Numerical Computations

Hamid Sarbazi-Azad &
Samira Hossein Ghorban
Department of Computer Engineering
Sharif University of Technology (SUT)
Tehran, Iran

Example of errors

$\Pi = 3.14159265\ldots$

$e = 2.71828182\ldots$

$\sqrt{5} = 2.2360679\ldots$

$^1/_3 = 0.3333333\ldots$

Error

- For many engineering problems, we cannot obtain analytical solutions.
- Numerical methods yield approximate results, results that are close to the exact analytical solution.
- How confident we are in our approximate result?
- The question is
    - How much error is present in our calculation and if it is tolerable…

Chapter Topics

- Why measure errors?
- Sources of error
- Types of errors
- Accuracy vs. Precision
- Error in arithmetic operations

Why measure errors?

- To determine the accuracy of numerical results
- To develop stopping criteria for iterative algorithms

Problems Created by Round Off Error

- 28 Americans were killed on February 25, 1991, by an Iraqi Scud missile in Dhahran, Saudi Arabia.*
- The patriot defense system failed to track and intercept the Scud. Why?

https://www.iro.umontreal.ca/~mignotte/IFT2425/Disasters.html

Problem with Patriot Missile

- The problem was in the differencing of floating-point numbers obtained by converting and scaling an integer timing register
- Clock cycle of 1/10 seconds was represented in 24-bit fixed-point register created an error of $\mathbf{9.5 \times 10^{-8}}$ seconds.
- The battery was on for **100** consecutive hours, thus causing an inaccuracy of

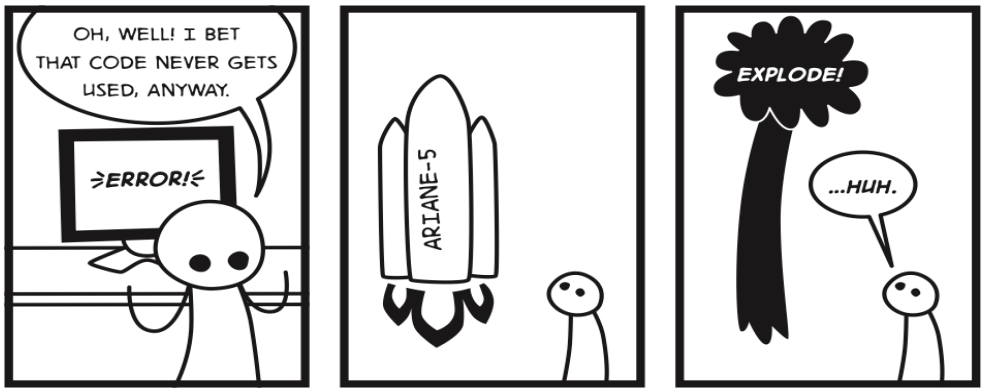
$$\mathbf{9.5 \times 10^{-8}} \frac{s}{0.1\ s} \times \underbrace{\mathbf{100}hr \times \frac{\mathbf{3600}s}{1hr}} = 0.342\boldsymbol{s}$$

The Short Flight of Ariane 5

- On June 4, 1996, the first Ariane 5 was launched. All went well for 36 seconds. Then the Ariane veered off course and self-destructed. Why?*

The Short Flight of Ariane 5

- The problem was in the Inertial Reference System, which produced an operation exception trying *to convert a 64-bit floating-point number to a 12-bit integer*. It sent a diagnostic word to the On-Board Computer, which interpreted it as flight data. Ironically, the computation was done by legacy software from the Ariane 4, and its results were not needed after lift-off.



The Vancouver Stock Exchange*

- In 1982, the Vancouver Stock Exchange instituted a new index initialized to a value of 1000.
- The index was updated after each transaction.
- Twenty two months later it had fallen to 520.
- The cause was that the updated value was truncated rather than rounded. The rounded calculation gave a value of 1098.892.

*The Wall Street Journal November 8, 1983, p.37.
*The Toronto Star, November 19, 1983.
*B.D. McCullough and H.D. Vinod, Journal of Economic Literature, Vol XXXVII (June 1999), pp. 633-665

Sources of error

- Measurement
  - Measurement contains error.
- Representation Manner
- Mathematical Models
  - Some parameters are ignored in the model.

