FAEN 301: NUMERICAL METHODS

Lecture 1: Introduction to Numerical Methods

Gifty Buah

August 2019

NUMERICAL METHODS AND ANALYSIS

- Engineers encounter real world problems everyday. Our role is to solve them as efficiently and accurately as possible.
- Unfortunately, real world problems do not usually come in single variable or second-degree equations!
- **Numerical analysis** is the study of algorithms that use numerical approximation (as opposed to general symbolic manipulations) for problems that need mathematical analysis.
- Approximations come with **errors** that must be factored into the mathematical analysis.

Basic Needs in the Numerical Methods:

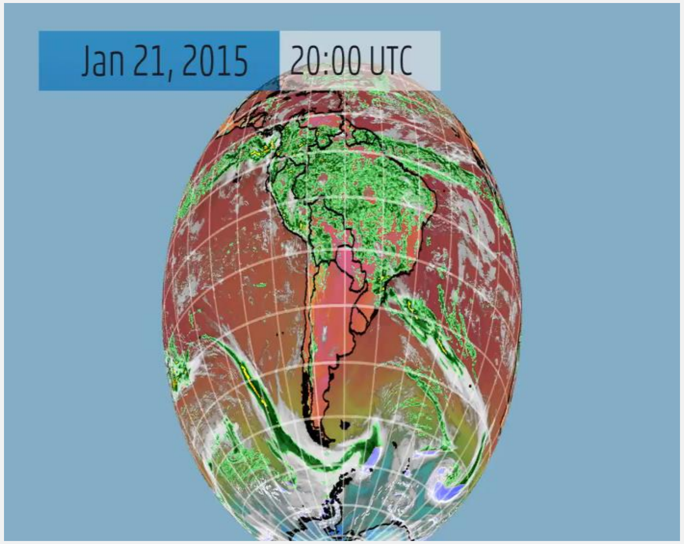
- Practical:

  Can be computed in a reasonable amount of time.

- Accurate:
  - Good approximate to the true value,
  - Information about the approximation error (Bounds, error order,…).

APPLICATIONS OF NUMERICAL ANALYSIS

- Making weather predictions based on many natural factors (wind velocity, solar radiation, humidity, etc)
- Gene simulations
- Crash safety simulations of cars.
- Private investment funds use numerical analysis to predict values of stock.
- Several forms of modelling and simulation involve numerical analysis.



NON-LINEAR EQUATIONS

- Some simple equations can be solved analytically:

$$x^2 + 4x + 3 = 0$$

$$\text{Analytic solution } roots = \frac{-4 \pm \sqrt{4^2 - 4(1)(3)}}{2(1)}$$

$$x = -1 \; and \quad x = -3$$

- Even some slightly complex ones cannot be solved analytically!

$$\left.\begin{array}{r} x^9 - 2x^2 + 5 = 0 \\ x = e^{-x} \end{array}\right\} \text{No analytic solution}$$

SYSTEM OF LINEAR EQUATIONS

- Consider the system below:

$$x_1 + x_2 = 3$$
$$x_1 + 2x_2 = 5$$

- Using Substitution, we solve as:

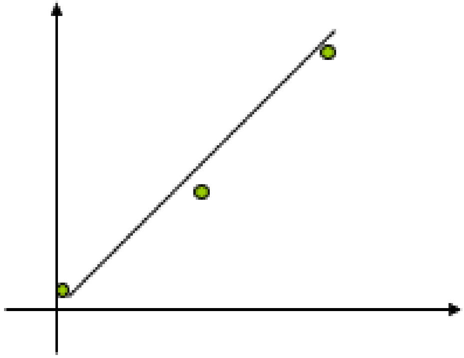
$$x_1 = 3 - x_2\,, \qquad 3 - x_2 + 2x_2 = 5$$
$$\Rightarrow x_2 = 2,\; x_1 = 3 - 2 = 1$$

- What if there are 70 equations with 70 unknowns?

