

Engineering Analysis ENG 3420 Fall 2009

Dan C. Marinescu

Office: HEC 439 B

Office hours: Tu-Th 11:00-12:00

Lecture 6

- Last time:

- Internal representations of numbers and characters in a computer.
- Arrays in Matlab
- Matlab program for solving a quadratic equation

- Today:

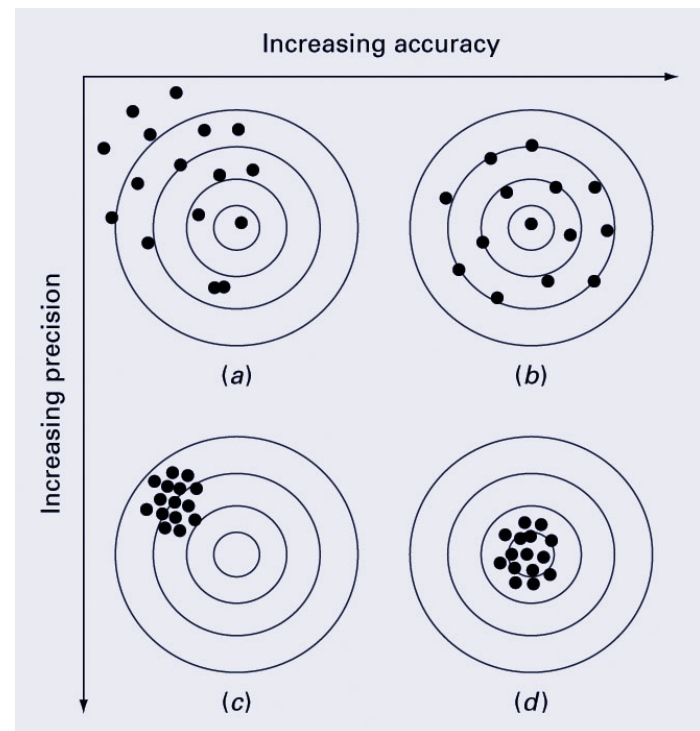
- Roundoff and truncation errors
- More on Matlab

- Next Time

- More on approximations.

Accuracy versus 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.
- a) inaccurate and imprecise
b) accurate and imprecise
c) inaccurate and precise
d) accurate and precise



Errors when we know the true value

- True error (E_t) → the difference between the true value and the approximation.
- Absolute error ($|E_t|$) → the absolute difference between the true value and the approximation.
- True fractional relative error → the true error divided by the true value.
- Relative error (ε_t) → the true fractional relative error expressed as a percentage.
- For iterative processes, the error can be approximated as the difference in values between successive iterations.

Stopping criterion

- Often, we are interested in whether the absolute value of the error is lower than a pre-specified tolerance ε_s . For such cases, the computation is repeated until $|\varepsilon_a| < \varepsilon_s$

Roundoff Errors

- *Roundoff errors* arise because
 - Digital computers have size and precision limits on their ability to represent numbers.
 - Some numerical manipulations are highly sensitive to roundoff errors.

Floating Point Representation

- By default, MATLAB has adopted the IEEE double-precision format in which eight bytes (64 bits) are used to represent floating-point numbers:

$$n = \pm(1+f) \times 2^e$$

- The sign is determined by a sign bit
- The mantissa f is determined by a 52-bit binary number
- The exponent e is determined by an 11-bit binary number, from which 1023 is subtracted to get e

Floating Point Ranges

- Values of -1023 and +1024 for e are reserved for special meanings, so the exponent range is -1022 to 1023.
- The largest possible number MATLAB can store has
 - f of all 1's, giving a significand of $2-2^{-52}$, or approximately 2
 - e of 11111111110_2 , giving an exponent of $2046-1023=1023$
 - This yields approximately $2^{1024}=1.7997 \times 10^{308}$
- The smallest possible number MATLAB can store with full precision has
 - f of all 0's, giving a significand of 1
 - e of 00000000001_2 , giving an exponent of $1-1023=-1022$
 - This yields $2^{-1022}=2.2251 \times 10^{-308}$

