Numerical Methods and Optimization

Topic 1:

Introduction to Numerical Methods

Lectures 1-4:

Lecture 1
Introduction to Numerical Methods

- What are **NUMERICAL METHODS**?
- Why do we need them?
- Topics covered in NMO.

Reading Assignment: Pages 3-10 of textbook

Numerical Methods

Numerical Methods:

Algorithms that are used to obtain numerical solutions of a mathematical problem.

Why do we need them?

1. No analytical solution exists,
2. An analytical solution is difficult to obtain or not practical.

What do we need?

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,...).

Outlines of the Course

- Number Representation
- Approximate solution of nonlinear Equations
- Solution of linear Equations (Direct methods)
- Solution of linear Equations (Iterative methods)
- Polynomial Interpolation
- Least Squares approximation
- Numerical Integration

Solution of Nonlinear Equations

- Some simple equations can be solved analytically:

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

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

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

- Many other equations have no analytical solution:

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

Methods for Solving Nonlinear Equations

- **Bisection Method**
- **Newton-Raphson Method**
- **Secant Method**

Solution of Systems of Linear Equations

$$x_1 + x_2 = 3$$

$$x_1 + 2x_2 = 5$$

We can solve it as:

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

$$\Rightarrow x_2 = 2,\ x_1 = 3 - 2 = 1$$

What to do if we have

1000 equations in 1000 unknowns.

Cramer's Rule is Not Practical

Cramer's Rule can be used to solve the system:

$$x_1 = \frac{\begin{vmatrix} 3 & 1 \\ 5 & 2 \end{vmatrix}}{\begin{vmatrix} 1 & 1 \\ 1 & 2 \end{vmatrix}} = 1, \qquad x_2 = \frac{\begin{vmatrix} 1 & 3 \\ 1 & 5 \end{vmatrix}}{\begin{vmatrix} 1 & 1 \\ 1 & 2 \end{vmatrix}} = 2$$

But Cramer's Rule is not practical for large problems.

To solve N equations with N unknowns, we need $(N+1)(N-1)N!$ multiplications.

To solve a 30 by 30 system, 2.3×10^{35} multiplications are needed.

A super computer needs more than 10^{20} years to compute this.

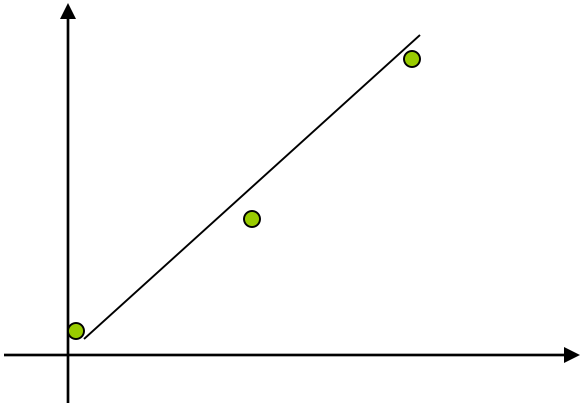
Methods for Solving Systems of Linear Equations

- **Naive Gaussian Elimination**
- **Gaussian Elimination with Scaled Partial Pivoting**
- **Algorithm for Tri-diagonal Equations**

Curve Fitting

- Given a set of data:

x	0	1	2
y	0.5	10.3	21.3

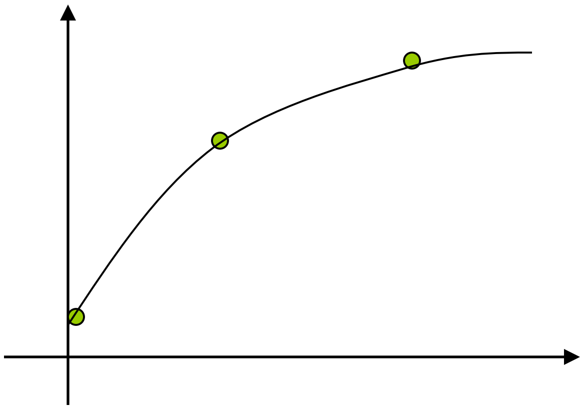


- Select a curve that best fits the data. One choice is to find the curve so that the sum of the square of the error is minimized.