CURVE FITTING

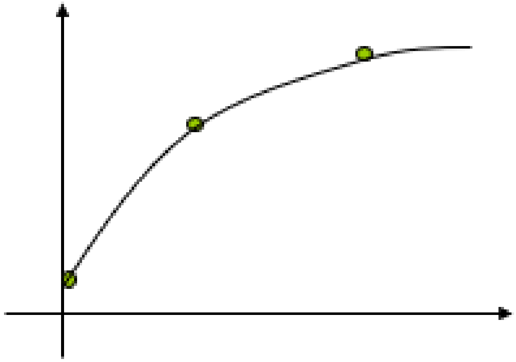
- Given a set of data:

x	0	1	2
y	0.5	10.3	21.3



- Which polynomial P(x) passes through all the points?

x_i	0	1	2
y_i	0.5	10.3	15.3



INTEGRATIONS

- Some functions can be integrated analytically.

$$\int_1^3 x\,dx \qquad = \frac{1}{2}x^2\Big|_1^3 = \frac{9}{2} - \frac{1}{2} = 4$$

- But Several functions cannot be solved analytically.

$$\int_0^a e^{-x^2}\,dx = ?$$

NUMBER REPRESENTATION

- You are familiar with the decimal (Base 10) system with digits from 0 – 9.
- Standard Representations:

±	3 1 2	.	4 5
sign	integral part		fraction part

$$312.45 = 3 \times 10^2 + 1 \times 10^1 + 2 \times 10^0 + 4 \times 10^{-1} + 5 \times 10^{-2}$$

NORMALIZED FLOATING POINT REPRESENTATION

- Exactly one non-zero digit appears before the decimal point.

$$\pm \quad \underbrace{d.\,f_1\ f_2\ f_3\ f_4}_{\text{mantissa}} \times \underbrace{10^{\pm n}}_{\text{exponent}}$$

sign mantissa exponent

$$d \neq 0, \qquad \pm n : \text{signed exponent}$$

- Hence 0.00000024 becomes 2.4×10^{-7}

- This is an efficient way of representing and storing very small or very large numbers.

NUMBER REPRESENTATION

- Computers store numbers with the binary (Base 2) system with digits from 0 and 1.
- Standard Representations:

$$1.101_2 = (1 + 1 \times 2^{-1} + 0 \times 2^{-2} + 1 \times 2^{-3}) = 1.625_{10}$$

- Conversion between bases is done by computers. Converting 0.375_{10} to binary:

$$0.375 \cdot 2 = 0 + 0.75$$
$$0.75 \cdot 2 = 1 + 0.5$$
$$0.5 \cdot 2 = 1 + 0$$

NUMBER REPRESENTATION

- Converting 0.375_{10} to binary:

$$0.375 \times 2 = 0 + 0.75$$

$$0.75 \times 2 = 1 + 0.5$$

$$0.5 \times 2 = 1 + 0$$

- Hence

$$0.375_{10} = 0.011$$

NUMBER REPRESENTATION

- Converting 0.1_{10} to binary:

$$
\begin{aligned}
0.1 \cdot 2 &= 0 + 0.2 \\
0.2 \cdot 2 &= 0 + 0.4 \\
0.4 \cdot 2 &= 0 + 0.8 \\
0.8 \cdot 2 &= 1 + 0.6 \\
0.6 \cdot 2 &= 1 + 0.2 \\
0.2 \cdot 2 &= 0 + 0.4 \\
0.4 \cdot 2 &= 0 + 0.8 \\
&\ldots
\end{aligned}
$$

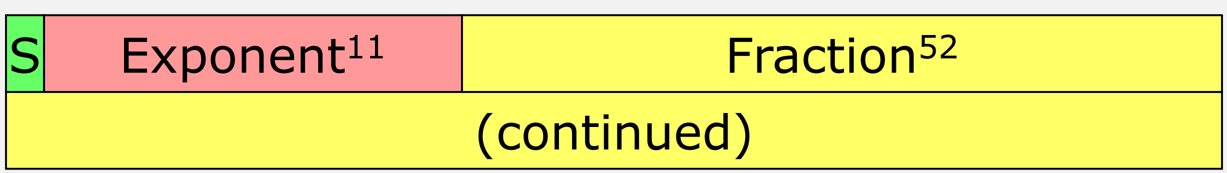
- Hence

$$1.1_{10} = 1.0001100\ldots_{2}$$

NUMBER REPRESENTATION

- Single Precision (32-bit representation)
  - 1-bit Sign + 8-bit Exponent + 23-bit Fraction

- Double Precision (64-bit representation)
  - 1-bit Sign + 11-bit Exponent + 52-bit Fraction



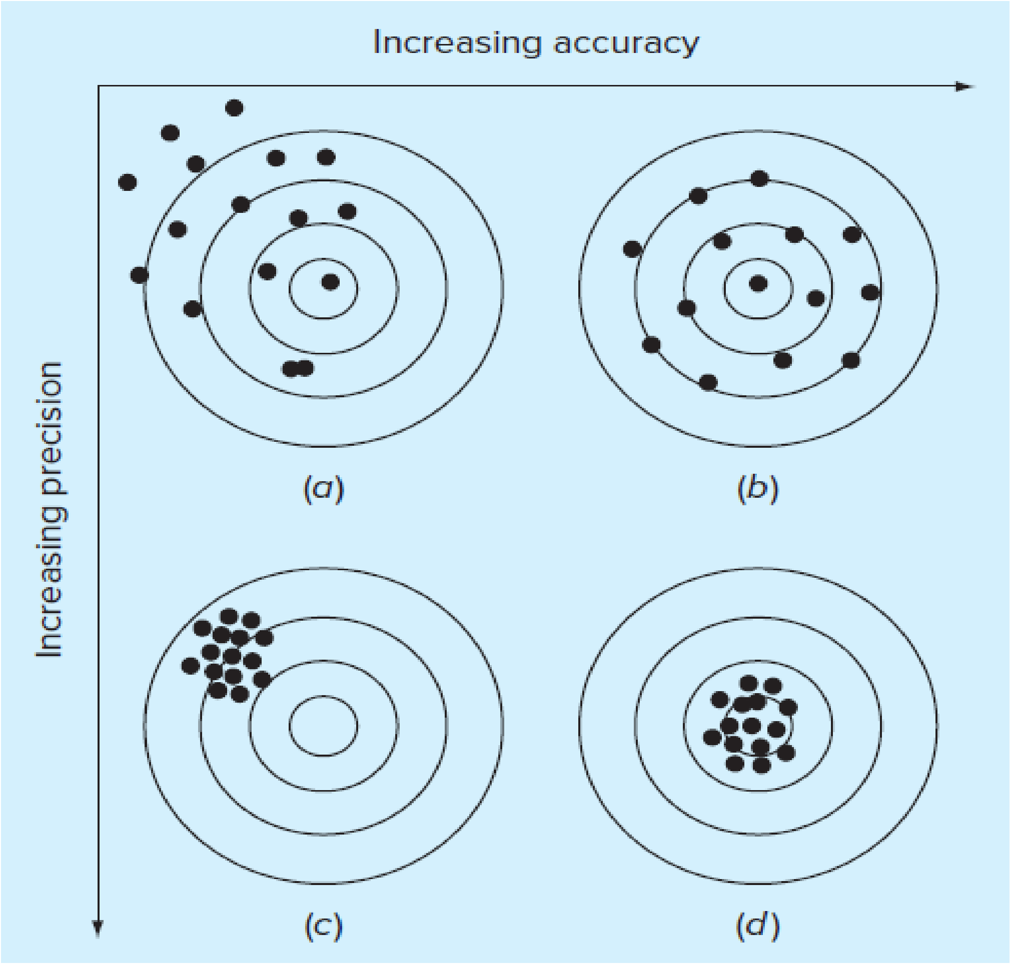
ROUNDING AND CHOPPING

- The computer’s way of representing numbers like 1.1
  - Rounding : Replacing the number with the nearest machine number.
  - Chopping: Throwing away all extra digits

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 called *bias*) is defined as systematic deviation from the truth.
- *Imprecision* (also called *uncertainty*) refers to the magnitude of the scatter.
- Numerical methods should be sufficiently accurate and precise.
- The term error is used to describe the inaccuracy and imprecision of solutions.

