} < 23.475$$

Types of Errors

Inherent **Round Off** Truncation

Decimal : $\frac{1}{3} = 0.\overline{3}$

Binary : $(0.1)_{10} = (0.0\overline{0011})_2$

Types of Errors

Inherent   **Round Off**   Truncation

	Chopping	Symmetric
0.00065	0.0006	0.0007

Rounded off to 4 digits (4D).

Example (symmetric): 1.23456

1.23 (2D)

1.235 (3D)

1.2346 (4D)

Maximum Round Off Error

Absolute:

Chopping	$\|e_x\| < 10^{-t}$
Symmetric	$\|e_x\| \le \frac{1}{2} \times 10^{-t}$

Rounded off to t digits.

Types of Errors

Inherent | Round Off | **Truncation**

$$\sin x = x - \frac{x^{۳}}{۳!} + \frac{x^{۵}}{۵!} - \cdots + (-۱)^{n} \frac{x^{۲n+۱}}{(۲n+۱)!} \pm \cdots$$

$$e^{x} = ۱ + \frac{x}{۱!} + \frac{x^{۲}}{۲!} + \cdots + \frac{x^{n}}{n!} + E_{n}(x)$$

$$f(x_{i+1}) = f(x_i) + f'(x_i)h + \frac{f''(x_i)}{2!}h^2 + \frac{f^{(3)}(x_i)}{3!}h^3 + \cdots + \frac{f^{(n)}(x_i)}{n!}h^n + E_n$$

Types of Errors

Inherent Round Off **Truncation**

مثال ۱۴. مقدار تقریبی تابع $\sin x$ را به ازای $x = \frac{\pi}{۷}$ و با خطای کمتر از $۱۰^{-۳}$ حساب کنید.

حل : داریم

$$\sin x = x - \frac{x^۳}{۳!} + \frac{x^۵}{۵!} - \cdots + (-۱)^n \frac{x^{۲n+۱}}{(۲n+۱)!} \pm \cdots$$

در اینجا قرار می‌دهیم $|E_n(x)| = \frac{x^{۲n+۱}}{(۲n+۱)!}$

$$x = \frac{\pi}{۷} = \pi \frac{۱}{۷} = ۳/۱۴۱۶ \times ۰/۱۴۲۹ = ۰/۴۴۸۹$$

Types of Errors

Inherent Round Off **Truncation**

بنابراین بایستی n را طوری تعیین کنیم که

$$\frac{(۰/۴۴۸۹)^{۲n+۱}}{(۲n+۱)!} \leq \frac{۱}{۲} \times ۱۰^{-۳} = ۵ \times ۱۰^{-۴}$$

برای $n \geq ۲$ نامساوی فوق برقرار می‌باشد، در نتیجه

$$\sin\frac{\pi}{۷} \simeq ۰/۴۴۸۹ - \frac{(۰/۴۴۸۹)^{۳}}{۳!} + \frac{(۰/۴۴۸۹)^{۵}}{۵!}$$

$$= ۰/۴۴۸۹ - ۰/۰۱۵۱ + ۰/۰۰۰۲$$

$$= ۰/۴۳۴۰ \quad (۴D)$$

$$\sin\frac{\pi}{۷} \simeq ۰/۴۳۴(۳D)$$

Error Propagation

When does error propagation occur?

1) When we want to substitute parameters of formulas with non-exact values.

$$s = \pi r^2$$

2) When we have two algebraically equivalent equations and like to discover which one is better for implementation.

$$a^2 - b^2$$

$$(a-b)(a+b)$$

Error Propagation

Absolute Error Relative Error

Addition (+):

$\bar{x}$ and $\bar{y}$ are approximations of x and y $(\bar{x}, \bar{y} > 0)$

$|x - \bar{x}| \leq e_{\bar{x}}$ and $|y - \bar{y}| \leq e_{\bar{y}}$

$\bar{x} - e_{\bar{x}} \leq x \leq \bar{x} + e_{\bar{x}}$

$\bar{y} - e_{\bar{y}} \leq y \leq \bar{y} + e_{\bar{y}}$

$\bar{x} + \bar{y} - (e_{\bar{x}} + e_{\bar{y}}) \leq x + y \leq \bar{x} + \bar{y} + (e_{\bar{x}} + e_{\bar{y}})$

$|(x + y) - (\bar{x} + \bar{y})| \leq (e_{\bar{x}} + e_{\bar{y}})$

$$e_{\bar{x}+\bar{y}} \leq e_{\bar{x}} + e_{\bar{y}}$$

Error Propagation

Absolute Error Relative Error

Subtraction (-):

$\bar{x}$ and $\bar{y}$ are approximations of x and y $(\bar{x}, \bar{y} > 0)$

$|x - \bar{x}| \leq e_{\bar{x}}$ and $|y - \bar{y}| \leq e_{\bar{y}}$

$\bar{x} - e_{\bar{x}} \leq x \leq \bar{x} + e_{\bar{x}}$

$\bar{y} - e_{\bar{y}} \leq y \leq \bar{y} + e_{\bar{y}}$

$-\bar{y} - e_{\bar{y}} \leq -y \leq -\bar{y} + e_{\bar{y}}$

$\bar{x} - \bar{y} - (e_{\bar{x}} + e_{\bar{y}}) \leq x - y \leq \bar{x} - \bar{y} + (e_{\bar{x}} + e_{\bar{y}})$

$$e_{\bar{x}-\bar{y}} \leq e_{\bar{x}} + e_{\bar{y}}$$

Error Propagation

Absolute Error

Relative Error

$$e_{\bar{x}+\bar{y}} \leq e_{\bar{x}} + e_{\bar{y}}$$

$$e_{\bar{x}-\bar{y}} \leq e_{\bar{x}} + e_{\bar{y}}$$

Error Propagation

Absolute Error

Relative Error

مثال ۸. هرگاه اعداد $\sqrt{۱۷}$ و $\sqrt{۵}$ را تا سه رقم اعشار گرد کنیم، مطلوبست محاسبه $\sqrt{۱۷} \pm \sqrt{۵}$ و محاسبه حداکثر خطای حاصل جمع و تفاضل.

حل: داریم: $\sqrt{۱۷} = ۴٫۱۲۳ + e_۱, \qquad \sqrt{۵} = ۲٫۲۳۶ + e_۲$

منظور از $e_۱$ و $e_۲$ خطای مرتکب شده در نمایش $\sqrt{۱۷}$ و $\sqrt{۵}$ می‌باشد. چون اعداد تا سه رقم اعشار گرد شده‌اند، پس

$$e_۱ \le \frac{۱}{۲} \times ۱۰^{-۳}, \qquad e_۲ \le \frac{۱}{۲} \times ۱۰^{-۳}$$

داریم: $\sqrt{۱۷} + \sqrt{۵} = (۴٫۱۲۳ + ۲٫۲۳۶) + e_۳ = ۶٫۳۵۹ + e_۳$

و چون $e_۳ \le e_۱ + e_۲$ لذا $e_۳ \le ۱۰^{-۳}$. در نتیجه

$$۶٫۳۵۹ - ۱۰^{-۳} \le \sqrt{۱۷} + \sqrt{۵} \le ۶٫۳۵۹ + ۱۰^{-۳}$$

همچنین $\sqrt{۱۷} - \sqrt{۵} = ۱٫۸۸۷ + e_۴$ که در اینجا نیز

$$e_۴ \le e_۱ + e_۲ \le ۱۰^{-۳}$$

بنابراین

$$۱٫۸۸۷ - ۱۰^{-۳} \le \sqrt{۱۷} - \sqrt{۵} \le ۱٫۸۸۷ + ۱۰^{-۳}$$

Error Propagation

Absolute Error | **Relative Error**

Addition (+):

$\bar{x}$ and $\bar{y}$ are approximations of x and y $(\bar{x}, \bar{y} > 0)$

$$\delta_{\bar{x}} \cong \frac{e_{\bar{x}}}{\bar{x}} \quad and \quad \delta_{\bar{y}} \cong \frac{e_{\bar{y}}}{\bar{y}}$$

$$\delta_{\bar{x}+\bar{y}} \leq \frac{e_{\bar{x}+\bar{y}}}{\bar{x}+\bar{y}} \leq \frac{e_{\bar{x}} + e_{\bar{y}}}{\bar{x}+\bar{y}} = \frac{\bar{x}}{\bar{x}+\bar{y}} * \frac{e_{\bar{x}}}{\bar{x}} + \frac{\bar{y}}{\bar{x}+\bar{y}} * \frac{e_{\bar{y}}}{\bar{y}} = \frac{\bar{x}}{\bar{x}+\bar{y}} \delta_{\bar{x}} + \frac{\bar{y}}{\bar{x}+\bar{y}} \delta_{\bar{y}}$$

$$\delta_{\bar{x}+\bar{y}} \leq \frac{\bar{x}}{\bar{x}+\bar{y}} \delta_{\bar{x}} + \frac{\bar{y}}{\bar{x}+\bar{y}} \delta_{\bar{y}}$$

Error Propagation

Absolute Error | **Relative Error**

Subtraction (-):

$\bar{x}$ and $\bar{y}$ are approximations of x and y $(\bar{x},\bar{y} > 0)$

$$\delta_{\bar{x}} \cong \frac{e_{\bar{x}}}{\bar{x}} \qquad and \qquad \delta_{\bar{y}} \cong \frac{e_{\bar{y}}}{\bar{y}}$$

$$\delta_{\bar{x}-\bar{y}} \leq \frac{e_{\bar{x}-\bar{y}}}{\bar{x}-\bar{y}} \leq \frac{e_{\bar{x}}+e_{\bar{y}}}{\bar{x}-\bar{y}} = \frac{\bar{x}}{\bar{x}-\bar{y}} * \frac{e_{\bar{x}}}{\bar{x}} + \frac{\bar{y}}{\bar{x}-\bar{y}} * \frac{e_{\bar{y}}}{\bar{y}} = \frac{\bar{x}}{\bar{x}-\bar{y}}\delta_{\bar{x}} + \frac{\bar{y}}{\bar{x}-\bar{y}}\delta_{\bar{y}}$$

$$\delta_{\bar{x}-\bar{y}} \leq \frac{\bar{x}}{\bar{x}-\bar{y}}\delta_{\bar{x}} + \frac{\bar{y}}{\bar{x}-\bar{y}}\delta_{\bar{y}}$$

Error Propagation

Absolute Error

Relative Error

$$\delta_{\bar{x}+\bar{y}} \leq \frac{\bar{x}}{\bar{x}+\bar{y}}\delta_{\bar{x}} + \frac{\bar{y}}{\bar{x}+\bar{y}}\delta_{\bar{y}} \qquad \bar{x},\bar{y} > 0$$

$$\delta_{\bar{x}-\bar{y}} \leq \frac{\bar{x}}{\bar{x}-\bar{y}}\delta_{\bar{x}} + \frac{\bar{y}}{\bar{x}-\bar{y}}\delta_{\bar{y}} \qquad \bar{x} > \bar{y} > 0$$

Nearly identical amounts for $\bar{x}$ and $\bar{y}$ increase error propagation.

Error Propagation

Absolute Error Relative Error

Multiplication (*):

$\bar{x}$ and $\bar{y}$ are approximations of x and y $(\bar{x}, \bar{y} > 0)$

$|x - \bar{x}| \leq e_{\bar{x}}$ and $|y - \bar{y}| \leq e_{\bar{y}}$

$\bar{x} - e_{\bar{x}} \leq x \leq \bar{x} + e_{\bar{x}}$

$\bar{y} - e_{\bar{y}} \leq y \leq \bar{y} + e_{\bar{y}}$

$\bar{x}\bar{y} - (\bar{y}e_{\bar{x}} + \bar{x}e_{\bar{y}}) + e_{\bar{x}}e_{\bar{y}} \leq xy \leq \bar{x}\bar{y} + (ye_{\bar{x}} + xe_{\bar{y}}) + e_{\bar{x}}e_{\bar{y}}$

$$e_{\bar{x}\bar{y}} \leq \bar{y}e_{\bar{x}} + \bar{x}e_{\bar{y}}$$

Error Propagation

Absolute Error   Relative Error

Division (/):

$\bar{x}$ and $\bar{y}$ are approximations of x and y $(\bar{x}, \bar{y} > 0)$

$|x-\bar{x}| \le e_{\bar{x}}$ and $|y-\bar{y}| \le e_{\bar{y}}$

$$\bar{x} - e_{\bar{x}} \le x \le \bar{x} + e_{\bar{x}}$$

$$\bar{y} - e_{\bar{y}} \le y \le \bar{y} + e_{\bar{y}}$$

$$\frac{\bar{x} - e_{\bar{x}}}{\bar{y} + e_{\bar{y}}} \le \frac{x}{y} \le \frac{\bar{x} + e_{\bar{x}}}{\bar{y} - e_{\bar{y}}}$$

$$\frac{\bar{x} - e_{\bar{x}}}{\bar{y} + e_{\bar{y}}} * \frac{\bar{y} - e_{\bar{y}}}{\bar{y} - e_{\bar{y}}} = \frac{\bar{x}\bar{y} - \bar{y}e_{\bar{x}} - \bar{x}e_{\bar{y}} + e_{\bar{x}}e_{\bar{y}}}{\bar{y}^2 - {e_{\bar{y}}}^2} = \frac{\bar{x}}{\bar{y}} - \frac{\bar{y}e_{\bar{x}} + \bar{x}e_{\bar{y}}}{\bar{y}^2}$$

$$\frac{\bar{x} + e_{\bar{x}}}{\bar{y} - e_{\bar{y}}} * \frac{\bar{y} + e_{\bar{y}}}{\bar{y} + e_{\bar{y}}} = \frac{\bar{x}\bar{y} + \bar{y}e_{\bar{x}} + \bar{x}e_{\bar{y}} + e_{\bar{x}}e_{\bar{y}}}{\bar{y}^2 - {e_{\bar{y}}}^2} = \frac{\bar{x}}{\bar{y}} + \frac{\bar{y}e_{\bar{x}} + \bar{x}e_{\bar{y}}}{\bar{y}^2}$$

$$e_{\frac{\bar{x}}{\bar{y}}} \le \frac{\bar{y}e_{\bar{x}} + \bar{x}e_{\bar{y}}}{\bar{y}^2}$$

Error Propagation

Absolute Error

Relative Error

$$e_{\bar{x}\times\bar{y}} \leq e_{\bar{x}} \times |\bar{y}| + e_{\bar{y}} \times |\bar{x}|$$

$$e_{\frac{\bar{x}}{\bar{y}}} \leq \frac{|\bar{y}|e_{\bar{x}} + |\bar{x}|e_{\bar{y}}}{|\bar{y}|^2}$$

$$\bar{x}, \bar{y} > 0$$

Absolute error is too sensitive to the value of the parameters shown:

- Large amounts for $\bar{x}, \bar{y}$ increase error propagation in multiplication
- Small amounts for $\bar{y}$ increase error propagation in division.

Error Propagation

Absolute Error   Relative Error

مثال ۱۰. مقدار $\pi\sqrt{۲}$ را با چهار رقم اعشار محاسبه نموده و حداکثر خطای این حاصل ضرب را نیز به دست آورید.

حل : داریم:

$$\pi = ۳/۱۴۱۶ + e_۱$$

$$\sqrt{۲} = ۱/۴۱۴۲ + e_۲$$

$$e_۱ \leq \frac{۱}{۲} \times ۱۰^{-۴}, \qquad e_۲ \leq \frac{۱}{۲} \times ۱۰^{-۴}$$

$$\pi\sqrt{۲} = (۳/۱۴۱۶ \times ۱/۴۱۴۲) + e_۳$$

Error Propagation

Absolute Error

Relative Error

$$e_{۳} \leq ۳٫۱۴۱۶ e_{۲} + ۱٫۴۱۴۲ e_{۱}$$

$$e_{۳} \leq \frac{۱}{۲} \times ۱۰^{-۴}(۳٫۱۴۱۶ + ۱٫۴۱۴۲)$$

$$e_{۳} \leq ۰٫۵ \times ۱۰^{-۴}(۴٫۵۵۵۸) = ۲٫۲۷۷۹ \times ۱۰^{-۴}$$

اما

$$\pi\sqrt{۲} = ۴٫۴۴۲۹ + e'_{۳}$$

Error Propagation

Absolute Error

Relative Error

چون حاصل‌ضرب اعداد ۳٫۱۴۱۶ و ۱٫۴۱۴۲ در محاسبهٔ $\pi\sqrt{۲}$ بیشتر از چهار رقم اعشار دارد، هنگام نمایش حاصل‌ضرب دو عدد مذکور با چهار رقم اعشار خطای دیگری مرتکب شده‌ایم و خطای حدی کل را با $e'_۴$ نشان داده‌ایم. برای $e'_۴$ داریم:

$$e'_۴ \leq \frac{۱}{۲} \times ۱۰^{-۴} + e_۴$$

$$e'_۴ \leq ۰٫۵ \times ۱۰^{-۴} + ۲٫۲۷۷۹ \times ۱۰^{-۴} = ۲٫۷۷۷۹ \times ۱۰^{-۴}$$

$$۴٫۴۴۲۹ - ۲٫۷۷۷۹ \times ۱۰^{-۴} \leq \pi\sqrt{۲} \leq ۴٫۴۴۲۹ + ۲٫۷۷۷۹ \times ۱۰^{-۴}$$

$$۴٫۴۴۲۶ \leq \pi\sqrt{۲} \leq ۴٫۴۴۳۲$$

Error Propagation

Absolute Error Relative Error

Multiplication (*):

$\bar{x}$ and $\bar{y}$ are approximations of x and y $(\bar{x}, \bar{y} > 0)$

$$\delta_{\bar{x}} \cong \frac{e_{\bar{x}}}{\bar{x}} \qquad and \qquad \delta_{\bar{y}} \cong \frac{e_{\bar{y}}}{\bar{y}}$$

$$\delta_{\bar{x}\bar{y}} \leq \frac{e_{\bar{x}\bar{y}}}{\bar{x}\bar{y}} \leq \frac{\bar{y}e_{\bar{x}} + \bar{x}e_{\bar{y}}}{\bar{x}\bar{y}} = \frac{e_{\bar{x}}}{\bar{x}} + \frac{e_{\bar{y}}}{\bar{y}} = \delta_{\bar{x}} + \delta_{\bar{y}}$$

$$\delta_{\bar{x}\bar{y}} \leq \delta_{\bar{x}} + \delta_{\bar{y}}$$

Error Propagation

Absolute Error | **Relative Error**

Division (/):

$\bar{x}$ and $\bar{y}$ are approximations of x and y $(\bar{x}, \bar{y} > 0)$

$$\delta_{\bar{x}} \cong \frac{e_{\bar{x}}}{\bar{x}} \qquad and \qquad \delta_{\bar{y}} \cong \frac{e_{\bar{y}}}{\bar{y}}$$

$$\delta_{\frac{\bar{x}}{\bar{y}}} \le \frac{\frac{e_{\bar{x}}}{\bar{y}}}{\frac{\bar{x}}{\bar{y}}} \le \frac{\frac{\bar{y}e_{\bar{x}}+\bar{x}e_{\bar{y}}}{\bar{y}^2}}{\frac{\bar{x}}{\bar{y}}} = \frac{\frac{\bar{x}}{\bar{y}}\left(\frac{e_{\bar{x}}}{\bar{x}}+\frac{e_{\bar{y}}}{\bar{y}}\right)}{\frac{\bar{x}}{\bar{y}}} = \frac{e_{\bar{x}}}{\bar{x}} + \frac{e_{\bar{y}}}{\bar{y}} = \delta_{\bar{x}} + \delta_{\bar{y}}$$

$$\delta_{\frac{\bar{x}}{\bar{y}}} \le \delta_{\bar{x}} + \delta_{\bar{y}}$$

Error Propagation

Absolute Error

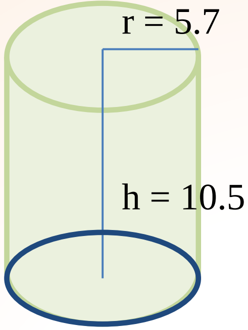
Relative Error

$\delta_{\frac{\bar{x}}{\bar{y}}} \leq \delta_{\bar{x}} + \delta_{\bar{y}}$ $\qquad \bar{y} \neq 0$

$\delta_{\bar{x}\bar{y}} \leq \delta_{\bar{x}} + \delta_{\bar{y}}$ $\qquad \bar{x}, \bar{y} > 0$

An Example

Suppose a cylinder with a radius of 5.7 cm and a height of 10.5 cm. Estimate the absolute and relative errors of calculating the volume of it considering that all the values have been rounded using chopping model.(consider $\pi = 3.14$)



An Example

Suppose a cylinder with a radius of 5.7 cm and a height of 10.5 cm. Estimate the absolute and relative errors of calculating the volume of it considering that all the values have been rounded using chopping model.(consider π = 3.14)

Solution :

V = $h\,\pi\,r^2$

π **is symmetrically rounded off into two floating digits,**

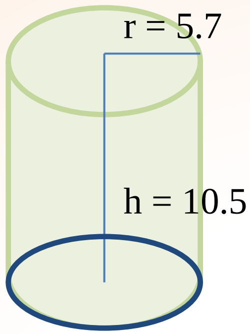
thus $\rightarrow e_\pi \leq 0.5 \times 10^{-2}$

h **and** r **are measured by a device with the maximum error of** 10^{-1}

thus $\rightarrow\ e_r \leq 10^{-1}$ **and** $e_h \leq 10^{-1}$

$$e(x \times y) \leq |x|e_y + |y|e_x \rightarrow \begin{cases} e(h\pi) \leq 10.5 \times 0.5 \times 10^{-2} + 3.14 \times 10^{-1} = 0.3665 \\ e(r \times r) \leq 2|r|e_r = 1.14 \end{cases}$$

$$\delta(xy) \leq \delta_x + \delta_y \rightarrow \begin{cases} \delta(h\pi) \leq 0.111162 \times 10^{-1} \\ \delta(r^2) \leq \delta_r + \delta_r = 0.35087 \times 10^{-1} \end{cases}$$



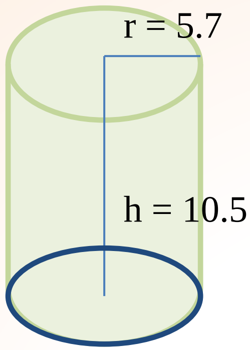
$$e_v = e_{h\pi \times r^2} \leq |h\pi|e_{r^2} + |r^2| \times e_{h\pi} = 0.494933 \times 10^2\ ,\ \delta_v \leq 0.46204 \times 10^{-1}$$

An Example

Suppose a cylinder with a radius of 5.7 cm and a height of 10.5 cm. Estimate the absolute and relative errors of calculating the volume of it considering that all the values have been rounded.

Solution :

	$value$	$max(e)$	$max(\delta)$
h	10.5	0.1	0.00952
π	3.14	0.005	0.00159
r	5.7	0.01	0.01754
$\mathrm{h}\pi$	32.97	0.3665	0.01111
$\mathrm{r}^2 = \mathrm{r} \times \mathrm{r}$	32.49	1.14	0.03508
$\mathrm{v} = (\mathrm{h}\pi).(\mathrm{r}^2)$	1071.1953	49.49338	0.046204

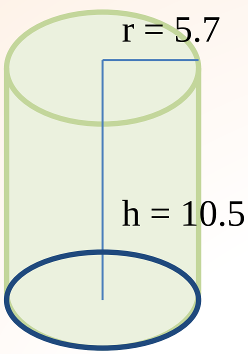


An Example

Suppose a cylinder with a radius of 5.7 cm and a height of 10.5 cm. Estimate the absolute and relative errors of calculating the volume of it considering that all the values have been rounded.

Solution :

	$value$	$max(e)$	$max(\delta)$
h	10.5	0.1	0.00952
π	3.14	0.005	0.00159
r	5.7	0.01	0.01754
$h\pi$	32.97	0.3665	0.01111
$r^2 = r \times r$	32.49	1.14	0.03508
$v = (h\pi).(r^2)$	1071.1953	49.49338	0.046204



What if our computer supports just 3 digits mantissa?

An Example

Suppose a cylinder with a radius of 5.7 cm and a height of 10.5 cm. Estimate the absolute and relative errors of calculating the volume of it considering that all the values have been rounded.

Solution :

	$value$	$max(e)$	$max(\delta)$
h	10.5	$0.1 + \lvert 10.5 - 10.5\rvert$	$0.00952 + \frac{\lvert 10.5 - 10.5\rvert}{10.5}$
π	3.14	$0.005 + \lvert 3.14 - 3.14\rvert$	$0.00159 + \frac{\lvert 3.14 - 3.14\rvert}{3.14}$
r	5.7	$0.1 + \lvert 5.7 - 5.7\rvert$	$0.01754 + \frac{\lvert 5.7 - 5.7\rvert}{5.7}$
$h\pi$	33	$0.3665 + \lvert 33 - 32.97\rvert$	$0.011116 + \frac{\lvert 33 - 32.97\rvert}{33}$
$r^2 = r \times r$	32.5	$1.14 + \lvert 32.5 - 32.49\rvert$	$0.035087 + \frac{\lvert 32.5 - 32.49\rvert}{32.5}$
$v = (h\pi).(r^2)$	107	$49.4933 + \lvert 107 - 1071.1953\rvert$	$0.04620 + \frac{\lvert 107 - 1071.1953\rvert}{107}$

Formula Error

The Taylor series of $f(x)$ at a number a:

$$f(x) = f(a) + (x-a)f'(a) + \frac{(x-a)^2}{2!}f''(a) + \cdots$$

$$|f(x) - f(a)| \cong |x-a||f'(a)| = e(a)|f'(a)|$$

$$e_f \leq e_a|f'(a)|$$

Formula Error

The Taylor series of $f(x_1, x_2)$ at (a_1, a_2):

$$f(x_1, x_2) = f(a_1, a_2) + (x_1 - a_1)\frac{\partial f(a_1, a_2)}{\partial x_1} + (x_2 - a_2)\frac{\partial f(a_1, a_2)}{\partial x_2} + \cdots$$

$$|f(x_1, x_2) - f(a_1, a_2)| \cong e(a_1)\left|\frac{\partial f(a_1, a_2)}{\partial x_1}\right| + e(a_2)\left|\frac{\partial f(a_1, a_2)}{\partial x_2}\right|$$

$$e_f \leq e_{a_1}\left|\frac{\partial f(a_1, a_2)}{\partial x_1}\right| + e_{a_2}\left|\frac{\partial f(a_1, a_2)}{\partial x_2}\right|$$

Formula Error

Error of $f(x_1, x_2, \dots, x_n)$ at $\bar{a}$=$(a_1, a_2, \dots, a_n)$:

$$e_f = |f(x_1, x_2, \dots, x_n) - f(a_1, a_2, \dots, a_n)| \leq$$

$$e_{a_1} \left| \frac{\partial f}{\partial x_1} \right|_{\bar{a}} + e_{a_2} \left| \frac{\partial f}{\partial x_2} \right|_{\bar{a}} + \dots + e_{a_n} \left| \frac{\partial f}{\partial x_n} \right|_{\bar{a}}$$

An Example

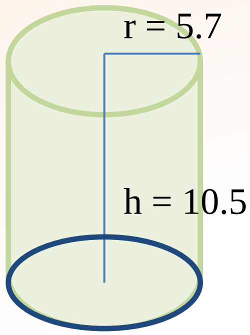
Suppose a cylinder with a radius of 5.7 cm and a height of 10.5 cm. Estimate the absolute and relative errors of calculating the volume of it considering that all the values have been rounded using chopping model.(consider π = 3.14)

Solution :

V = $h\,\pi\,r^2$

$f = x\,y\,z^2$

$$e_h = e_x \le 10^{-1}$$
$$e_\pi = e_y \le 0.5 \times 10^{-2}$$
$$e_r = e_z \le 10^{-1}$$



$$e_f \le e_x\, yz^2 + e_y\, xz^2 + e_z\, 2xyz =$$

$$10^{-1} \times 3.14 \times (5.7)^2 + 0.5 \times 10^{-2} \times 10.5 \times (5.7)^2 +$$

$$10^{-1} \times 2 \times 10.5 \times 3.14 \times 5.7 = 0.494933 \times 10^2$$

$$\delta_f \le \frac{e_f}{|f(\bar{a})|} = 0.46204 \times 10^{-1}$$

An Example

Compute the following expression with 4 digits mantissa and symmetric round-off for $x = 3.209$.

$$1.076x^3 + 0.319x^2 - 0.017x + 1.107$$

a) From left to right.

b) From right to left.

c) Compute the exact value.

d) What is the difference and why?

An Example

Compute the following expression with 4 digits mantissa and symmetric round-off for $x = 3.209$.

$$1.076x^3 + 0.319x^2 - 0.017x + 1.107$$

a) From left to right.

b) From right to left.

c) Compute the exact value.

d) What is the difference and why?

Solution:

$$((1.076 \times 3.209) \times 3.209) \times 3.209 = 35.56$$

$$(0.319 \times 3.209) \times 3.209 = 3.286$$

$$0.017 \times 3.209 = 0.054553 \rightarrow 0.05455$$

$$1.107 \rightarrow 1.107$$

An Example

$$1.076x^3 + 0.319x^2 - 0.017x + 1.107$$

a) From left to right $\rightarrow 39.91$
b) From right to left
c) Compute the exact value
d) What is the difference and why?

Solution:

$$1.076x^3 \rightarrow 35.56$$
$$0.319x^2 \rightarrow 3.286$$
$$0.017x \rightarrow 0.05455$$
$$1.107 \rightarrow 1.107$$

$$35.56 + 3.286 = 38.85$$
$$38.85 - 0.05455 = 38.80$$
$$38.80 + 1.107 = 39.91$$

An Example

$$1.076x^3 + 0.319x^2 - 0.017x + 1.107$$

a) From left to right $\rightarrow 39.91$

b) From right to left

c) Compute the exact value

d) What is the difference and why?

Solution:

$$\big(3.209 \times (3.209 \times 3.209)\big) \times 1.076 = 35.56$$

$$(3.209 \times 3.209) \times 0.319 = 3.286$$

$$3.209 \times 0.017 = 0.054553 \rightarrow 0.05455$$

$$1.107 \rightarrow 1.107$$

An Example

$$1.076x^3 + 0.319x^2 - 0.017x + 1.107$$

a) From left to right $\rightarrow 39.91$
b) From right to left $\rightarrow 39.90$
c) Compute the exact value
d) What is the difference and why?

Solution:

$1.076x^3 \rightarrow 35.56$

$0.319x^2 \rightarrow 3.286$

$0.017x \rightarrow 0.05455$

$1.107 \rightarrow 1.107$

$1.107 - 0.05455 = 1.052$

$1.052 + 3.286 = 4.338$

$4.338 + 35.56 = 39.90$

An Example

$$1.076x^3 + 0.319x^2 - 0.017x + 1.107$$

a) From left to right $\rightarrow 39.91$
b) From right to left $\rightarrow 39.90$
c) Compute the exact value $\rightarrow 39.894105201004$
d) What is the difference and why?

Solution:

$1.076x^3 \rightarrow 35.56$
$0.319x^2 \rightarrow 3.285$
$0.017x \rightarrow 0.05455$
$1.107 \rightarrow 1.107$

$$1.076x^3 + 0.319x^2 - 0.017x + 1.107 = 39.894105201004$$

An Example

$$1.076x^3 + 0.319x^2 - 0.017x + 1.107$$

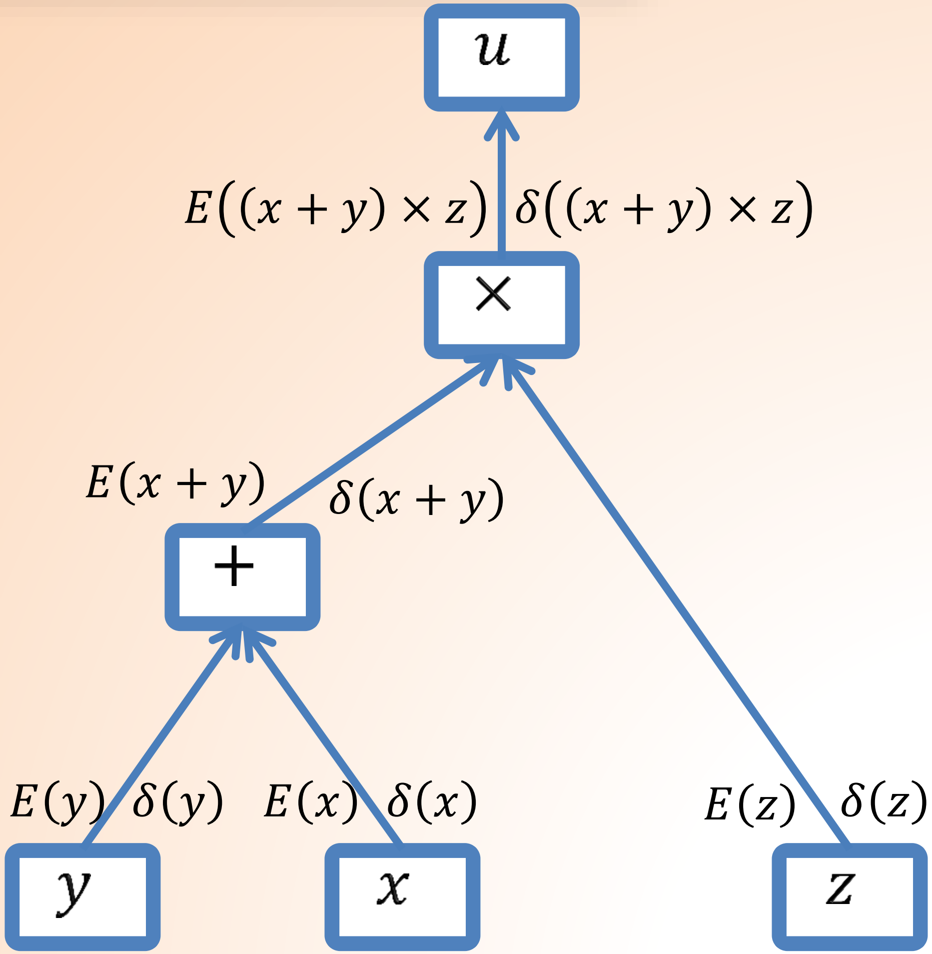
a) From left to right $\rightarrow 39.91$
b) From right to left $\rightarrow 39.90$
c) Compute the exact value
d) What is the difference and why?

Solution:

We better initially deal with the least significant numbers in any computational system where the number of digits are limited, i.e. small numbers show themselves better if used prior to others.

Process Graph

$$u = (x + y) * z$$



An Example

Draw process graph of $\mathrm{v} = \pi r^2 h$

1. From left to right
2. From right to left

An Example

Draw process graph of $\mathrm{v} = \pi r^2 h$

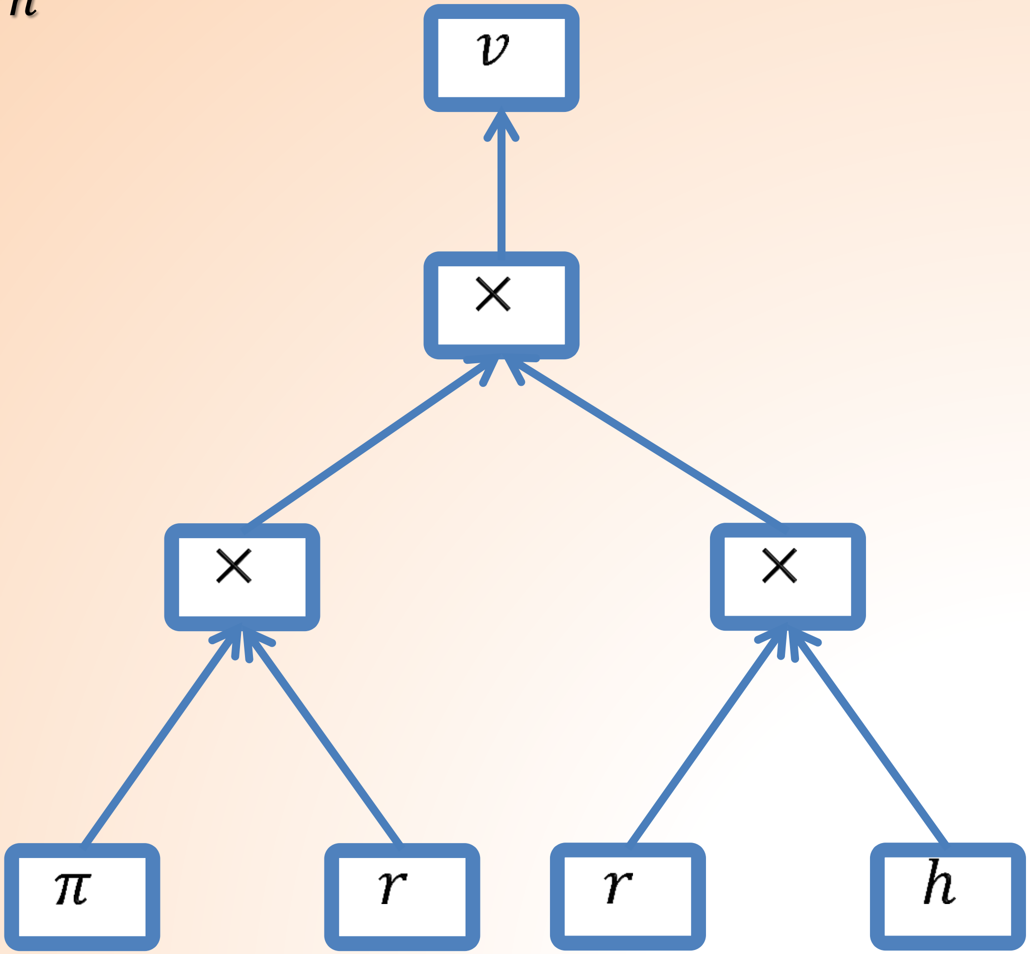
1. From left to right
2. From right to left

Solution :

$\mathrm{v} = \pi r^2 h$

$= \pi r r h$

$= (\pi . r) . (r . h)$



An Example

Draw process graph of $v = \pi r^2 h$

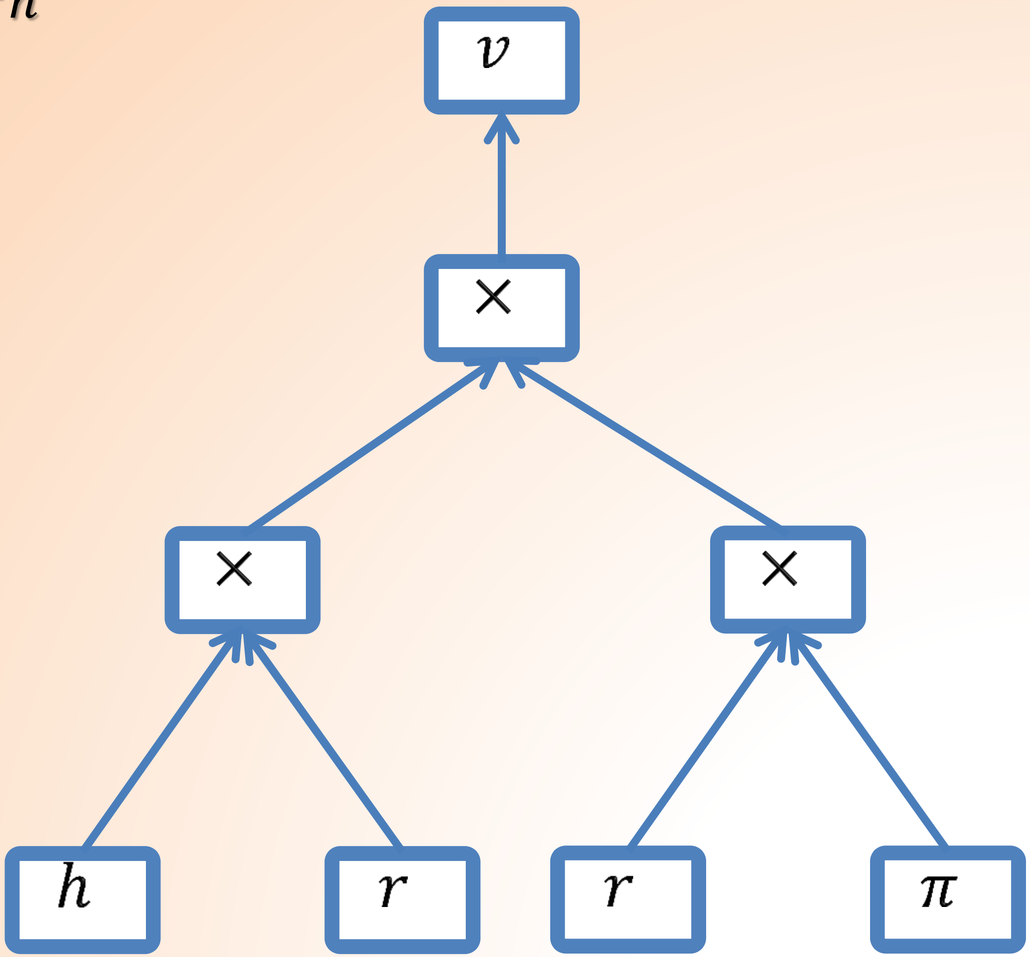
1. From left to right
2. From right to left

Solution :

$v = \pi r^2 h$

$= hrr\pi$

$= (h.r).(r.\pi)$



Stability

Algorithm (method)

- Stable : $E_n \approx cE_0$ (linearly)
- Unstable : $E_n \approx c^n E_0$ c≥ 1 (exponentially)

problem

- Inherent unstable ⟶ Example: Wilkinson problem,
- Induced unstable

Wilkinson problem: roots of

$$P_{20}(x) = (x-1)(x-2)\ldots(x-20) = x^{20} - 210x^{19} + \cdots + 20!$$


Any questions?