



Numerical Methods

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

©2022

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 Ebrahimpour

Mohammad Pourtaheri

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

Course Introduction

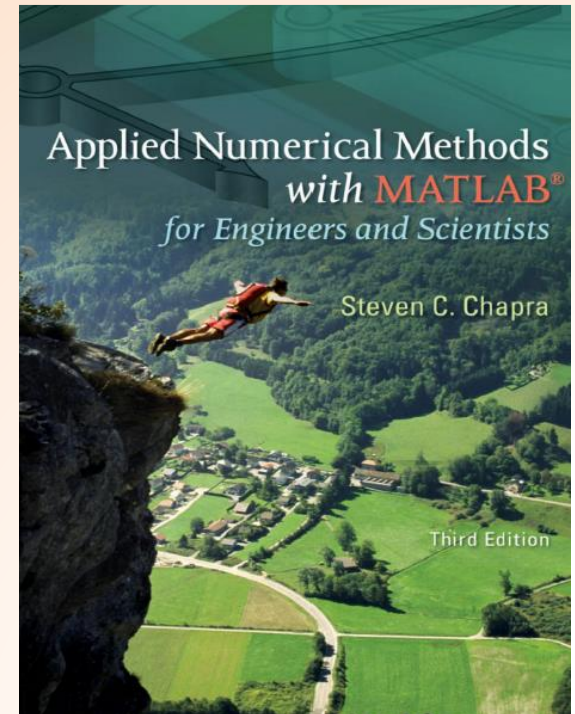
Webpage

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.

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

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

Numerical Methods

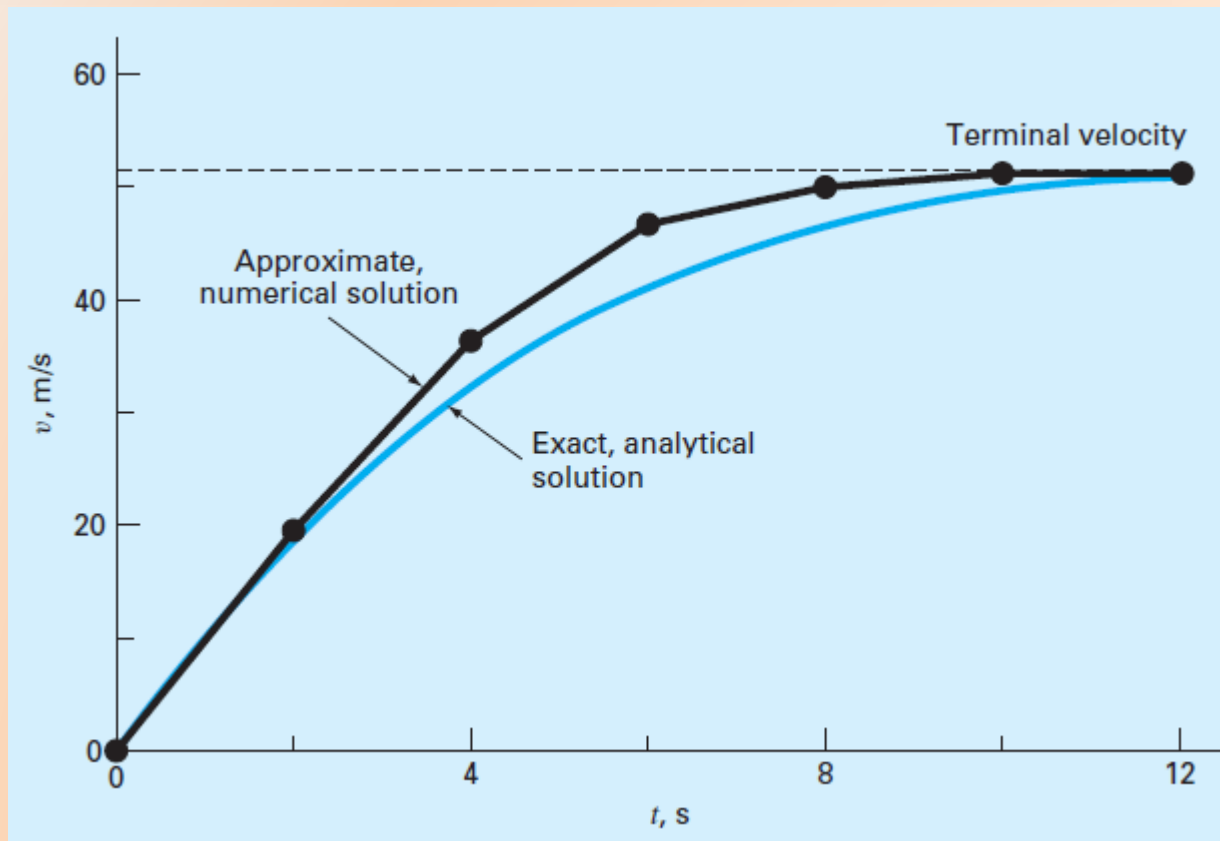
NM Methods:

- Iterative
- Direct

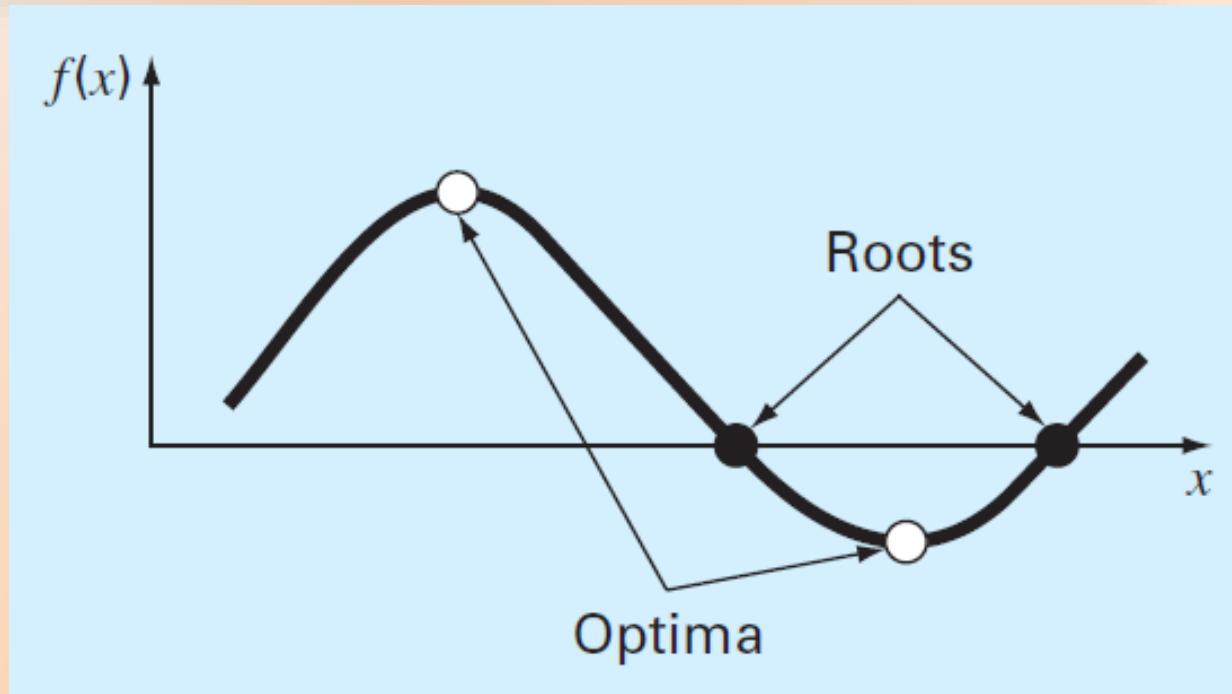
Contents

- 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



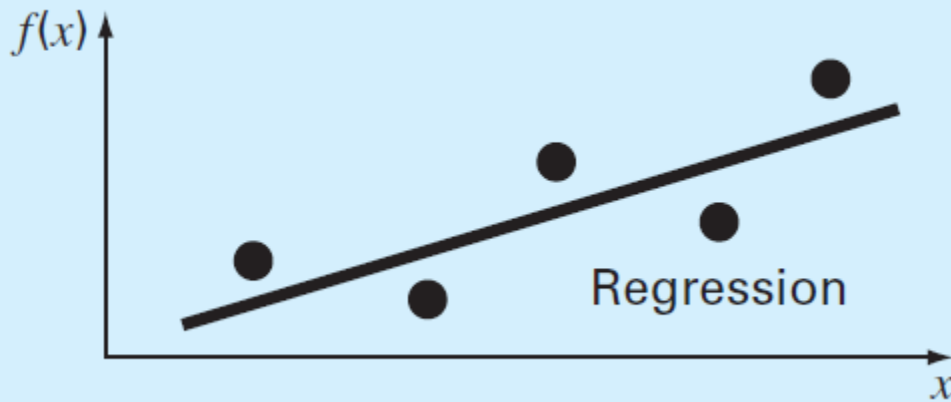
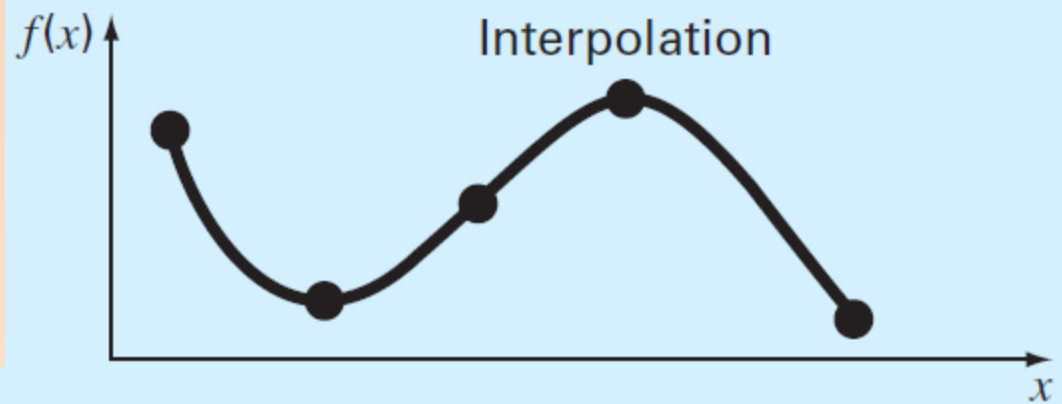
Roots of Nonlinear Equations



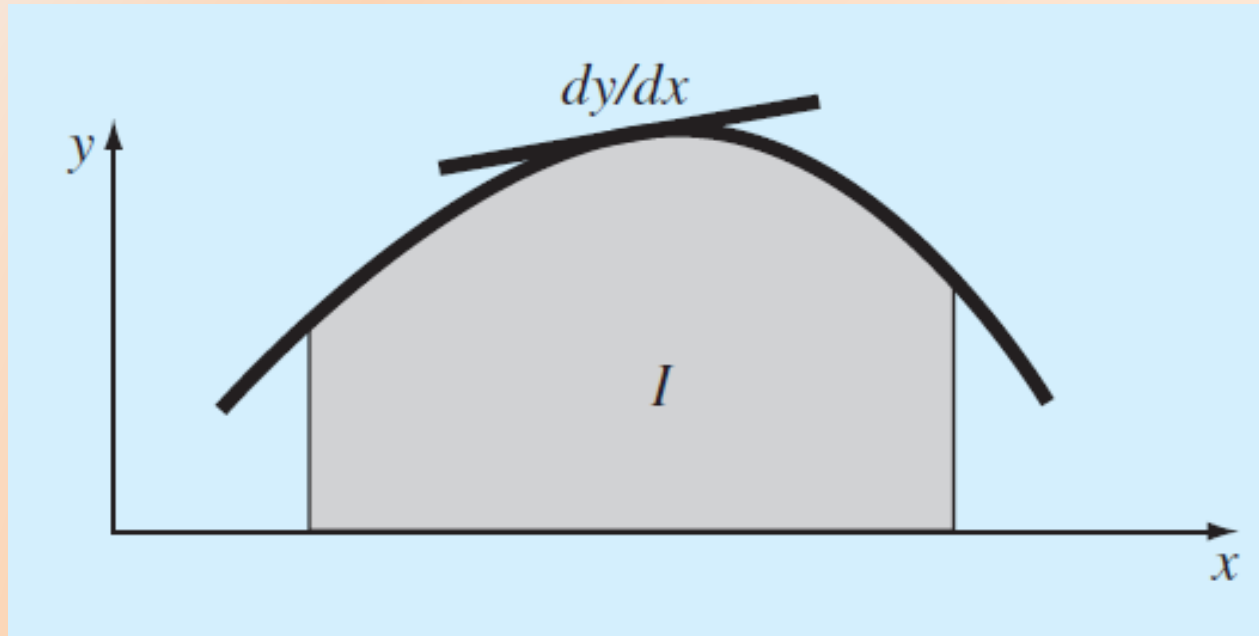
Roots: Solve for x so that $f(x) = 0$

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

Interpolation, Extrapolation, Curve Fitting



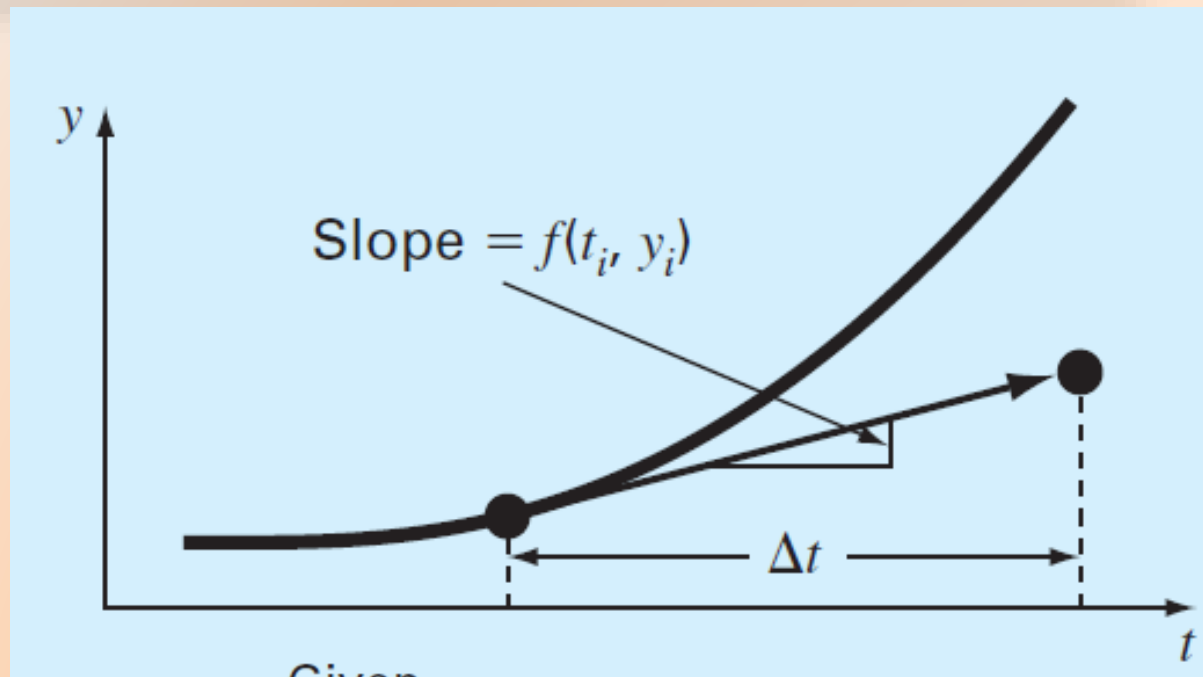
Integration and Differentiation



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

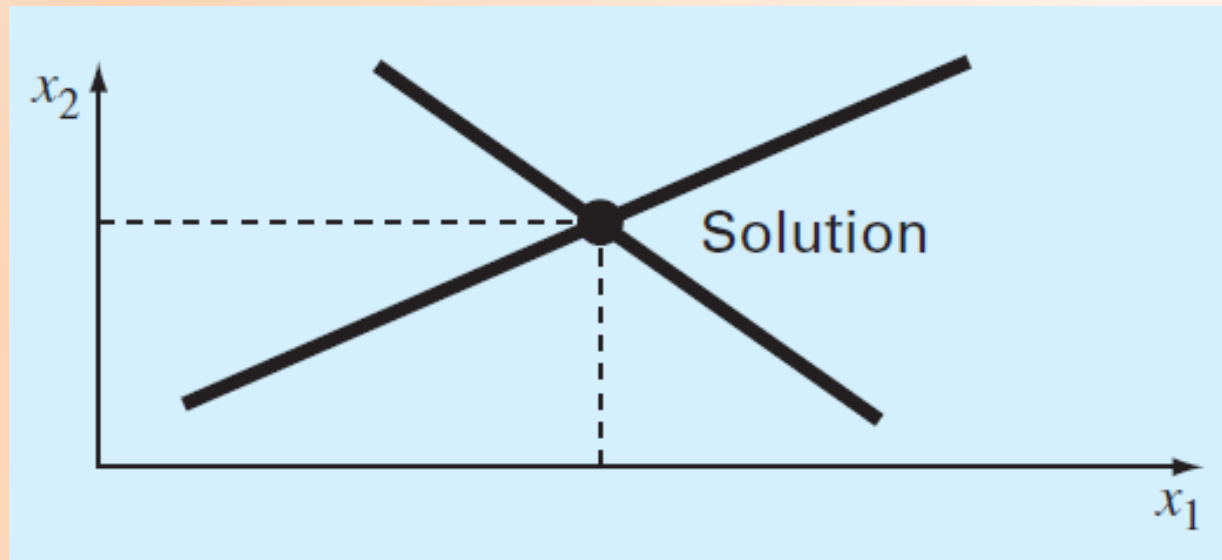
$$y_{i+1} = y_i + f(t_i, y_i)\Delta t$$

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:

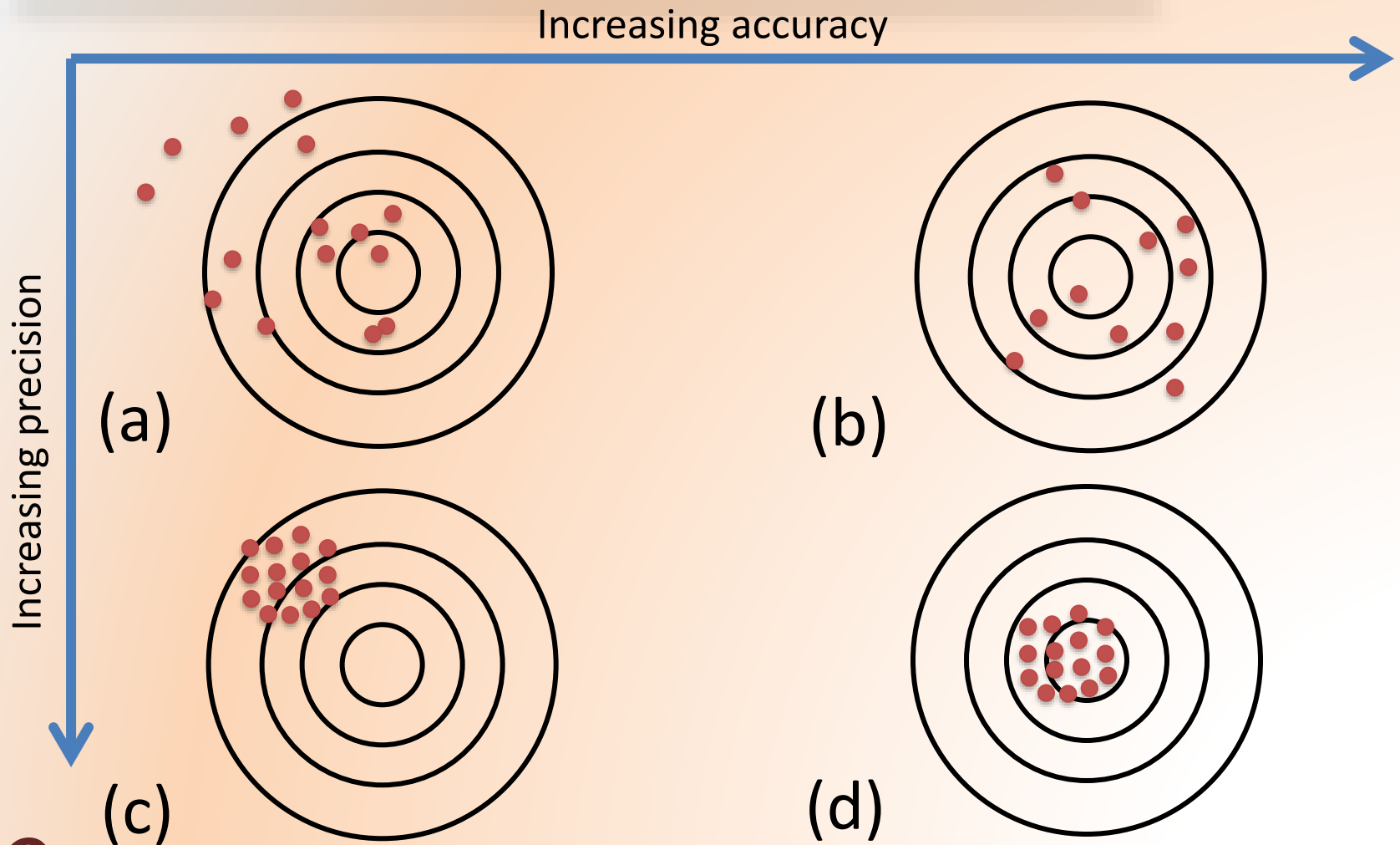
Accuracy refers to how closely a computed or measured value agrees with the true value.

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



1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023, 2024, 2025, 2026, 2027, 2028, 2029, 2030, 2031, 2032, 2033, 2034, 2035, 2036, 2037, 2038, 2039, 2040, 2041, 2042, 2043, 2044, 2045, 2046, 2047, 2048, 2049, 2050, 2051, 2052, 2053, 2054, 2055, 2056, 2057, 2058, 2059, 2060, 2061, 2062, 2063, 2064, 2065, 2066, 2067, 2068, 2069, 2070, 2071, 2072, 2073, 2074, 2075, 2076, 2077, 2078, 2079, 2080, 2081, 2082, 2083, 2084, 2085, 2086, 2087, 2088, 2089, 2090, 2091, 2092, 2093, 2094, 2095, 2096, 2097, 2098, 2099, 2100, 2101, 2102, 2103, 2104, 2105, 2106, 2107, 2108, 2109, 2110, 2111, 2112, 2113, 2114, 2115, 2116, 2117, 2118, 2119, 2120, 2121, 2122, 2123, 2124, 2125, 2126, 2127, 2128, 2129, 2130, 2131, 2132, 2133, 2134, 2135, 2136, 2137, 2138, 2139, 2140, 2141, 2142, 2143, 2144, 2145, 2146, 2147, 2148, 2149, 2150, 2151, 2152, 2153, 2154, 2155, 2156, 2157, 2158, 2159, 2160, 2161, 2162, 2163, 2164, 2165, 2166, 2167, 2168, 2169, 2170, 2171, 2172, 2173, 2174, 2175, 2176, 2177, 2178, 2179, 2180, 2181, 2182, 2183, 2184, 2185, 2186, 2187, 2188, 2189, 2190, 2191, 2192, 2193, 2194, 2195, 2196, 2197, 2198, 2199, 2200, 2201, 2202, 2203, 2204, 2205, 2206, 2207, 2208, 2209, 2210, 2211, 2212, 2213, 2214, 2215, 2216, 2217, 2218, 2219, 2220, 2221, 2222, 2223, 2224, 2225, 2226, 2227, 2228, 2229, 2230, 2231, 2232, 2233, 2234, 2235, 2236, 2237, 2238, 2239, 2240, 2241, 2242, 2243, 2244, 2245, 2246, 2247, 2248, 2249, 2250, 2251, 2252, 2253, 2254, 2255, 2256, 2257, 2258, 2259, 2260, 2261, 2262, 2263, 2264, 2265, 2266, 2267, 2268, 2269, 2270, 2271, 2272, 2273, 2274, 2275, 2276, 2277, 2278, 2279, 2280, 2281, 2282, 2283, 2284, 2285, 2286, 2287, 2288, 2289, 2290, 2291, 2292, 2293, 2294, 2295, 2296, 2297, 2298, 2299, 2300, 2301, 2302, 2303, 2304, 2305, 2306, 2307, 2308, 2309, 2310, 2311, 2312, 2313, 2314, 2315, 2316, 2317, 2318, 2319, 2320, 2321, 2322, 2323, 2324, 2325, 2326, 2327, 2328, 2329, 2330, 2331, 2332, 2333, 2334, 2335, 2336, 2337, 2338, 2339, 2340, 2341, 2342, 2343, 2344, 2345, 2346, 2347, 2348, 2349, 2350, 2351, 2352, 2353, 2354, 2355, 2356, 2357, 2358, 2359, 2360, 2361, 2362, 2363, 2364, 2365, 2366, 2367, 2368, 2369, 2370, 2371, 2372, 2373, 2374, 2375, 2376, 2377, 2378, 2379, 2380, 2381, 2382, 2383, 2384, 2385, 2386, 2387, 2388, 2389, 2390, 2391, 2392, 2393, 2394, 2395, 2396, 2397, 2398, 2399, 2400, 2401, 2402, 2403, 2404, 2405, 2406, 2407, 2408, 2409, 2410, 2411, 2412, 2413, 2414, 2415, 2416, 2417, 2418, 2419, 2420, 2421, 2422, 2423, 2424, 2425, 2426, 2427, 2428, 2429, 2430, 2431, 2432, 2433, 2434, 2435, 2436, 2437, 2438, 2439, 2440, 2441, 2442, 2443, 2444, 2445, 2446, 2447, 2448, 2449, 2450, 2451, 2452, 2453, 2454, 2455, 2456, 2457, 2458, 2459, 2460, 2461, 2462, 2463, 2464, 2465, 2466, 2467, 2468, 2469, 2470, 2471, 2472, 2473, 2474, 2475, 2476, 2477, 2478, 2479, 2480, 2481, 2482, 2483, 2484, 2485, 2486, 2487, 2488, 2489, 2490, 2491, 2492, 2493, 2494, 2495, 2496, 2497, 2498, 2499, 2500, 2501, 2502, 2503, 2504, 2505, 2506, 2507, 2508, 2509, 2510, 2511, 2512, 2513, 2514, 2515, 2516, 2517, 2518, 2519, 2520, 2521, 2522, 2523, 2524, 2525, 2526, 2527, 2528, 2529, 2530, 2531, 2532, 2533, 2534, 2535, 2536, 2537, 2538, 2539, 2540, 2541, 2542, 2543, 2544, 2545, 2546, 2547, 2548, 2549, 2550, 2551, 2552, 2553, 2554, 2555, 2556, 2557, 2558, 2559, 2560, 2561, 2562, 2563, 2564, 2565, 2566, 2567, 2568, 2569, 2570, 2571, 2572, 2573, 2574, 2575, 2576, 2577, 2578, 2579, 2580, 2581, 2582, 2583, 2584, 2585, 2586, 2587, 2588, 2589, 2590, 2591, 2592, 2593, 2594, 2595, 2596, 2597, 2598, 2599, 2600, 2601, 2602, 2603, 2604, 2605, 2606, 2607, 2608, 2609, 2610, 2611, 2612, 2613, 2614, 2615, 2616, 2617, 2618, 2619, 2620, 2621, 2622, 2623, 2624, 2625, 2626, 2627, 2628, 2629, 2630, 2631, 2632, 2633, 2634, 2635, 2636, 2637, 2638, 2639, 2640, 2641, 2642, 2643, 2644, 2645, 2646, 2647, 2648, 2649, 2650, 2651, 2652, 2653, 2654, 2655, 2656, 2657, 2658, 2659, 2660, 2661, 2662, 2663, 2664, 2665, 2666, 2667, 2668, 2669, 2670, 2671, 2672, 2673, 2674, 2675, 2676, 2677, 2678, 26

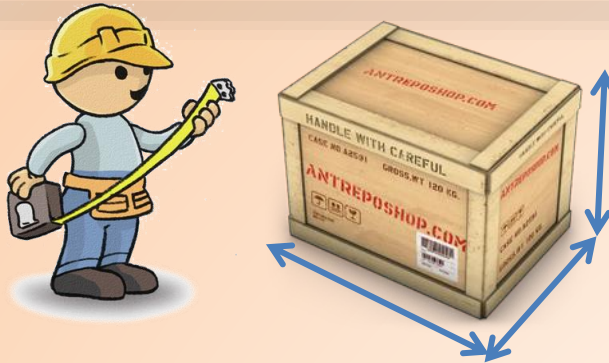
Examples of Errors

$$\sqrt{5} \stackrel{!}{=} 2.23.$$

Examples of Errors

$$\sqrt{5} \stackrel{!}{=} 2.2360679\dots$$

Examples of Errors



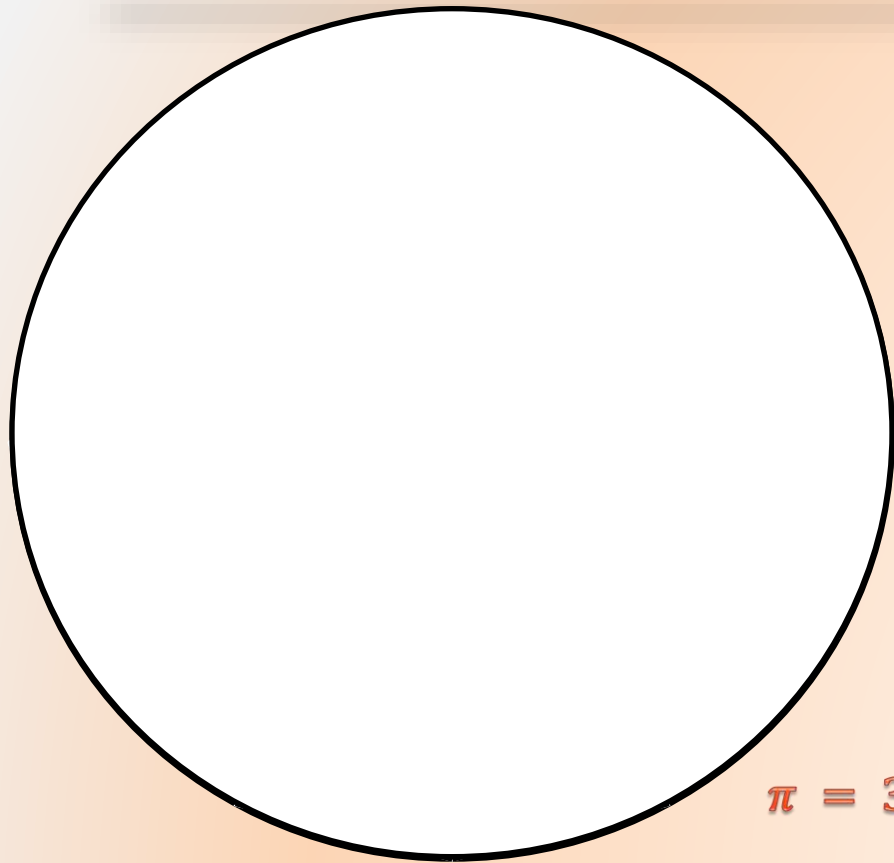
$$\pi = 3.14159265\dots$$

$$e = 2.71828182\dots$$

$$\frac{1}{3} = 0.3333333\dots$$

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.

$$\pi = 3.1415926535897932384626433832795028 \dots$$

Examples of Errors



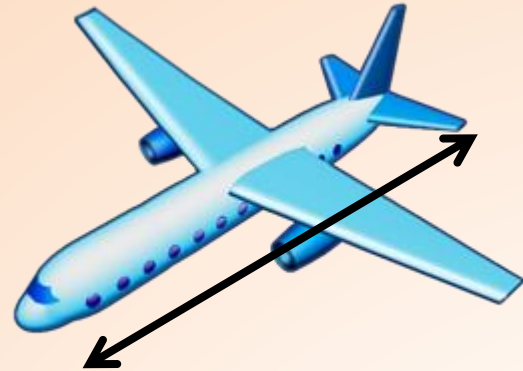
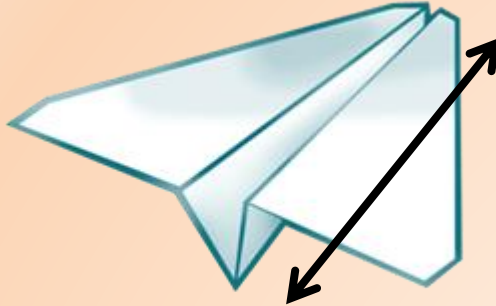
In Euclidean plane geometry, π is defined as the ratio of a circle's circumference (C) to its diameter (d).

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

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.

$$\pi = 3.1415926535897932384626433832795028 \dots$$

Examples of Errors



Actual Size	17.2 <i>cm</i>	82.49 <i>m</i>
Measured Size	18.2 <i>cm</i>	82.5 <i>m</i>

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

Approximate

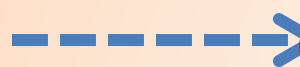
Absolute Error



x



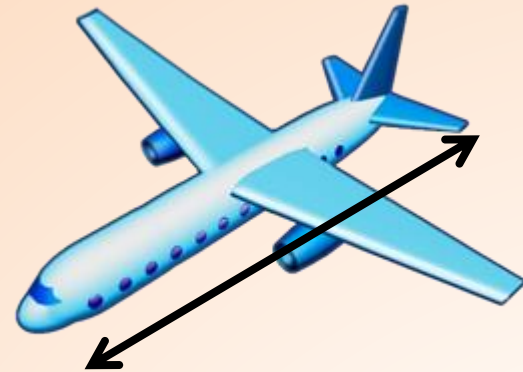
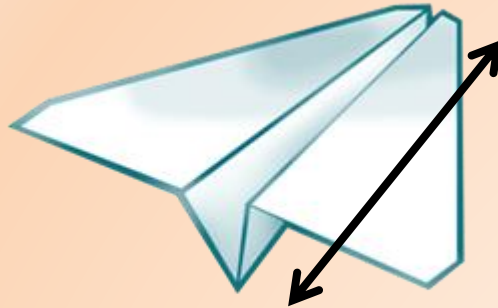
\bar{x}



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

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

An Example



Actual Size	17.2 <i>cm</i>	82.49 <i>m</i>
Measured Size	18.2 <i>cm</i>	82.5 <i>m</i>
Absolute Error	1 <i>cm</i>	1 <i>cm</i>

Error Representation

Absolute

Relative

Example

True Value

Approximate

Relative Error



x



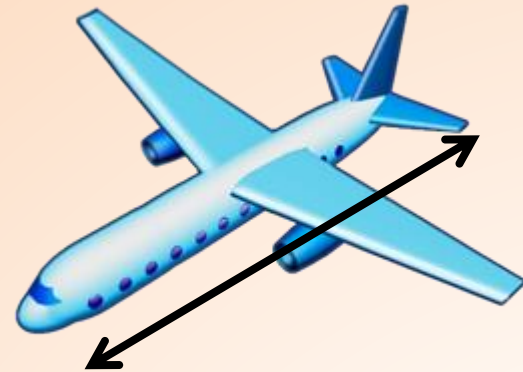
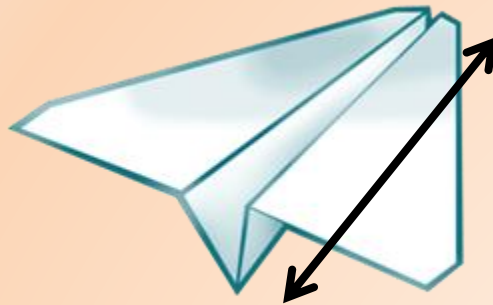
\bar{x}



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

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

An Example



Actual Size	17.2 <i>cm</i>	82.49 <i>m</i>
Measured Size	18.2 <i>cm</i>	82.5 <i>m</i>
Absolute Error	1 <i>cm</i>	1 <i>cm</i>
Relative Error	0.05	0.00012

Error Representation

Absolute

Relative

Example

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

Example: $x = \sqrt{2}$

$$\bar{x} = 1.41$$

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

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

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

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

σ is the sign (+ or -).

β is the base, e is the exponent.

binary : $\beta = 2$

decimal : $\beta = 10$

m is mantissa (*significant*):

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

binary: $1 \leq m < 2$

decimal: $1 \leq m < 10$

Normalized Representation

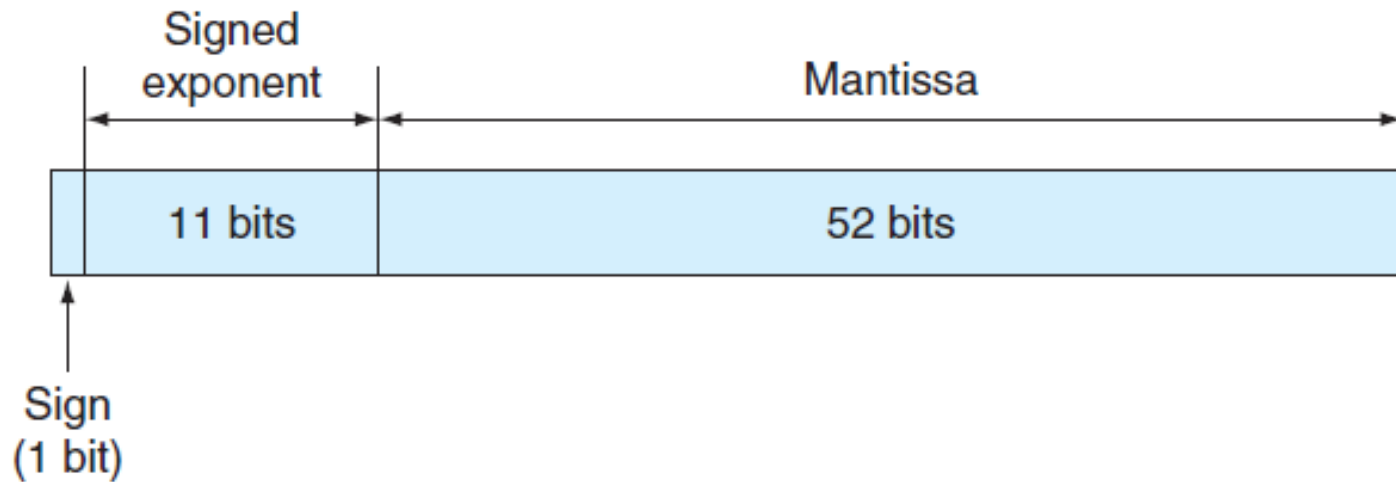
Example: $z = 0.005678$

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

$$x = 3.5 \text{ (2 significant digits)} \rightarrow 3.45 \leq x < 3.55$$

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:

$$1.1011 \times 2^{-3} \quad \text{or} \quad (1 + 0.1011) \times 2^{-3}$$

only have to store the four fractional bits instead 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

$$\text{Largest value} = +1.1111 \dots 1111 \times 2^{+1023}$$

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

$$\text{Smallest value} = +1.0000 \dots 0000 \times 2^{-1022}$$

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



$23.465 < \text{length} < 23.475$

Types of Errors

Inherent

Round Off

Truncation

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

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

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 \leq \frac{1}{2} \times 10^{-t}$

Rounded off to t digits.

Types of Errors

Inherent

Round Off

Truncation

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots + (-1)^n \frac{x^{2n+1}}{(2n+1)!} \pm \cdots$$

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!} + E_n(x)$$

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

Types of Errors

Inherent Round Off Truncation

مثال ۱۴. مقدار تقریبی تابع $\sin x$ را به ازای $x = \frac{\pi}{5}$ و با خطای کمتر از 10^{-3} حساب کنید.

حل : داریم

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots + (-1)^n \frac{x^{2n+1}}{(2n+1)!} \pm \cdots$$

$$|E_n(x)| = \frac{x^{2n+1}}{(2n+1)!} \text{ در اینجا قرار می دهیم}$$

$$x = \frac{\pi}{5} = \pi \frac{1}{5} = 3,1416 \times 0,1429 = 0,4489$$

Types of Errors

Inherent Round Off Truncation

بنابراین بایستی n را طوری تعیین کنیم که

$$\frac{(0,4489)^{n+1}}{(n+1)!} \leq \frac{1}{2} \times 10^{-2} = 5 \times 10^{-3}$$

برای $n \geq 2$ نامساوی فوق برقرار می باشد، در نتیجه

$$\begin{aligned} \sin \frac{\pi}{4} &\simeq 0,4489 - \frac{(0,4489)^2}{2!} + \frac{(0,4489)^3}{3!} - \frac{(0,4489)^4}{4!} + \frac{(0,4489)^5}{5!} \\ &= 0,4489 - 0,101 + 0,0002 \\ &= 0,4340 \quad (4D) \end{aligned}$$

$$\sin \frac{\pi}{4} \simeq 0,434(3D)$$

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

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

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

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

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

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

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

Error Propagation

Absolute Error

Relative Error

Subtraction (-):

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

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

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

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

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

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

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

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 $\sqrt{17}$ و $\sqrt{5}$ را تا سه رقم اعشار گرد کنیم، مطلوبست محاسبه $\sqrt{17} \pm \sqrt{5}$ و محاسبه حداکثر خطای حاصل جمع و تفاضل.

حل: داریم: $\sqrt{17} = 4,123 + e_1$, $\sqrt{5} = 2,236 + e_2$
منظور از e_1 و e_2 خطای مرکب شده در نمایش $\sqrt{17}$ و $\sqrt{5}$ می باشد. چون اعداد تا سه رقم اعشار گرد شده اند، پس

$$e_1 \leq \frac{1}{2} \times 10^{-2}, \quad e_2 \leq \frac{1}{2} \times 10^{-2}$$

$$\sqrt{17} + \sqrt{5} = (4,123 + 2,236) + e_2 = 6,359 + e_2 \quad \text{داریم:}$$

و چون $e_2 \leq e_1 + e_2$ لذا $e_2 \leq 10^{-2}$ در نتیجه

$$6,359 - 10^{-2} \leq \sqrt{17} + \sqrt{5} \leq 6,359 + 10^{-2}$$

همچنین $\sqrt{17} - \sqrt{5} = 1,887 + e_2$ که در اینجا نیز

$$e_2 \leq e_1 + e_2 \leq 10^{-2}$$

$$1,887 - 10^{-2} \leq \sqrt{17} - \sqrt{5} \leq 1,887 + 10^{-2}$$

بنابراین

Error Propagation

Absolute Error

Relative Error

Addition (+):

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

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

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

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

Error Propagation

Absolute Error

Relative Error

Subtraction (-):

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

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

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

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

Error Propagation

Absolute Error

Relative Error

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

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

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

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

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

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

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

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

Error Propagation

Absolute Error

Relative Error

Division (/):

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

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

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

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

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

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

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

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

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

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

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

Error Propagation

Absolute Error

Relative Error

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

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

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

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 $\pi\sqrt{2}$ را با چهار رقم اعشار محاسبه نموده و حداکثر خطای این حاصل ضرب را نیز به دست آورید.

حل : داریم:

$$\pi = 3,1416 + e_1$$

$$\sqrt{2} = 1,4142 + e_2$$

$$e_1 \leq \frac{1}{2} \times 10^{-4}, \quad e_2 \leq \frac{1}{2} \times 10^{-4}$$

$$\pi\sqrt{2} = (3,1416 \times 1,4142) + e_r$$

Error Propagation

Absolute Error

Relative Error

$$e_r \leq 3,1416 e_r + 1,4142 e_1$$

$$e_r \leq \frac{1}{r} \times 10^{-r} (3,1416 + 1,4142)$$

$$e_r \leq 0,5 \times 10^{-r} (4,5558) = 2,2779 \times 10^{-r}$$

لما

$$\pi\sqrt{2} = 4,4429 + e'_r$$

Error Propagation

Absolute Error

Relative Error

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

$$e'_r \leq \frac{1}{2} \times 10^{-4} + e_r$$

$$e'_r \leq 0,5 \times 10^{-4} + 2,2779 \times 10^{-4} = 2,7779 \times 10^{-4}$$

$$4,4429 - 2,7779 \times 10^{-4} \leq \pi\sqrt{2} \leq 4,4429 + 2,7779 \times 10^{-4}$$

$$4,4426 \leq \pi\sqrt{2} \leq 4,4432$$

Error Propagation

Absolute Error

Relative Error

Multiplication (*):

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

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

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

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

Error Propagation

Absolute Error

Relative Error

Division (/):

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

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

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

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

Error Propagation

Absolute Error

Relative Error

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

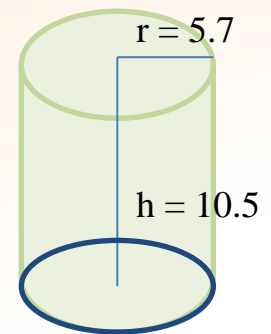
$$\bar{y} \neq 0$$

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

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

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 $\pi = 3.14$)

Solution :

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

π is symmetrically rounded off into two floating digits,

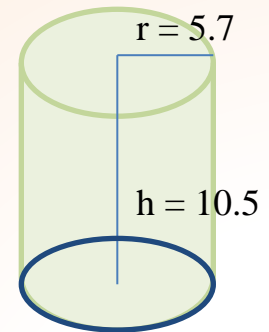
$$\text{thus} \rightarrow e_{\pi} \leq 0.5 \times 10^{-2}$$

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

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

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

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



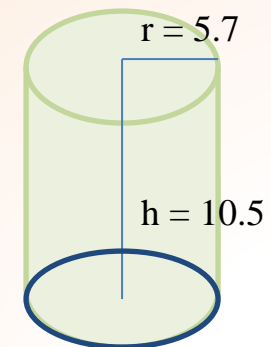
$$e_v = e_{h\pi \times r^2} \leq |h\pi|e_{r^2} + |r^2| \times e_{h\pi} = 0.494933 \times 10^2, \delta_v \leq 0.46204 \times 10^{-1}$$

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 :

	<i>value</i>	<i>max(e)</i>	<i>max(δ)</i>
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
$v = (h\pi) \cdot (r^2)$	1071.1953	49.49338	0.046204

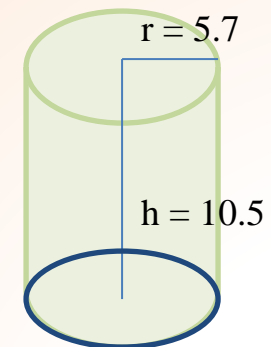


An Example

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 :

	<i>value</i>	<i>max(e)</i>	<i>max(δ)</i>
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
$v = (h\pi) \cdot (r^2)$	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 :

	<i>value</i>	<i>max(e)</i>	<i>max(δ)</i>
<i>h</i>	10.5	$0.1 + 10.5 - 10.5 $	$0.00952 + \frac{ 10.5 - 10.5 }{10.5}$
<i>π</i>	3.14	$0.005 + 3.14 - 3.14 $	$0.00159 + \frac{ 3.14 - 3.14 }{3.14}$
<i>r</i>	5.7	$0.1 + 5.7 - 5.7 $	$0.01754 + \frac{ 5.7 - 5.7 }{5.7}$
<i>hπ</i>	33	$0.3665 + 33 - 32.97 $	$0.011116 + \frac{ 33 - 32.97 }{33}$
$r^2 = r \times r$	32.5	$1.14 + 32.5 - 32.49 $	$0.035087 + \frac{ 32.5 - 32.49 }{32.5}$
$v = (h\pi) \cdot (r^2)$	107	$49.4933 + 107 - 1071.1953 $	$0.04620 + \frac{ 107 - 1071.1953 }{107}$

Formula Error

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

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

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

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

Formula Error

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

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

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

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

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 $\pi = 3.14$)

Solution :

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

$$f = x y z^2$$

$$e_h = e_x \leq 10^{-1}$$

$$e_\pi = e_y \leq 0.5 \times 10^{-2}$$

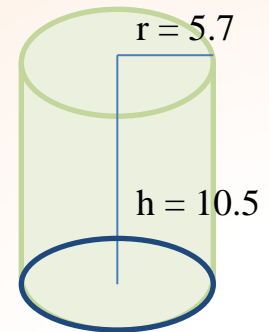
$$e_r = e_z \leq 10^{-1}$$

$$e_f \leq e_x y z^2 + e_y x z^2 + e_z 2 x y z =$$

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

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

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



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?

$$\begin{aligned}1.076x^3 &\rightarrow 35.56 \\0.319x^2 &\rightarrow 3.286 \\0.017x &\rightarrow 0.05455 \\1.107 &\rightarrow 1.107\end{aligned}$$

Solution:

$$35.56 + 3.286 = 38.85$$

$$38.85 - 0.05455 = 38.80$$

$$38.80 + 1.107 = 39.91$$

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:

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

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

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

$$1.107 \rightarrow 1.107$$

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:

$$1.107 - 0.05455 = 1.052$$

$$1.052 + 3.286 = 4.338$$

$$4.338 + 35.56 = 39.90$$

$$1.076x^3 \rightarrow 35.56$$

$$0.319x^2 \rightarrow 3.286$$

$$0.017x \rightarrow 0.05455$$

$$1.107 \rightarrow 1.107$$

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 $\rightarrow 39.894105201004$
- d) What is the difference and why?

Solution:

$$\begin{aligned}1.076x^3 &\rightarrow 35.56 \\0.319x^2 &\rightarrow 3.285 \\0.017x &\rightarrow 0.05455 \\1.107 &\rightarrow 1.107\end{aligned}$$

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

An Example

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

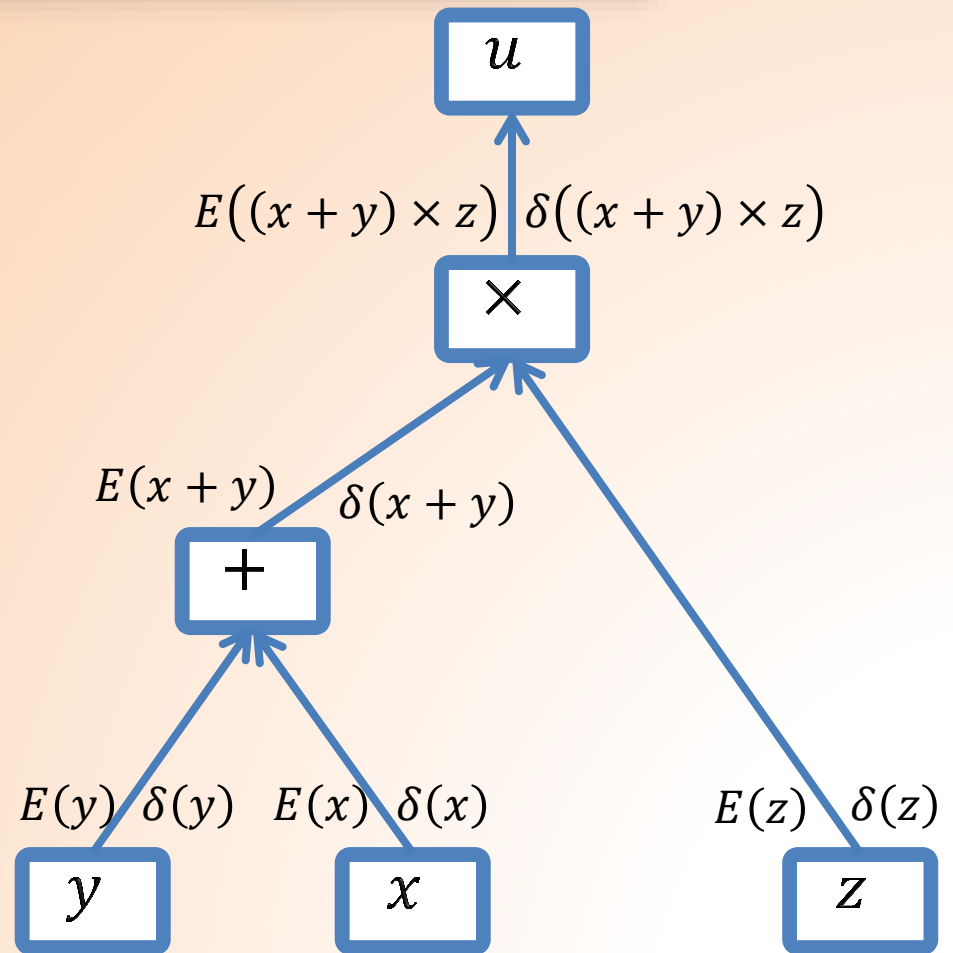
- a) From left to right $\rightarrow 39.91$
- b) From right to left $\rightarrow 39.90$
- c) Compute the exact value
- d) What is the difference and why?

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 $v = \pi r^2 h$

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

An Example

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

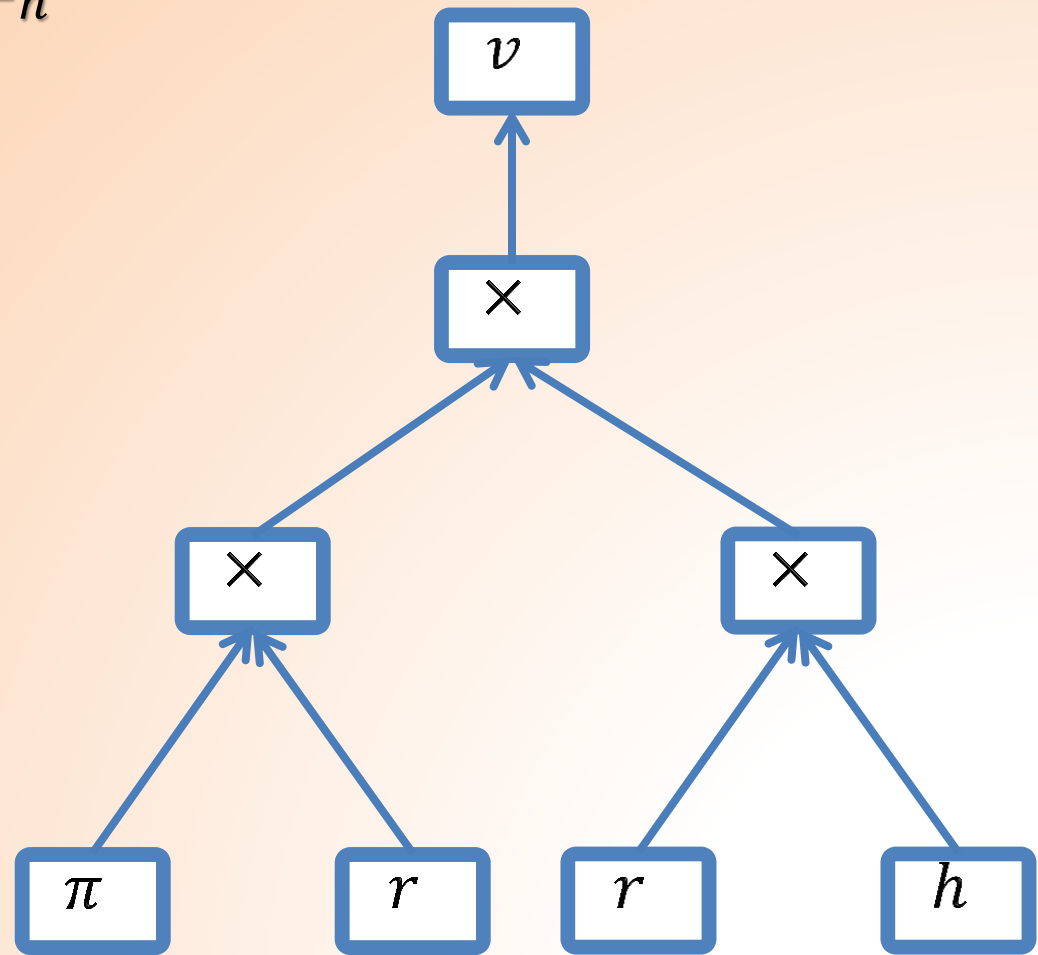
1. From left to right
2. From right to left

Solution :

$$v = \pi r^2 h$$

$$= \pi r r h$$

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



An Example

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

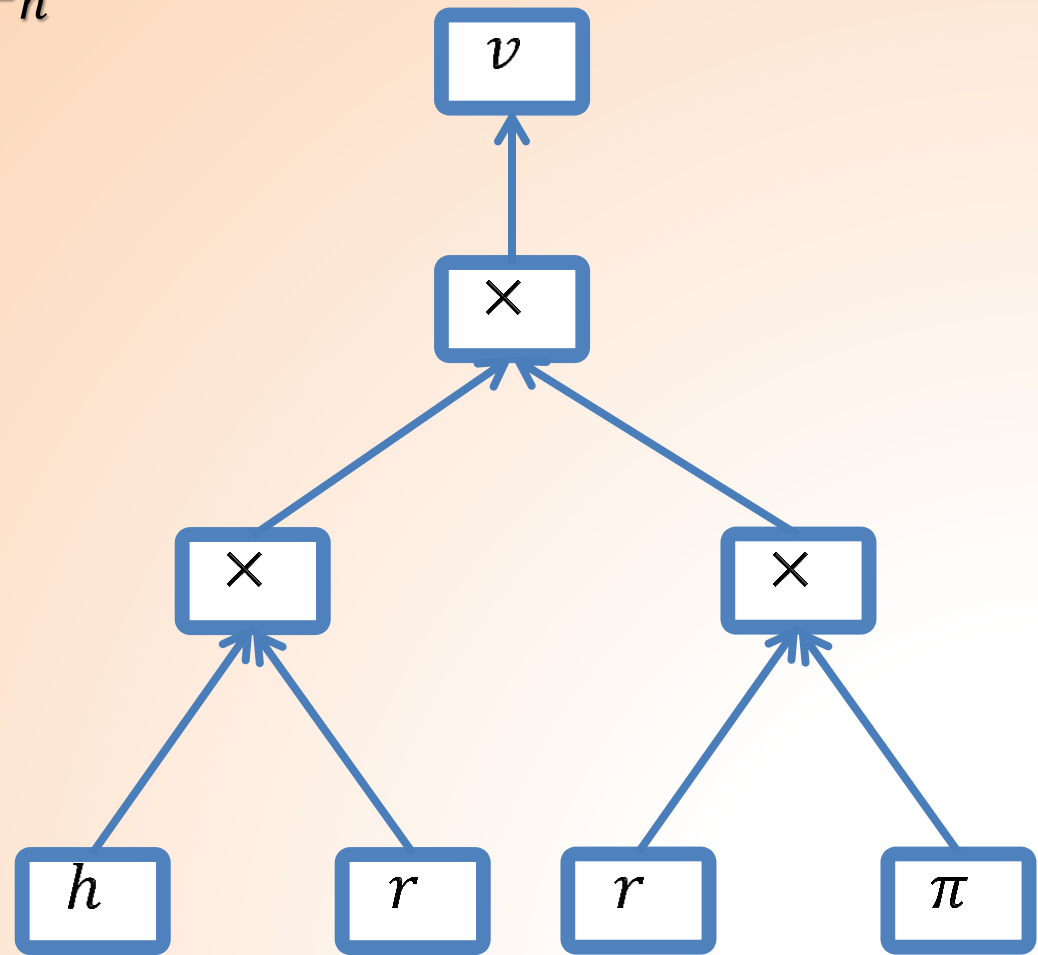
1. From left to right
2. From right to left

Solution :

$$v = \pi r^2 h$$

$$= h r r \pi$$

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



Stability

Algorithm (method) {
Stable : $E_n \approx cE_0$ (linearly)
Unstable : $E_n \approx c^n E_0$ $c \geq 1$ (exponentially)

problem {
Inherent unstable → Example: Wilkinson problem,
Induced unstable

Wilkinson problem: roots of

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

Any questions?