ACCURACY AND PRECISION



(a) Inaccurate and imprecise

(b) Accurate and imprecise

(c) Inaccurate and precise

(d) Accurate and precise

THE ISSUE OF ERRORS

- Error is the value of inaccuracy on a measure/quantity.
- It is important to account for errors in numerical computations.
  - Since solutions are approximations and not exact values
  - In iterative methods, errors accumulate into a large value
  - Errors must be as small as possible implying high accuracy
- The point of numerical analysis is to analyze methods that are used to give approximate number solutions to situations where it is unlikely to find the real solution quickly.
- We improve upon these methods to reduce the error generated by computer calculation.

CALCULATING ERRORS

- True Error: Calculated if the true outcome of a computation is known.
- Estimated Error: Calculated if the true outcome of a computation is not known.
- Absolute Error: Error given with the same units as the measure itself.
- Relative Error: Error expressed in relation to the measured value.

	True Error	Estimated Error
Absolute	Absolute true error	Absolute estimated error
Relative	Relative true error	Relative estimated error

CALCULATING ERRORS

Absolute True Error:

$$E_t = |\, true\ value\ - approximated\ value\, |$$

Relative True Error:

$$E_{rel} = |\frac{true\ value\ - approximated\ value}{true\ value}|$$

CALCULATING ERRORS

Absolute Estimated Error:

$$E_t = |\, current\ estimate\ - previous\ estimate|$$

Relative Estimated Error:

$$E_{rel} = |\frac{current\ estimate\ - previous\ estimate}{current\ estimate}|$$

CALCULATING ERRORS

- We normally want the error in a calculation e_a to be lower than a prespecified value e_s
- In these cases, computation is repeated until:

$$\boxed{|e_a| < e_s}$$

- This relationship is normally called the stopping criterion.

CHOOSING A STOPPING CRITERION

- To choose a stopping criterion such that the relative estimated error is to at least n significant figures, we can use the equation:

$$e_s = (0.5 \times 10^{2-n})\%$$

- Hence calculating for 3 significant figure correctness:

$$e_s = (0.5 \times 10^{2-3})\%$$

$$e_s = (0.5 \times 0.1)\%$$

$$e_s = 0.05\%$$

EXAMPLE: CALCULATING $e^{0.5}$

- The expression e^x can be expressed as the infinite series:

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

- Calculate $e^{0.5}$ to 3 significant figure correctness

STEP 1:

$$e^{0.5} = 1$$

- Relative True Error:

$$e_t = \left|\frac{1.648721 - 1}{1.648721}\right| \times 100$$

$$e_t = 39.3\%$$

EXAMPLE: CALCULATING $e^{0.5}$

STEP 2:

$$e^{0.5} = 1 + 0.5$$

$$e^{0.5} = 1.5$$

- Relative True Error:

$$e_t = \left|\frac{1.648721 - 1.5}{1.648721}\right| \times 100$$

$$e_t = 9.02\%$$

- Relative Estimated Error:

$$e_t = \left|\frac{1.5 - 1}{1.5}\right| \times 100$$

$$e_t = 33.3\%$$

EXAMPLE: CALCULATING $e^{0.5}$

STEP 3:

$$e^{0.5} = 1 + 0.5 + \frac{0.5^2}{2}$$

$$e^{0.5} = 1.625$$

- Relative True Error:

$$e_t = \left|\frac{1.648721 - 1.625}{1.648721}\right| \times 100$$

$$e_t = 1.438\%$$

- Relative Estimated Error:

$$e_t = \left|\frac{1.625 - 1.5}{1.625}\right| \times 100$$

$$e_t = 7.692\%$$

EXAMPLE: CALCULATING $e^{0.5}$

- Continuing the process gives:

Terms	Result	ε_t, %	ε_a, %
1	1	39.3	
2	1.5	9.02	33.3
3	1.625	1.44	7.69
4	1.645833333	0.175	1.27
5	1.648437500	0.0172	0.158
6	1.648697917	0.00142	0.0158

- At the sixth step: $|e_a| < e_s$

- Hence $e^{0.5} = 1.648697917$