Polynomial Interpolation

- Given a set of data:

x_i	0	1	2
y_i	0.5	10.3	15.3



- Find a polynomial *P(x)* whose graph passes through all tabulated points.

$$y_i = P(x_i) \quad if\ x_i \text{ is in the table}$$

Methods for Curve Fitting

- **Least Squares**
  - **Linear Regression**
  - **Nonlinear Least Squares Problems**

- **Interpolation**
  - **Newton Polynomial Interpolation**
  - **Lagrange Interpolation**

Integration

- Some functions can be integrated analytically:

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

But many functions have no analytical solutions :

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

Methods for Numerical Integration

- **Upper and Lower Sums**
- **Trapezoid Method**
- **Romberg Method**
- **Gauss Quadrature**

Summary

- **Numerical Methods:** Algorithms that are used to obtain numerical solution of a mathematical problem.
- **We need them when** No analytical solution exists or it is difficult to obtain it.

Topics Covered in the Course

- Solution of Nonlinear Equations
- Solution of Linear Equations
- Curve Fitting
  - Least Squares
  - Interpolation
- Numerical Integration

Lecture 2

Number Representation and Accuracy

- Number Representation
- Normalized Floating Point Representation
- Significant Digits
- Accuracy and Precision
- Rounding and Chopping

Reading Assignment: Chapter 3

Representing Real Numbers

- You are familiar with the decimal system:

$$312.45 = 3\times10^{2} + 1\times10^{1} + 2\times10^{0} + 4\times10^{-1} + 5\times10^{-2}$$

- Decimal System: Base = 10 , Digits (0,1,…,9)

- Standard Representations:

$\pm$	3 1 2	.	4 5
sign	integral part		fraction part

Normalized Floating Point Representation

- Normalized Floating Point Representation:

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

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

- Scientific Notation: Exactly one non-zero digit appears before decimal point.
- Advantage: Efficient in representing very small or very large numbers.

Binary System

- Binary System: Base = 2, Digits {0,1}

$$\underset{\text{sign}}{\pm} \quad \underbrace{1.\ f_1\ f_2\ f_3\ f_4}_{\text{mantissa}} \times 2^{\pm n}$$

signed exponent

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

Fact

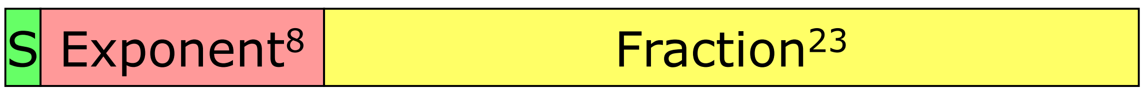
- Numbers that have a finite expansion in one numbering system may have an infinite expansion in another numbering system:

$$(1.1)_{10} = (1.00011001100110 0\ldots)_2$$

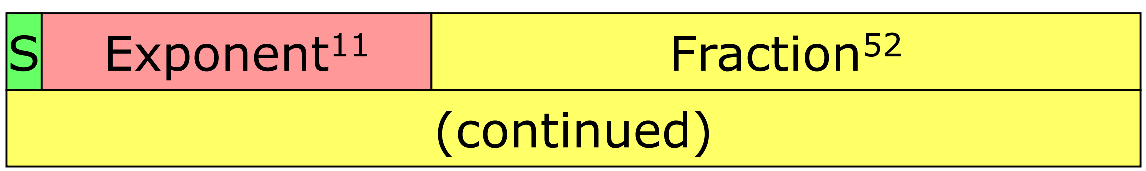
- You can never represent 1.1 exactly in binary system.

IEEE 754 Floating-Point Standard

- 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



Significant Digits

- Significant digits are those digits that can be used with confidence.
- Single-Precision: 7 Significant Digits

  $1.175494\ldots \times 10^{-38}$ to $3.402823\ldots \times 10^{38}$
- Double-Precision: 15 Significant Digits

  $2.2250738\ldots \times 10^{-308}$ to $1.7976931\ldots \times 10^{308}$

Remarks

- Numbers that can be exactly represented are called machine numbers.
- Difference between machine numbers is not uniform
- Sum of machine numbers is not necessarily a machine number

Calculator Example

- Suppose you want to compute:

  3.578 * 2.139

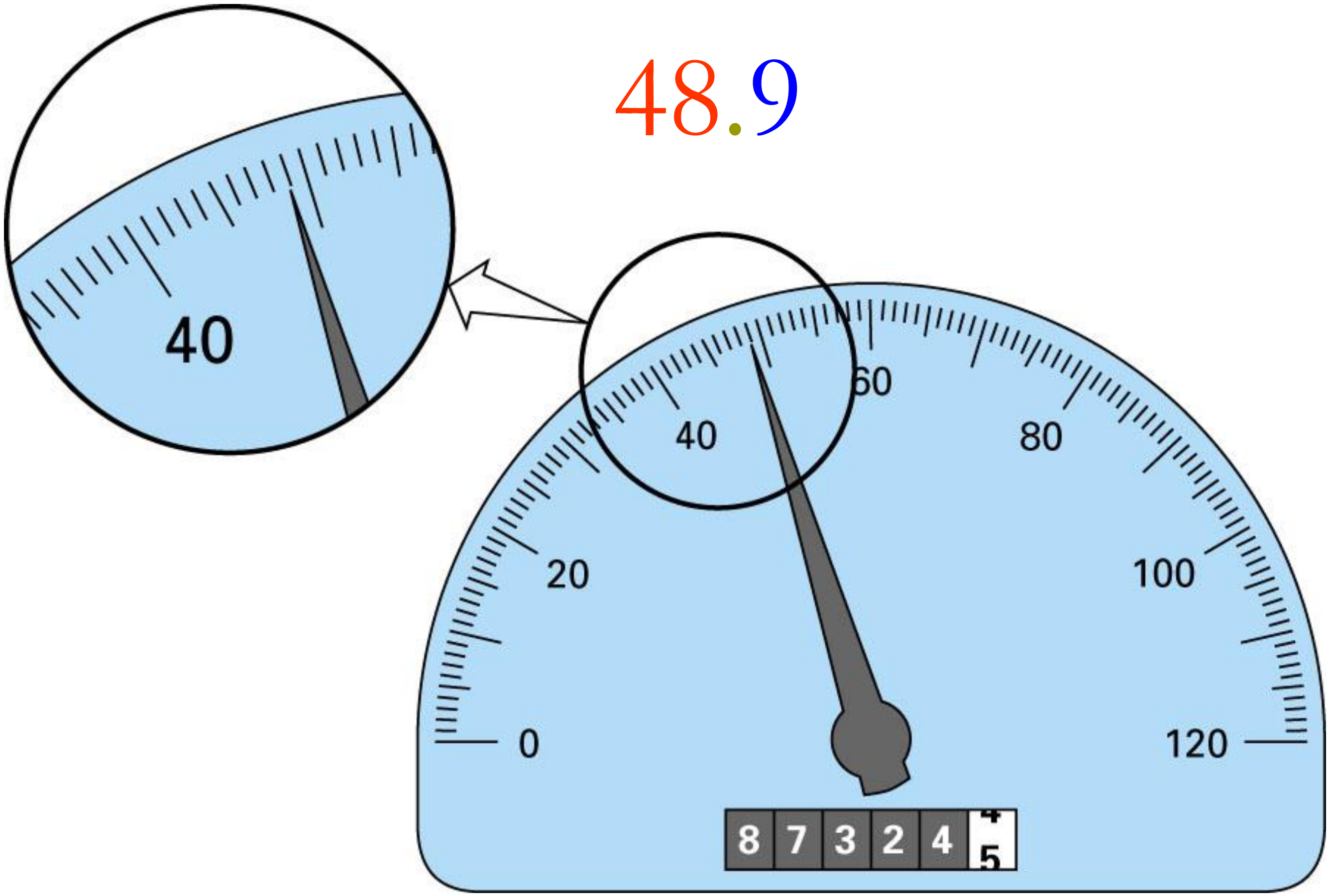
  using a calculator with two-digit fractions

3.57	*	2.13	=	7.60

True answer: 7.653342

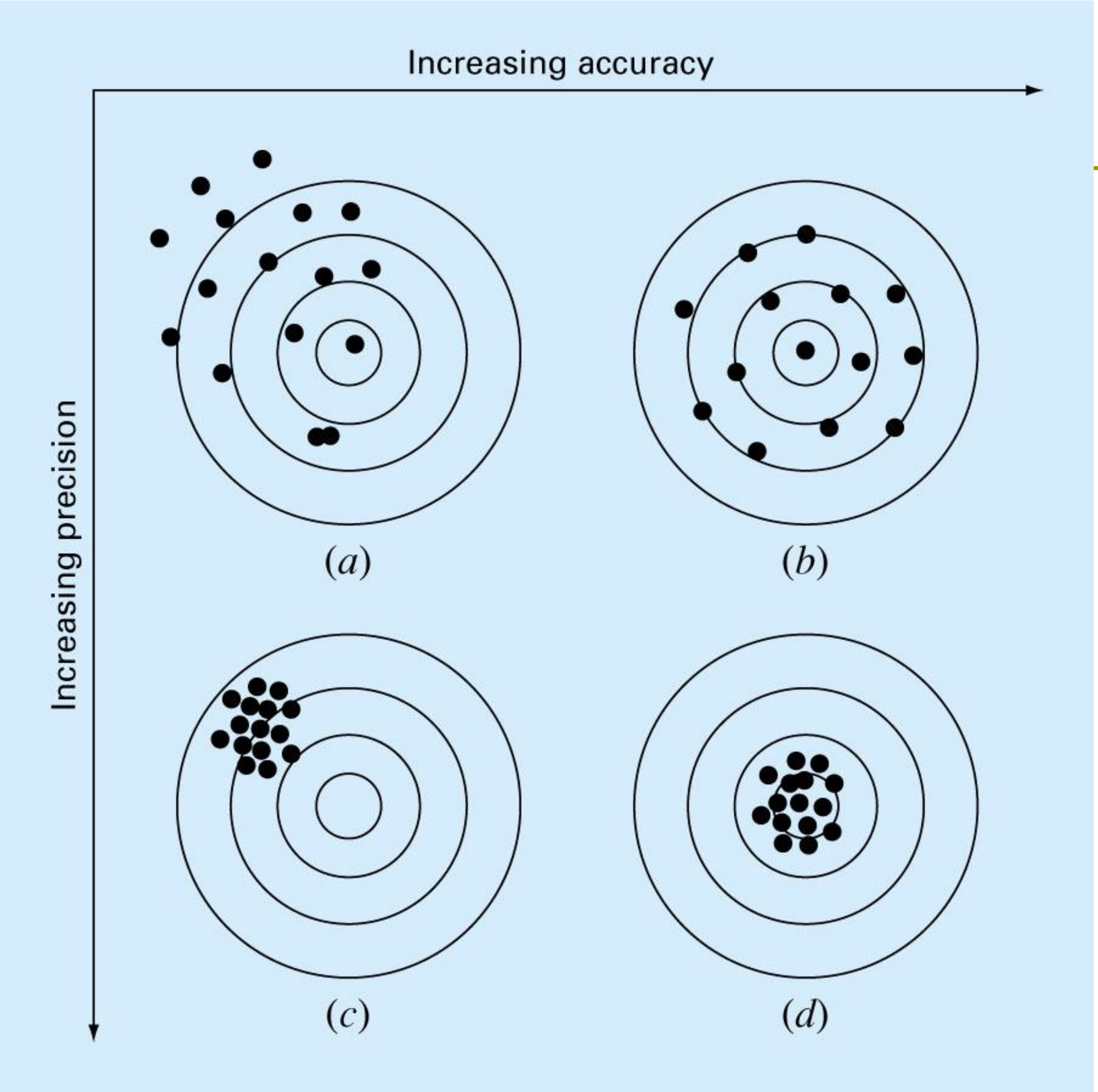
Significant Digits - Example

48.9



Accuracy and Precision

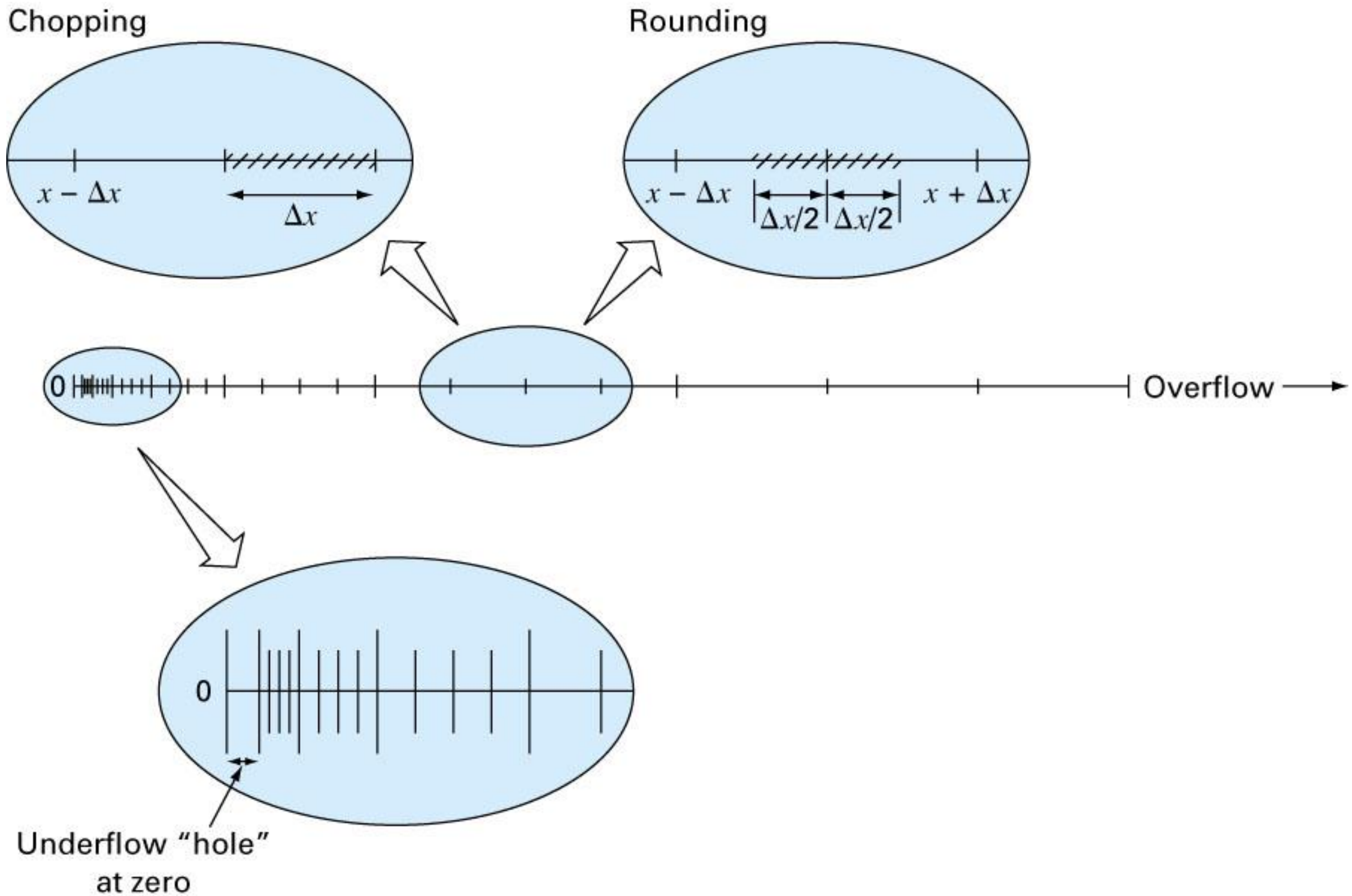
- Accuracy is related to the closeness to the true value.
- Precision is related to the closeness to other estimated values.


Increasing accuracy
Increasing precision
(a)
(b)
(c)
(d)

Rounding and Chopping

- Rounding: Replace the number by the nearest machine number.
- Chopping: Throw all extra digits.

Rounding and Chopping



Error Definitions – True Error

Can be computed if the true value is known:

Absolute True Error

$$E_t = \left| \text{true value} - \text{approximation} \right|$$

Absolute Percent Relative Error

$$\varepsilon_t = \left| \frac{\text{true value} - \text{approximation}}{\text{true value}} \right| * 100$$

Error Definitions – Estimated Error

When the true value is not known:

$$\text{Estimated Absolute Error}$$

$$E_a = \left| \text{current estimate} - \text{previous estimate} \right|$$

$$\text{Estimated Absolute Percent Relative Error}$$

$$\varepsilon_a = \left| \frac{\text{current estimate} - \text{previous estimate}}{\text{current estimate}} \right| *100$$

Notation

We say that the estimate is correct to *n* decimal digits if:

$$|\,\mathrm{Error}\,| \le 10^{-n}$$

We say that the estimate is correct to *n* decimal digits **rounded** if:

$$|\,\mathrm{Error}\,| \le \frac{1}{2} \times 10^{-n}$$

Summary

- Number Representation

  Numbers that have a finite expansion in one numbering system may have an infinite expansion in another numbering system.

- Normalized Floating Point Representation
  - Efficient in representing very small or very large numbers,
  - Difference between machine numbers is not uniform,
  - Representation error depends on the number of bits used in the mantissa.