Floating Point Precision

- The 52 bits for the mantissa f correspond to about 15 to 16 base-10 digits.
- The machine epsilon - the maximum relative error between a number and MATLAB's representation of that number, is thus
 $2^{-52} = 2.2204 \times 10^{-16}$

Roundoff Errors in Arithmetic Operations

- Roundoff error occur :
 - *Large computations* - if a process performs a large number of computations, roundoff errors may build up to become significant
 - *Adding a Large and a Small Number* - Since the small number's mantissa is shifted to the right to be the same scale as the large number, digits are lost
 - *Smearing* - Smearing occurs whenever the individual terms in a summation are larger than the summation itself.
 - $(x+10^{-20})-x = 10^{-20}$ mathematically, but $x=1$; $(x+1e-20)-x$ gives a 0 in MATLAB!

Truncation Errors

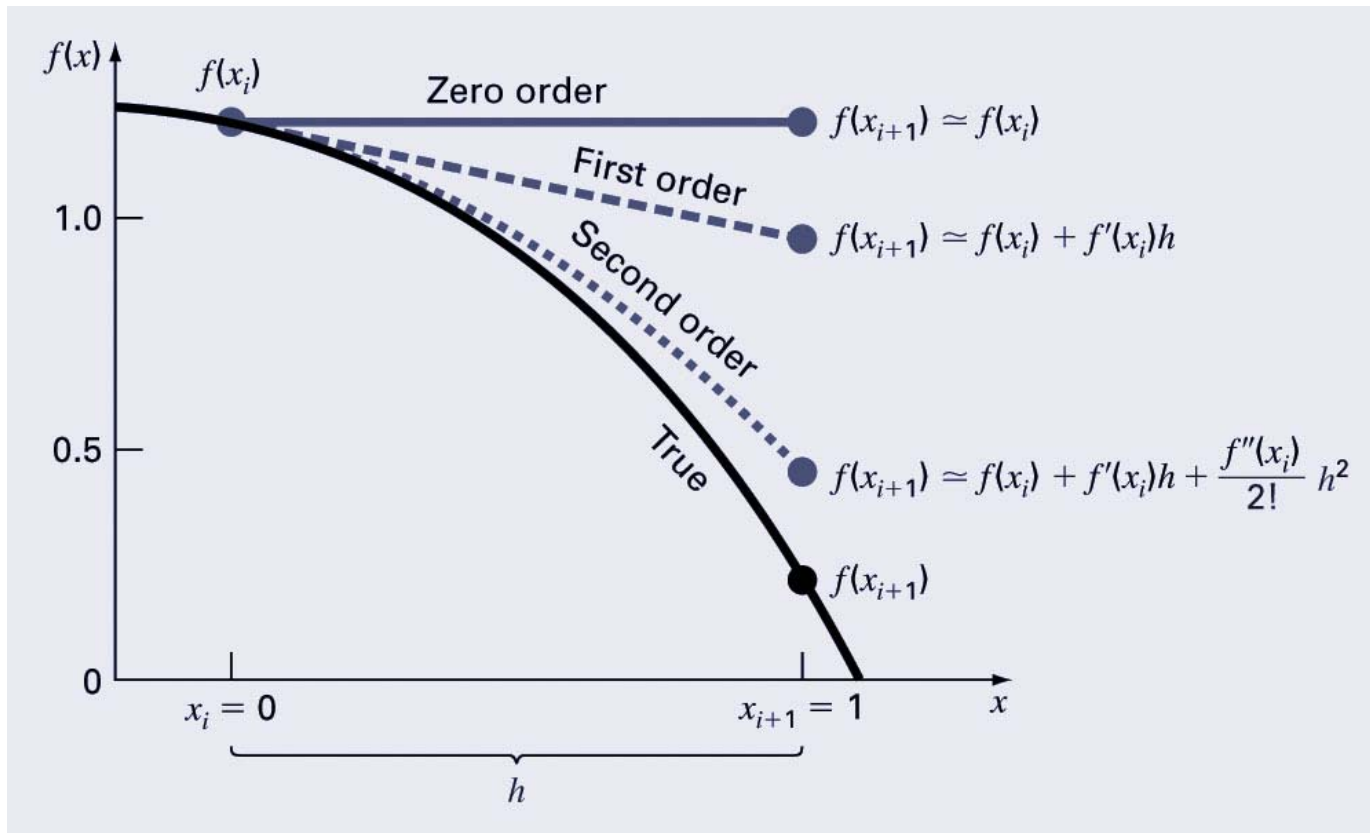
- *Truncation errors* result from using an approximation in place of an exact mathematical procedure.
- Example 1: approximation to a derivative using a finite-difference equation:

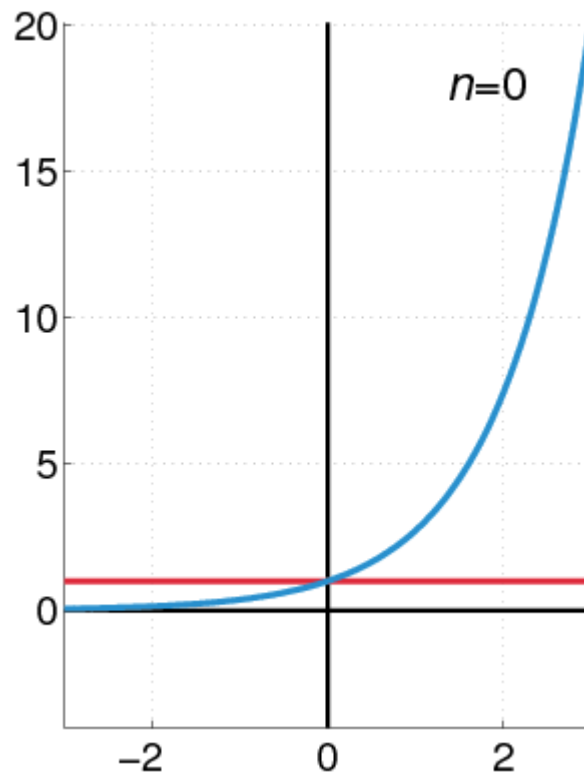
$$\frac{dv}{dt} \cong \frac{\Delta v}{\Delta t} = \frac{v(t_{i+1}) - v(t_i)}{t_{i+1} - t_i}$$

Example 2: The Taylor Series

The Taylor Series

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





Truncation Error

- In general, the n th order Taylor series expansion will be exact for an n th order polynomial.
- In other cases, the remainder term R_n is of the order of h^{n+1} , meaning:
 - The more terms are used, the smaller the error, and
 - The smaller the spacing, the smaller the error for a given number of terms.

Functions

Function [out1, out2, ...] = funname(in1, in2, ...)

- function funname that

- accepts inputs in1, in2, etc.
- returns outputs out1, out2, etc.
- Example: function [r1,r2,i1,i2] = quadroots(a,b,c)

- Before calling a function you need to

- Use the edit window and create the function
- Save the edit window in a .m file e.g., quadroots.m

- Example

```
function [mean,stdev] = stat(x)
n = length(x);
mean = sum(x)/n;
stdev = sqrt(sum((x-mean).^2/n));
```

Subfunctions

- A function file can also contain a *primary function* and one or more *subfunctions*
- The primary function → is listed first in the M-file - its function name should be the same as the file name.
- Subfunctions
 - are listed below the primary function.
 - *only* accessible by the main function and subfunctions within the same M-file and *not* by the command window or any other functions or scripts.

Example of a subfunction

- `function [mean,stdev] = stat(x)`
 `n = length(x);`
 `mean = avg(x,n);`
 `stdev = sqrt(sum((x-avg(x,n)).^2)/n);`

 `function mean = avg(x,n)`
 `mean = sum(x)/n;`

`avg` is a subfunction within the file `stat.m`: