Introduction to Computer Programming with MATLAB®

Calculation and Programming Errors

Selis Önel, PhD

Today you will learn

- Numbers, Significant figures
- Error analysis
- Absolute error
- Relative error
- Chopping off versus Rounding off

Significant Figures (Digits)

- A concept developed to formally designate the reliability of a numerical value
- Digits of a number that can be used with confidence
- Zeros are **not always** significant figures depending on whether the zeros are known with confidence
- The number of significant digits in an answer to a calculation will depend on the number of significant digits in the given data.
- *Approximate* calculations (order-of-magnitude estimates) always result in answers with only one or two significant digits.

Significant Figures (Digits)

Value	# of significiant figures	Value	# of significiant figures
52	2	0.1845	4
52.0	3	0.01845	4
52.1	3	0.0001845	4
52.1485	6	45300	if $= 4.53 \times 10^4 \rightarrow 3$
52.1485745	9	45300	if $= 4.530 \times 10^4 \rightarrow 4$
52.1485000	9	45300	if $= 4.5300 \times 10^4 \rightarrow 5$

Zeros…

A. placed before other digits are not significant; 0.051 has two significant digits.
B. placed between other digits are always significant; 6002 kg has four significant digits.
C. placed after other digits but behind a decimal point are significant; 2.40 has three significant digits.
D. at the end of a number are significant only if they are behind a decimal point as in C. Otherwise, it is impossible to tell if they are significant.

D. Zeros at the end of a number …

3200→it is not clear if the zeroes are significant or not.

The number of significant digits in 3200 is at least two, but could be three or four.

To avoid uncertainty, use scientific notation to place significant zeroes behind a decimal point:

- 3.200×10^3 has four significant digits
- 3.20×10^3 has three significant digits
- 3.2×10^3 has two significant digits

Significant Digits in Multiplication, Division, Trig. functions, etc

- # of significant digits in an answer = the least number of significant digits in any one of the numbers being multiplied, divided etc.
- Ex: sin(kx), where k = 0.081 m^{-1} (two significant digits) and x = 5.21 m (three significant digits), the answer should have two significant digits.
- Remember: Whole numbers have essentially an unlimited number of significant digits.
- Ex: if a hair dryer uses 1.4 kW of power, then 2 identical hairdryers use 2.8 kW:

1.4 kW {2 sig.dig.} x2 {unlimited sig.dig.} = 2.8 kW {2 sig.dig.}

Significant Digits in Addition and Subtraction

- **number of *decimal places* (not significant digits) in the answer = the least number of decimal places in any of the numbers being added or subtracted.**

Example:

5.67 J (two decimal places)
1.1 J (one decimal place)
0.9378 J (four decimal places)
7.7 J (one decimal place)

When doing multi-step calculations

- **Keep at least one more significant digit in intermediate results than needed in your final answer**.

Example:

If a final answer requires two significant digits, then carry at least three significant digits in calculations.

If you round-off all your intermediate answers to only two digits, you are discarding the information contained in the third digit, and as a result the *second* digit in your final answer might be incorrect.

This phenomenon is known as "**round-off error**")

THINGS NOT TO DO!

- Writing more digits in a final answer than justified by the number of digits in the data.
- Rounding-off, say, to two digits in an intermediate answer, and then writing three digits in the final answer.

Exercises

1) e^{kt} = ?, where k = 0.0286 yr^{-1}, and t = 15 yr

2) ab/c = ?, where a = 256 J, b = 33.56 J, and c = 11.42

3) x + y + z = ?, where x = 48.1, y = 77, and z = 65.789

4) m - n - p = ?, where m = 25.6, n = 21.1, and p = 2.43

Exercises

1) e^{kt} = ?, where k = 0.0286 yr^{-1}, and t = 15 yr.
   [Ans. 0.4290]

2) ab/c = ?, where a = 256 J, b = 33.56 J, and c = 11.42
   [Ans. 9.811×10^4 J^2]

3) x + y + z = ?, where x = 48.1, y = 77, and z = 65.789
   [Ans. 191]

4) m - n - p = ?, where m = 25.6, n = 21.1, and p = 2.43
   [Ans. 2.1]

Machine Numbers

- We do arithmetic using the decimal (base 10) number system
- Computers do arithmetic using the binary (base 2) number system
- Computers convert the numbers we enter in base 10 to base 2, performs base 2 arithmetic and then translates the answer to base 10 before it displays it as a result

Base 10 and Base 2 Numbers

Base 10	Base 2			
	2^3	2^2	2^1	2^0
1	0	0	0	1
2	0	0	1	0
3	0	0	1	1
4	0	1	0	0
5	0	1	0	1
6	0	1	1	0
7	0	1	1	1
8	1	0	0	0
9	1	0	0	1
10	1	0	1	0

Reference: Gilat, Subramaniam: *Numerical Methods with Matlab*

Base 10 and Base 2 Numbers

$$\begin{array}{cccccccccc} 10^4 & 10^3 & 10^2 & 10^1 & 10^0 & & 10^{-1} & 10^{-2} & 10^{-3} & 10^{-4} \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & & \downarrow & \downarrow & \downarrow & \downarrow \\ 6 & 0 & 7 & 2 & 4 & . & 3 & 1 & 2 & 5 \end{array}$$

$$6\times10^4+0\times10^3+7\times10^2+2\times10^1+4\times10^0+3\times10^{-1}+1\times10^{-2}+2\times10^{-3}+5\times10^{-4}=60{,}724.3125$$

$$\begin{array}{cccccccccc} 2^4 & 2^3 & 2^2 & 2^1 & 2^0 & & 2^{-1} & 2^{-2} & 2^{-3} \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & & \downarrow & \downarrow & \downarrow \\ 1 & 0 & 0 & 1 & 1 & . & 1 & 0 & 1 \end{array}$$

$$1\times2^4+0\times2^3+0\times2^2+1\times2^1+1\times2^0+1\times2^{-1}+0\times2^{-2}+1\times2^{-3}$$

$$1\times16+0\times8+0\times4+1\times2+1\times1+1\times0.5+0\times0.25+1\times0.125=19.625$$

Reference: Gilat, Subramaniam: *Numerical Methods with Matlab*

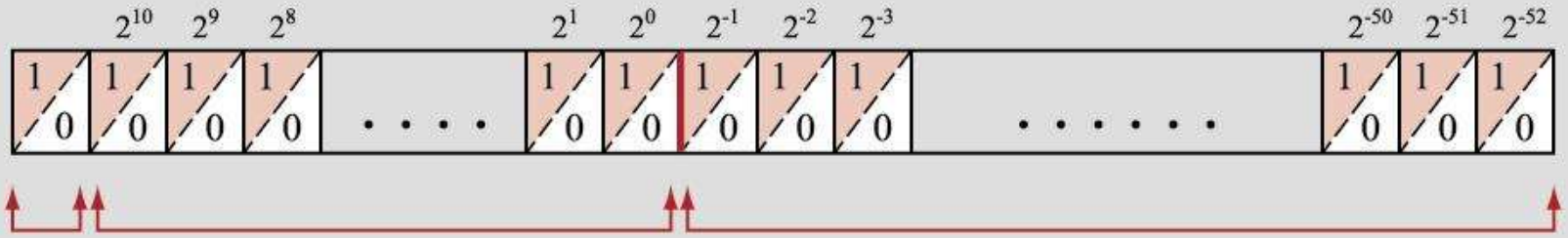
Machine Numbers

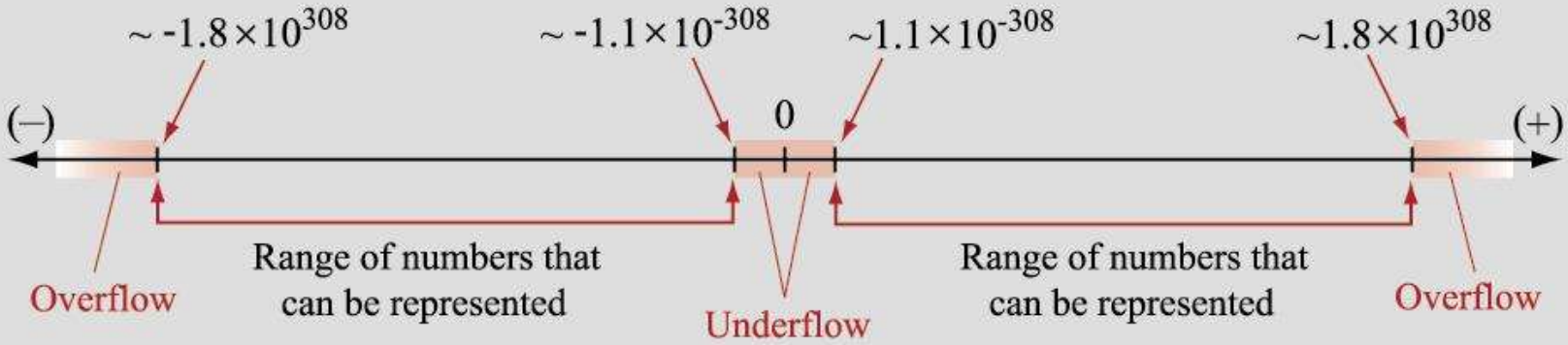
- Computers use normalized floating-point binary representation for real numbers.
- This means mathematical quantity *X* is not actually stored in the computer.
- Instead computer stores: $X \approx \pm q \times 2^n$

q: mantissa, where $\frac{1}{2} \leq q < 1$

n: exponent

Machine Numbers

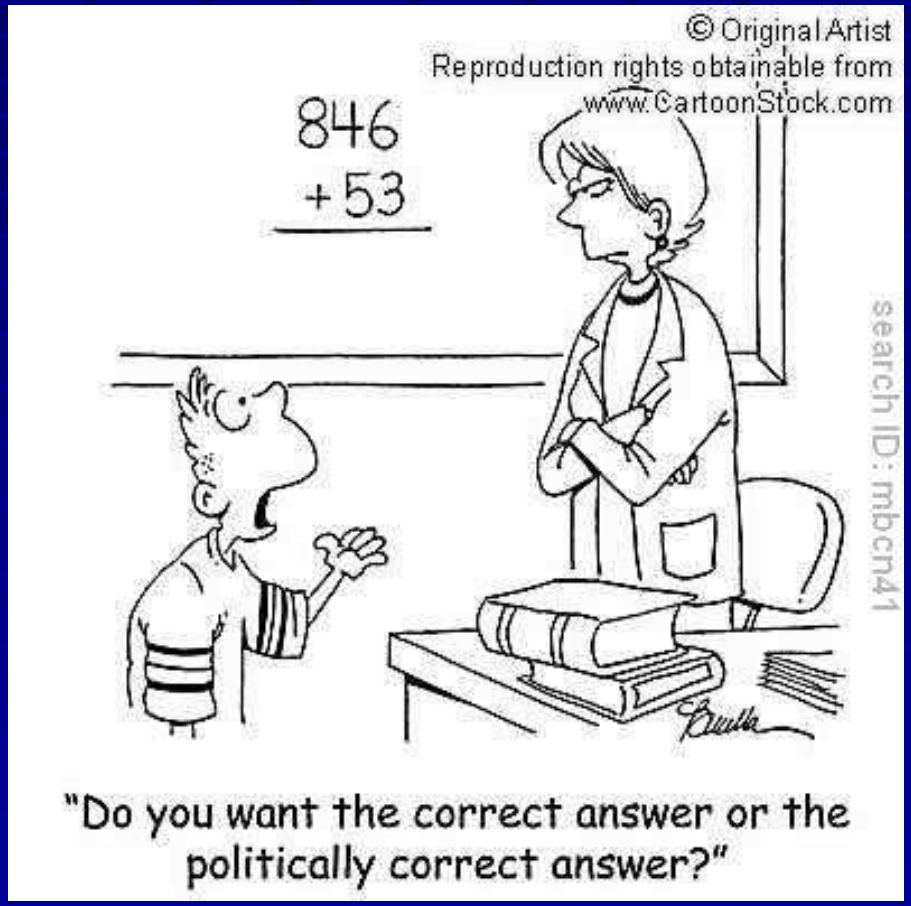




Reference: Gilat, Subramaniam: *Numerical Methods with Matlab*

Quality control in computing

- We are problem solving engineers and our work will be used by clients and sponsors, so our programs must be reliable
- When preparing or executing a program **ERRORS** will occur→ called **BUGS** in computer jargon (remember Admiral Grace Hopper,1945)

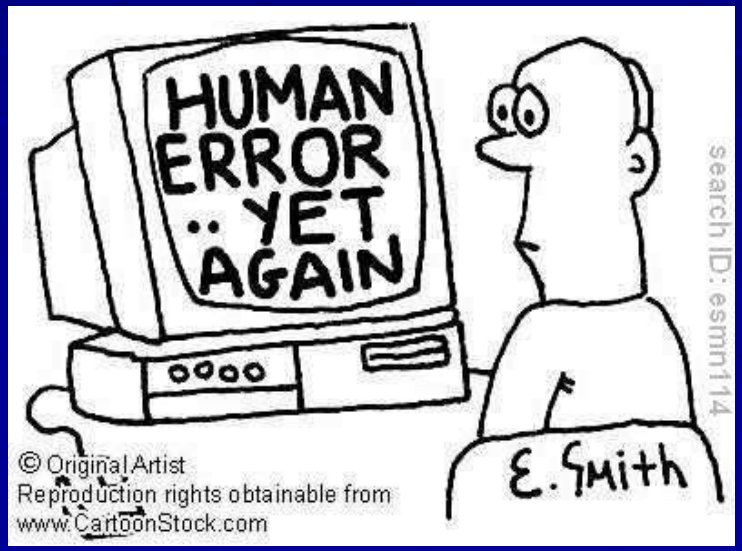


Quality control in computing

The fact that a computer program prints out information is no guarantee that these answers are correct!

When running or developing a program: BUGS

- **Syntax errors:** Violate rules of language such as spelling, number formation, etc. Ex: REED vs READ
- **Link or build errors:** Occur during link phase. Ex: Misspelling the name of an intrinsic function
- **Run-time errors:** Occur during program execution. Ex: Insufficient number of data entries for the number of variables in an input statement
- **Logic errors:** Occur due to faulty program logic. Dangerous because program may work properly but the output will be incorrect!



Debugging and …

- Debug: Correct known errors

If a program runs and prints out reasonable results →

Do I know it is correct?

You need to TEST the program and check the results

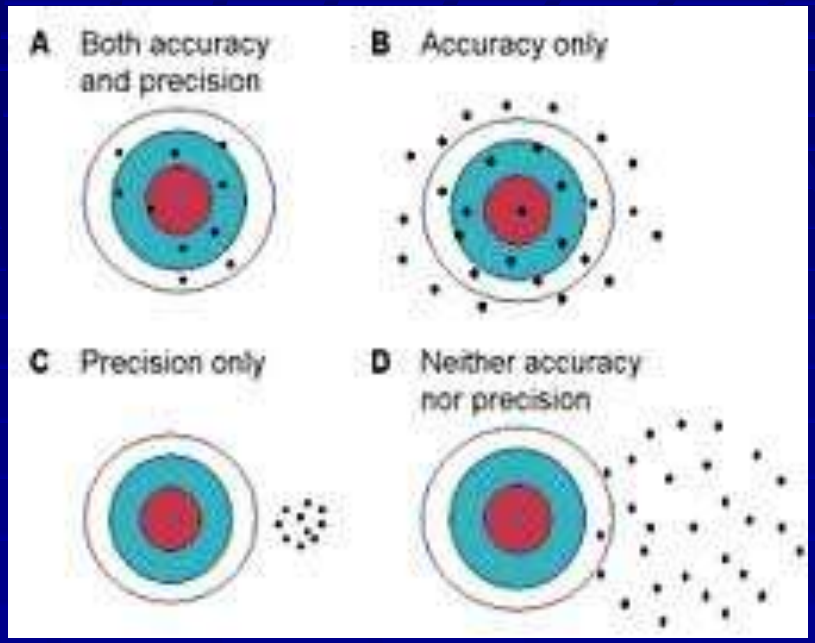
…Testing: To ensure program is correct

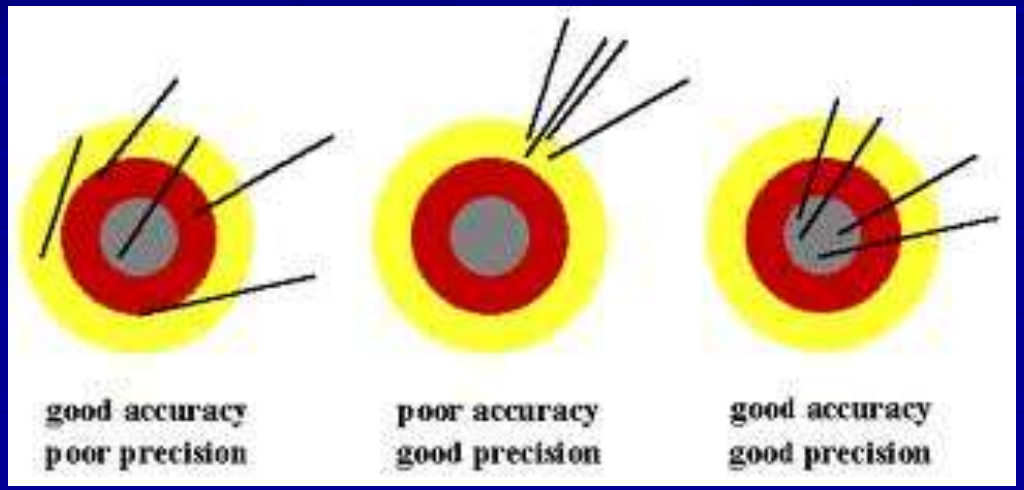
Debug and test modules prior to integrating them into the total program

- **Module tests:** Use sample input data to test each function called in the program
- **Developmental tests:** Perform a test after integrating each module (function) one by one
- **Whole system tests:** After the whole program is assembled run the program using:
  - Typical data
  - Unusual but valid data
  - Incorrect data to check if the program can handle errors

Accuracy and Precision

- Accuracy→ how closely a computed or measured value agrees with the true value
- Precision→ how closely individual computed or measured values agree with each other





Error Analysis

An approximation error can occur because:

- Measurement of data is not precise (due to the instruments), or
- Approximations are used instead of the real data (e.g., 3.14 instead of π)

Absolute error is:

$$\epsilon = |b - a|$$

If **a≠0**, the **relative error** is:

$$\eta = \frac{|b - a|}{|a|},$$

Percent error is:

$$\delta = \frac{|b - a|}{|a|} \times 100\%.$$

Approximation Errors

- Round-off errors
- Truncation errors

Round-off Errors

- Due to use of numbers with limited significant figures to represent exact numbers.

ex: e, π, $\sqrt{7}$ (no fixed number of significant figures)

ex: Computer base-2 representation cannot precisely represent certain exact base-10 numbers.

ex: $1/10=0.0\overline{0011}_{two}$ →

actual number in the computer may undergo chopping or rounding of the last digit

- Computer's representation of real numbers is limited to the **fixed precision of the mantissa**

Round-off Errors

Double-precision

uses 16 digits

```
>> format long e
>> pi
ans =
   3.14159265358979e+000
>> sqrt(7)
ans =
   2.64575131106459e+000
```

Floating-point Representation: Used for fractional quantities in computers.

$m \cdot b^x \rightarrow$ ***m*: mantissa (significand)**

***b*: base of number system**

***x*: exponent**

Mantissa holds only a finite number of significant figures

Truncation Errors

Derivative of velocity of a car

$$\frac{dv}{dt} \cong \frac{\Delta v}{\Delta t} = \frac{v(t_{i+1}) - v(t_i)}{t_{i+1} - t_i}$$

Truncation error

(or *discretization error)* :

- Due to use of approximations to represent exact mathematical procedures
- Introduced when a more complicated mathematical expression is replaced with a more elementary formula
- Due to using finite number of steps in computation
- Present even with infinite-precision arithmetic, because it is caused by truncation of the infinite Taylor series to form the algorithm

Truncation Errors and Taylor Series

Why is Taylor series important in the study of Numerical Methods?

- Provides ways to predict a function value at one point in terms of the function value and its derivatives at another point
- States that any smooth function can be approximated as a polynomial

Reference: S. C. Chapra and R. P. Canale, Numerical Methods for Engineers, 3rd Ed., WCB/McGraw-Hill, 1998, p.79

Truncation Errors and Taylor Series

- A Taylor series of a real (or complex) function *f(x)* is infinitely differentiable in a neighborhood of a real (or complex) number *a*, i.e. it is the power series:

$$f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f^{(3)}(a)}{3!}(x-a)^3 + \cdots,$$

or $$\sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n,$$

- *f*(*x*) is usually equal to its Taylor series evaluated at *x* for all *x* sufficiently close to *a*
- If *a* = 0 → Maclaurin series

Why Use Approximating Functions?

- Replace f(x) (ex: transcendental functions ln x, sin x, erf x, …) with g(x) (ex: a power series) which can handle arithmetic operations
- Ex:

$$e^{x^2} = 1 + x^2 + \frac{x^4}{2!} + \frac{x^6}{3!} + \frac{x^8}{4!} + \ldots + \frac{x^{2n}}{n!} + \ldots$$

Using just 5 terms to simplify gives:

$$e^{x^2} = 1 + x^2 + \frac{x^4}{2!} + \frac{x^6}{3!} + \frac{x^8}{4!}$$

Example Problem (Mathews & Fink p.26)

Given that $\int_0^{1/2} e^{x^2}\,dx = 0.544987104184 = p$ determine the accuracy

of the approximation obtained by replacing the integrand $f(x) = e^{x^2}$

with the truncated Taylor series $P_8(x) = 1 + x^2 + \frac{x^4}{2!} + \frac{x^6}{3!} + \frac{x^8}{4!}$

Term by term integration gives:

$$\int_0^{1/2}\left(1 + x^2 + \frac{x^4}{2!} + \frac{x^6}{3!} + \frac{x^8}{4!}\right)dx = \left(x + \frac{x^3}{3} + \frac{x^5}{5(2!)} + \frac{x^7}{7(3!)} + \frac{x^9}{9(4!)}\right)_{x=0}^{x=\frac{1}{2}}$$

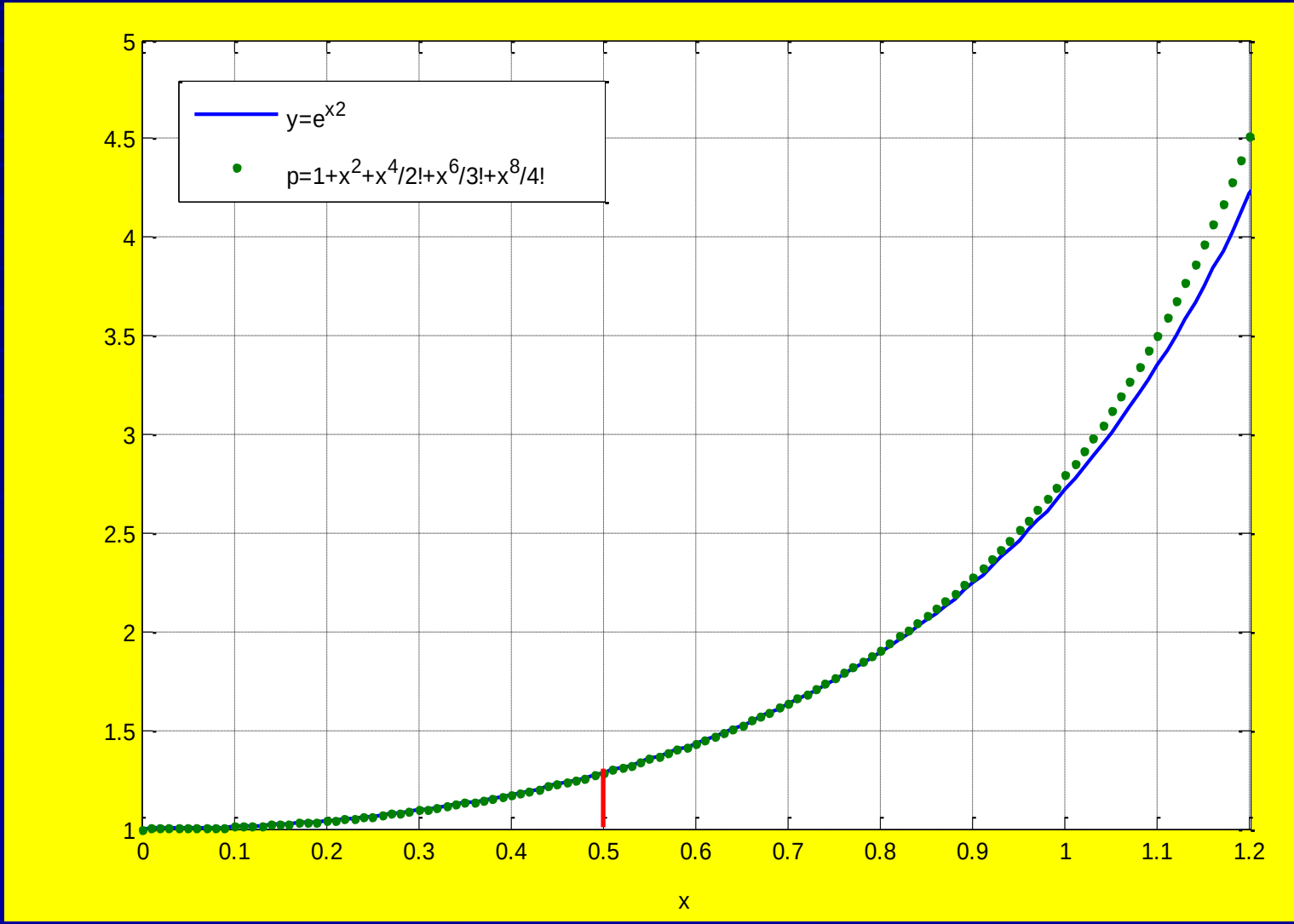
$$= \frac{1}{2} + \frac{1}{24} + \frac{1}{320} + \frac{1}{5376} + \frac{1}{110,592} = \frac{2,109,491}{3,870,720} = 0.544986720817 = p_{new}$$

$$error\% = \frac{|p - p_{new}|}{|p|} = 7.03442 \times 10^{-7} \qquad \frac{10^{-5}}{2} > error\% > \frac{10^{-6}}{2}$$

The approximation p_{new} agrees with the true answer p to five significant figures

Example Problem (Mathews & Fink p.26)

The graphs show the area under the curves between x=0 and 0.5



Chopping off versus Rounding off

- Consider real number p expresses in <u>normalized decimal form</u>:

$$p = \pm 0.d_1d_2d_3....d_kd_{k+1}.....*10^n$$

where $1 \le d_1 \le 9$ and
for $j \ge 1$, $0 \le d_j \le 9$

k: maximum number of decimal digits carried in the floating-point computations of a computer.

I. CHOPPING

$$p_{chopped} = \pm 0.d_1d_2d_3....d_k*10^n$$

chopped floating-point representation of p

Chopping off versus Rounding off

- Consider real number p expresses in <u>normalized decimal form</u>:

$$p = \pm 0.d_1 d_2 d_3 \ldots . d_k d_{k+1} \ldots . * 10^n$$

where $1 \le d_1 \le 9$ and
for $j \ge 1$, $0 \le d_j \le 9$

k: maximum number of decimal digits carried in the floating-point computations of a computer.

I. ROUNDING

$$p_{rounded} = \pm 0.d_1 d_2 d_3 \ldots . r_k * 10^n$$

rounded floating-point representation of p

r_k : last digit is obtained by rounding the number $d_k d_{k+1} d_{k+2} \ldots$ to the nearest integer

Ex: Rounding vs Chopping

Real number $p = \frac{22}{7} = 3.14285714285714285...$

6-digit representation

$P_{chopped} = 0.314285 * 10^1$

$p_{rounded} = 0.314286 * 10^1$

Essentially all computers use some form of rounded floating-point representation method

Loss of significance

The final computed answer may be different depending on your calculation steps

Ex1: Consider functions f(x) and g(x)

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

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

Loss of significance Ex1 cont.d

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

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

Functions f(x) and g(x) are identical:

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

$$= \frac{x\left[(\sqrt{x+1})^2 - (\sqrt{x})^2\right]}{\sqrt{x+1} + \sqrt{x}} \qquad = \frac{x}{\sqrt{x+1} + \sqrt{x}} = g(x)$$

Loss of significance Ex1 cont.d

Calculate f(500) and g(500):

$$f(500) = 500(\sqrt{501} - \sqrt{500})$$

$$= 500(22.3830 - 22.3607) = 500(0.0223) = 11.1500$$

$$g(500) = \frac{500}{\sqrt{501} - \sqrt{500}}$$

$$= \frac{500}{22.3830 + 22.3607} = \frac{500}{44.7437} = 11.1748$$

Matlab results (format long) with 15 decimal digits	Hand calculation with 4 decimal digits
f(500) = 11.174755300746853	f(500) = 11.1500
g(500)= 11.174755300747199	g(500) = 11.1748

Loss of significance Ex2

Compare results of f(0.01) and g(0.01) using 6 digits and rounding

$$f(x) = \frac{e^x - 1 - x}{x^2} \quad and \quad g(x) = \frac{1}{2} + \frac{x}{6} + \frac{x^2}{24}$$

g(x) is the Taylor polynomial of degree n=2 for f(x) expanded about x=0

$$f(0.01) = \frac{e^{0.01} - 1 - 0.01}{(0.01)^2} = \frac{1.010050 - 1 - 0.01}{0.0001} = 0.5$$

$$g(0.01) = \frac{1}{2} + \frac{0.01}{6} + \frac{0.001}{24} = 0.5 + 0.001667 + 0.000004 = 0.501671$$

Matlab results for x = 0.01;

f = (exp(x) -1-x)/x^2 → f = 0.501670841679489

g = 1/2 + x/6 + x^2/24 →g = 0.501670833333333

0.501671 = g(0.01) contains less error and is the same as that obtained by rounding the true answer 0.501670841679489 to six digits.

Loss of Significance in Polynomial Evaluation

For polynomials:

Rearrangement of terms into nested multiplication form may produce better results

Example 3: Loss of significance

Let $P(x) = x^3 - 3x^2 + 3x - 1$ and the nested form is

$Q(x) = ((x - 3)x + 3)x - 1$

$P(x) = Q(x)$

Loss of significance Ex3 contd

Use 3-digit rounding arithmetic to compute P and Q using x = 2.19

$P(2.19) \approx (2.19)^3 - 3(2.19)^2 + 3(2.19) - 1 = 10.5 - 14.4 + 6.57 - 1 = \mathbf{1.67}$

$Q(2.19) \approx ((2.19 - 3)2.19 + 3)2.19 - 1 = \mathbf{1.69}$

Matlab result for P = 1.685158999999999

P = x^3 – 3 * x^2 + 3*x – 1 **~ 1.685159**

Matlab result for Q= 1.685159000000000

Q = ((x - 3)*x + 3)*x – 1 **~ 1.685159**

Errors : 0.015159 and -0.004481

Q has less error!

Errors

- Once an error is generated, it will generally propagate through the calculation.

Ex:

Operation (+) on a calculator (or a computer) is inexact.

It follows that a calculation of the type a+b+c+d+e is even more inexact.

Homework II: Groups of 2 students

Draw a flowchart and write a Matlab program to calculate the total of exam grades and the average

- Repeat operations (iterate) until the input of 101 as a grade, which will end the program
- The program should give an error message with the input of a negative value for grade and should ask for a new grade
- The output of the program should display number of students and the average grade on the command window
- Save your m-file as hw2_lastnames and email to kmu206@gmail.com

Quiz-Error Analysis

1. Find the error *E* and relative error *R* for *x* and x_{app}, determine the number of significant digits in the approximation $x = 2.71828182$ $x_{app} = 2.7182$
2. Consider data p1=1.414 and p2=0.09125. How many significant digits do they have? What would be the proper answers for p1+p2 and p1*p2?