Types of Error

Round off errors

- Due to the fact that computers can work only with a finite representation of numbers

$$\frac{1}{3} \cong 0.33333 \qquad \sqrt{2} \cong 1.4142$$

Truncation errors

- Approximating functions using polynomials
- For example, using Taylor series to approximate function

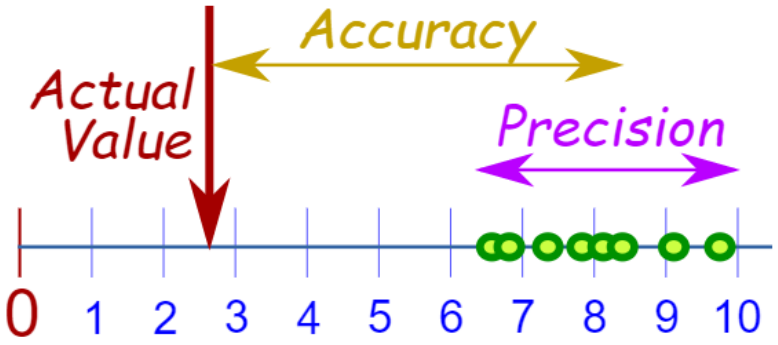
for $f(x) = e^x$ about point 0 is given by:

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

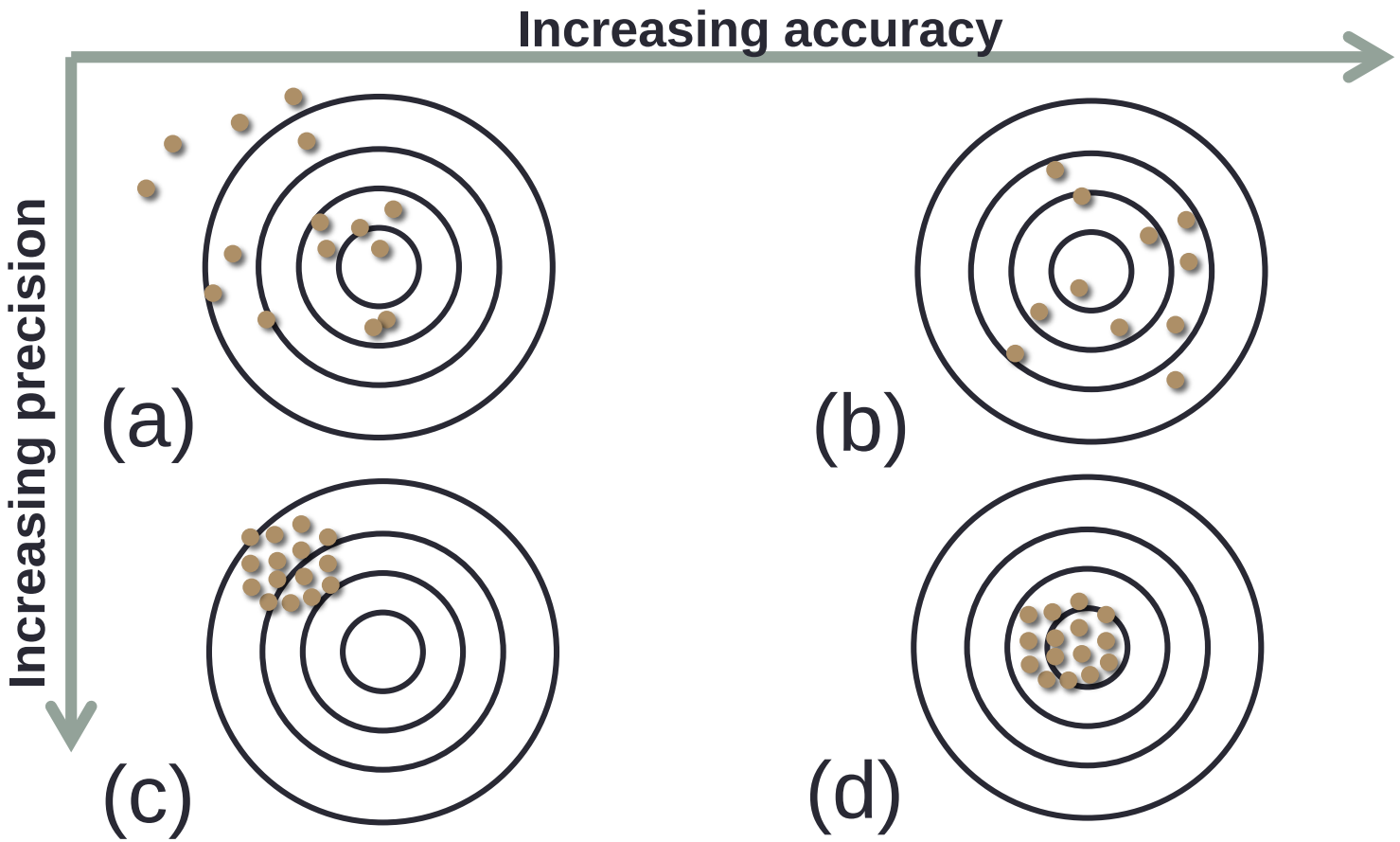
Accuracy and Precision

The errors associated with measurements can be characterized by their accuracy and precision.

- Accuracy refers to how close the value is to the **true value**.
- Precision refers to how close the measured values are to **each other**.

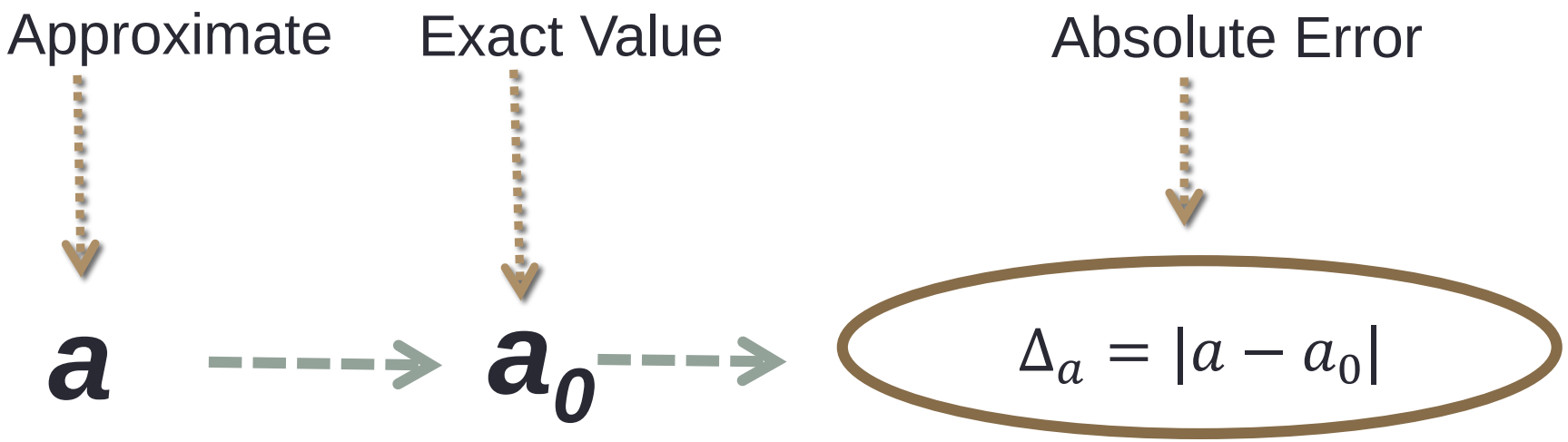


An example on a Target for Accuracy and Precision



Error Representation

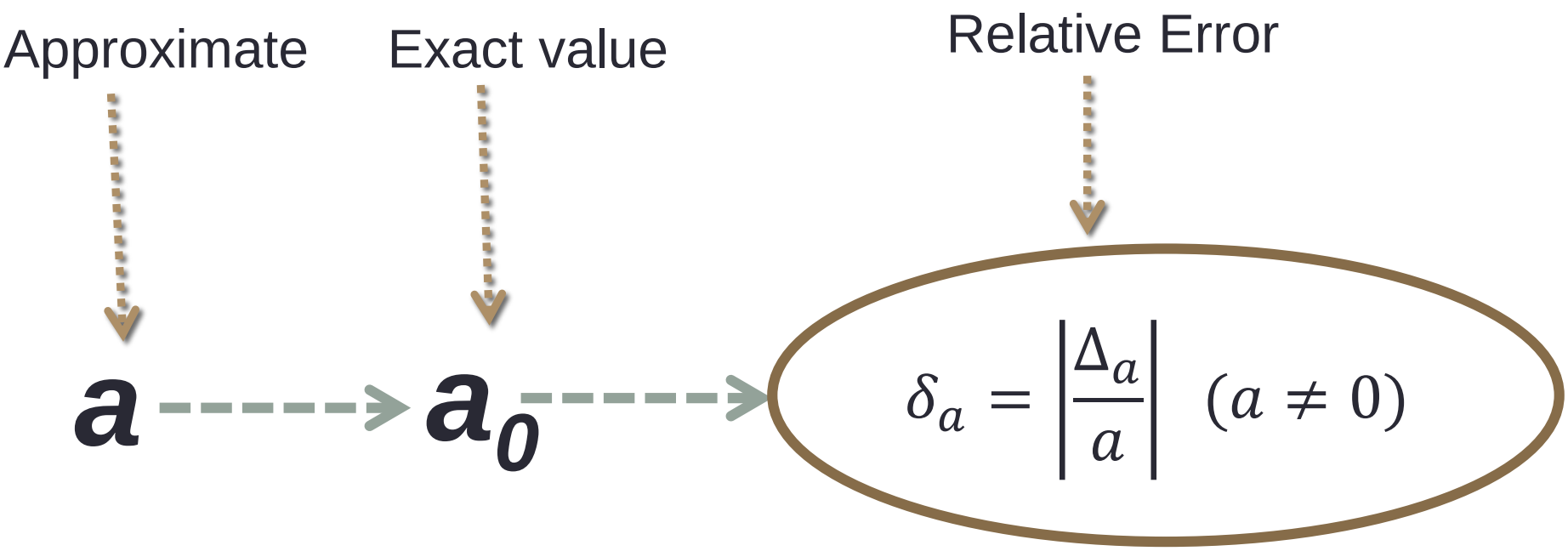
Absolute Error : The discrepancy between an exact value and some approximation to it.



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

Error Representation

Relative Error: The ratio of the absolute error of the number a to the exact number a_0.



Accuracy of approximate numbers

In many applied sciences and technology, the accuracy of approximate numbers is usually characterized by their relative error.

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

Example:

The length and width of a room *measured* *accurate* to $1cm$ are:

$$a = 5.43\,m,\ \ b = 3.82\,m.$$

Estimate the error in determining the area of the room

$$S = a \times b = 20.7426\,m^2.$$

Solution:

Base on the accuracy of measurement, the exact values for length and wide are bounded as follows:

$$-0.01 \le a_0 - 5.43 \le 0.01$$
$$-0.01 \le b_0 - 3.82 \le 0.01$$

$\Delta_a = 0.01m$
$\Delta_b = 0.01m$

So the extreme possible values of area:

$$(a + 0.01)(b + 0.01) = 20.8352\,m\text{^}2$$

$$(a - 0.01)(b - 0.01) = 20.6502\,m^2$$

- Upper bound for Absolute Error: $|S - S_0| \le 0.0926$
- Relative Error

$$\delta_s = \frac{0.0926}{20.7426} = 0.0045 = 0.45\%$$

Adding and Subtracting Approximate Numbers

The absolute error of an algebraic sum of several approximate numbers is equal to the sum of the absolute errors of the numbers. That means if

$$S = a_1 + a_2 + \cdots + a_n,$$

Then

$$\Delta_S = \Delta_{a_1} + \Delta_{a_2} + \cdots + \Delta_{a_n}$$

If among the terms there is a number, whose absolute error exceeds essentially the absolute errors of the rest of the terms, then the absolute error of the sum is considered to be equal to this greatest error.

Example:

Find the sum of the approximate numbers 0.348; 0.1834; 345.4; 235.2; 11.75; 9.27; 0.0849; 0.0214; 0.000354, assuming all their digits being correct, i.e. assuming that the absolute error of each term does not exceed *half the unit of the junior digit retained*.

Solution:

$$
\begin{array}{r}
345.4 \\
235.2 \\
11.75 \\
9.27 \\
0.35 \\
0.18 \\
0.08 \\
0.02 \\
0.00 \\
\hline
\end{array}
$$

$$Sum = 602.25 \simeq 602.2$$

These have the maximum absolute error $\Delta = 0.05$ $2\Delta = 0.1$.

The rounding-off error: $\Delta_{sum} = 0.1 + 0.05 = 0.15$

Remark

If the terms $a_1, \dots, a_n$ have the same sign, then

$$\delta_{min} \leq \delta_S \leq \delta_{max}$$

where $S = a_1 + \cdots + a_n$.

Proof.

Subtracting Approximate Numbers

The relative error of a difference of two positive numbers is more than the relative errors of these numbers, especially if they are nearly equal.

$$c = a - b$$

Example:

$a = 1.137$ and $b = 1.073$ with the absolute errors $\Delta_a = \Delta_b = 0.011$

Solution:

$$c = 1.137 - 1.073 = 0.064$$

$$\Delta_c = \Delta_a + \Delta_b = 0.022$$

$$\delta_c = \frac{22}{64} = 35\%$$

$$\delta_a = \delta_b = 1\%$$

Multiplying and Dividing Approximate Numbers

The relative error of the expression

$$r = \frac{a_1 a_2 \dots a_m}{b_1 b_2 \dots b_n}$$

is estimated by the quantity

$$\delta_r = \delta_{a_1} + \delta_{a_2} + \cdots + \delta_{a_m} + \delta_{b_1} + \delta_{b_2} + \cdots + \delta_{b_n}$$

Relative error in multiplication

Proposition. If $p = ab$, then $\delta_p = \delta_a + \delta_b$.

Proof.

$$p_{max} = (a + \Delta_a)(b + \Delta_b) = ab + b\Delta_a + a\Delta_b + \Delta_a\Delta_b$$

$$p_{min} = (a - \Delta_a)(b - \Delta_b) = ab - b\Delta_a - a\Delta_b + \Delta_a\Delta_b$$

$$\Delta_p \leq |p_{max} - p|$$

$$\Delta_p \leq |p - p_{min}|$$

$\Delta_a, \Delta_b \approx 0$

So

$$\delta_p \leq \frac{b\Delta a + a\Delta b}{ab} = \delta_a + \delta_b$$

■

Relative error in division

Proposition. If $p = \frac{a}{b}$ then $\delta_p = \delta_a + \delta_b$.

Proof. $p_{max} = \frac{a+\Delta_a}{b-\Delta_b}$ and $p_{min} = \frac{a-\Delta a}{b+\Delta b}$

$$p_{max} - p = \left|\frac{a+\Delta a}{b-\Delta b} - \frac{a}{b}\right| = \left|\frac{ab + b\Delta a - ab + a\Delta b}{b(b-\Delta b)}\right|$$

$$p_{min} - p = \left|\frac{a-\Delta a}{b+\Delta b} - \frac{a}{b}\right| = \left|\frac{ab - b\Delta a - ab - a\Delta b}{b(b+\Delta b)}\right|$$

$$\Delta p \le \frac{b\Delta a + a\Delta b}{b^2} \quad \rightarrow \quad \delta_p = \frac{\Delta p}{\frac{a}{b}} \le \delta_a + \delta_b \ \blacksquare$$

Δb *is* 0 *against* b

Multiplying and Dividing Approximate Numbers

If the number $m + n$ is rather large, then it is more expedient to use the statistic estimate which takes into account a partial compensation of errors of different signs: if all the numbers a_i, b_j $(i = 1, 2, \ldots, m;\ j = 1, 2, \ldots, n)$ have roughly the same relative error δ, then the relative error is equal to

$$\delta_r = \delta\sqrt{3(n+m)} \qquad (m+n) > 10$$

The absolute error can be estimated as:

$$\Delta_r = |r|\delta_r$$

Example:

Compute r and its absolute error in

$$r = \frac{3.2\times357.7\times0.04811}{7.1948\times34.56} = 0.221468424$$

Solution:

$$\delta_a = \frac{0.05}{3.2} = 0.016$$

$$r = \frac{3.2 \times 357 \times 0.0481}{7.19 \times 34.6} = 0.221$$

$$\Delta_r = |r|\delta_r = 0.221 \times 0.016 = 0.0036$$

$$r = 0.22$$

Rounding off the result to correct digits

Errors in Computing the Value of a Function

Functions of one variable :

The absolute error of a differentiable function y= f(x) due to a sufficiently small error of the argument Δx is estimated by the quantity.

$$\Delta_y = |f'(x)|\Delta_x, \qquad f'(x) \neq 0$$

$$\delta_y = \frac{|f'(x)|}{|f(x)|}\Delta_x = |[\ln f(x)]'|\Delta_x$$

Errors in Computing the Values of a Function

Functions of several variables:

The absolute error of a differentiable function $y = f(x_1, \dots, x_n)$ caused by sufficiently small errors $\Delta_{x_1}, \dots, \Delta_{x_n}$ of the arguments $x_1, \dots, x_n$ is estimated by the quantity:

$$\Delta_y = \sum_{i=1}^{n} \left| \frac{\partial f}{\partial x_i} \right| \Delta_{x_i}$$

If the values of the function are positive, then the relative error is estimated by the formula

$$\delta_y = \sum_{i=1}^{n} \frac{1}{f} \frac{|\partial f|}{|\partial x_i|} \Delta_{x_i} = \sum_{i=1}^{n} \left| \frac{\partial \ln f}{\partial x_i} \right| \Delta_{x_i}$$

Example:

Compute the value $z = \ln(10.3 + \sqrt{4.4})$ considering all the digits of the approximate numbers $x = 10.3$ and $y = 4.4$ as correct.

Solution:

$$\delta_y = \frac{0.5}{44} = 1.2\%$$

The number y has the relative error

$$\sqrt{y} = \sqrt{4.4} = 2.10$$

$\sqrt{y}$ has the relative error

$$\delta_y = \frac{\left|\frac{1}{2\sqrt{4.4}}\right|}{\sqrt{4.4}} 0.05 = 0.6\%$$

The absolute error of this root being equal to

$$\Delta_{\sqrt{y}} = 2.10 \times 0.006 = 0.013$$

$$x + \sqrt{y} = 10.3 + 2.10 = 12.4$$

$$\delta_z = \frac{0.05 + 0.013}{12.4} = 0.5\%$$

The relative error

$$z = \ln(10.3 + 2.10) = \ln(12.4) = 2.517$$

The absolute error of the sum

Functions of One Variable

- Logarithmic function $y = \ln x$

$$\Delta_y = \frac{\mathbf{1}}{\boldsymbol{x}} \Delta_x = \boldsymbol{\delta_x}$$

- Trigonometric functions

$$\Delta_{\sin x} = |\cos x| \Delta_x \leq \Delta_x$$

$$\Delta_{\cos x} = |\sin x| \Delta_x \leq \Delta_x$$

$$\Delta_{\tan x} = (1 + \tan^2 x)\Delta_x \geq \Delta_x$$

$$\Delta_{\cot x} = (1 + \cot^2 x)\Delta_x \geq \Delta_x$$

Example:

The diameter of a circle measured to within 1 mm amounts to $d = 0.842m$. Compute the area of the circle and its error.

Solution:

The area of a circle $S=\pi d^2/4$. The number π can be taken with any degree of accuracy, the error of computation of the area is determined by the error of computation of d^2. The relative error of d^2 is

$$\delta_{d^2} = 2\delta_d = 2.\frac{1}{842} = 0.24\%$$

$$\delta_s = \delta(\frac{\pi}{4}) + \delta_d$$

The number π should be taken at least with four correct digits.

$$S = \frac{3.1416}{4} \times 0.8422m^2 = 0.7854 \times 0.7090m^2 = 0.5568m^2$$

$$\Delta_S = S\delta_s = 0.557 \times 0.0024 = 0.0014$$

$$S = 0.557m^2 \qquad \Delta_s = 0.002$$

Determining the Error of Arguments

Functions of one variable, $y = f(x)$

$$\Delta_x = \frac{1}{|f'(x)|}\Delta_y \quad , f'(x) \neq 0$$

Functions of several variable, $y = f(x_1, \dots, x_n)$

$$\Delta_{x_i} = \frac{\Delta_y}{n\left|\frac{\partial f}{\partial x_i}\right|}\Delta_y \, , (i = 1,2, \dots, n)$$

Example:

To what accuracy must an angle x be measured in the first quadrant to get the value of sin(x) with five correct digits?

Solution.

It is known if the angle $\boldsymbol{x} > \mathbf{6}°$ so that $\sin(\boldsymbol{x}) > \mathbf{0.1}$.

Then, it is necessary to determine Δ_x such that the inequality

$\Delta_{sin(x)} < 0.5.10^{-5}$ is fulfilled.

$$\Delta_x = \frac{1}{\cos x}\Delta_{\sin x} > 2 \times 0.5 \times 10^{-5} = 10^{-5}$$

ANY QUESTIONS?