CS 3331 Numerical Methods

Lecture 1: Introduction

Cherung Lee

About this course

- Text: *Applied Numerical Analysis using Matlab, 2nd edition*
  - Laurene V. Fausett. (LVF)
- TA: TBA
- Website:

  `http://www.cs.nthu.edu.tw/~cherung/teaching/cs3331/cs3331.html`
- Office hours: Tuesday 2:00-3:00, Friday 3:00-4:00 (or by appointment).

Tentative agenda

1	Introduction	(chap 1)
2	Functions of one variable	(chap 2)
3	Linear systems	(chap 3, 4)
4	Linear least square problems	(chap 4,9)
5	Eigenvalues/eigenvectors	(chap 5)
6	Iterative methods for solving linear systems	(chap 6)
7	Interpolation	(chap 8)
8	Approximation	(chap 9)
9	Fourier methods	(chap 10)
10	Numerical differentiation and integration	(chap 11)
11	Numerical optimization	(chap 2, 7)

Pre-requirements

- Calculus: mean value theory, Taylor expansion ...

- Linear algebra: symmetric matrix, orthogonal matrix, eigen-values/eigenvectors ...

- Computer science: floating-point arithmetic, algorithm ...

- Programming: Matlab, c/c++

Grading

- Quiz (50%)
  - every 1-2 weeks
- Assignment (50%)
  - 4-5 (programming) projects

QUESTIONS?

Introduction

Numerical methods

- Numerical vs. Analytical

- Continuous vs. Discrete

- Examples: solving nonlinear equations, linear systems, numerical integration ...

Nonlinear equations LVF pp.4-5

- Solve $f(x) = x^2 - 3 = 0$ ($x = \pm\sqrt{3}$).

- Fixed point iterations:

  – rewrite $x^2 - 3 = 0$ as $x = \frac{1}{2}(x + \frac{3}{x})$

$$
\begin{aligned}
x_0 &= 1 \\
x_1 &= \frac{1}{2}\left(1 + \frac{3}{1}\right) = 2 \\
x_2 &= \frac{1}{2}\left(2 + \frac{3}{2}\right) = \frac{7}{4} \\
\vdots &= \vdots
\end{aligned}
$$

Linear system LVF pp.6-7

$$\begin{array}{lcrcrcl} L_1 & : & 4x_1 & + & x_2 & = & 6 \\ M_1 & : & -x_1 & + & 5x_2 & = & 9 \end{array}$$

- Gaussian elimination.
  - Solve $\begin{pmatrix} 4 & 1 \\ -1 & 5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 6 \\ 9 \end{pmatrix}$

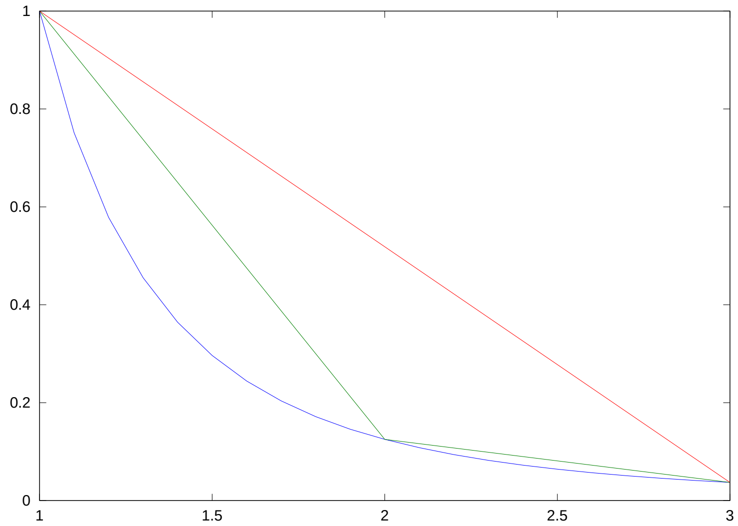
$$\Rightarrow \begin{pmatrix} 1 & 0 \\ 1/4 & 1 \end{pmatrix} \begin{pmatrix} 4 & 1 \\ -1 & 5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1/4 & 1 \end{pmatrix} \begin{pmatrix} 6 \\ 9 \end{pmatrix}$$

$$\Rightarrow \begin{pmatrix} 4 & 1 \\ 0 & 5.25 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 6 \\ 10.5 \end{pmatrix}$$

$\Rightarrow$ Back-substitution: $x_2 = 2, x_1 = \frac{1}{4}(6 - 1 * 2) = 1$

Numerical integration LVF pp.8-9

- Compute $I = \int_1^3 \frac{1}{x^3} dx$
- Trapezoid rule.



Method	Formula	Result
analytical solution	$\frac{-1}{2x^2}\vert_1^3$	0.444444
one subdivision	$2/2[1+1/27]$	1.037037
two subdivisions	$1/2[1+1/8]+1/2[1/8+1/27]$	0.643518

Basic issues of numerical methods

- Accuracy: (errors)

- Speed: (cost)
  - Time complexity (operation counts).
  - Converge rate.
  - Machine/software properties.

Real Numbers in Computer

Real number in computer LVF pp.13

- Binary representation

$$
\begin{aligned}
N &= (d_k d_{k-1} \cdots d_1 d_0 . d_{-1} \cdots d_{-p})_b \\
&= d_k 2^k + d_{k-1} 2^{k-1} \cdots 2 d_1 + d_0 + d_{-1}\frac{1}{2} + d_{-2}\frac{1}{4} + \cdots + d_{-p}\frac{1}{2^p}
\end{aligned}
$$

  - $d_1, d_2 \cdots, d_p$ are in $\{0, 1\}$.
  - Alternative representation $(d_k . d_{k-1} \cdots d_1 d_0 d_{-1} \cdots d_{-p})_b \times 2^k$

- Floating-point vs. fixed-point

- Multiprecision and arbitrary precision

IEEE 754 LVF pp.13-15

s	exponent	mantissa

$$a = (-1)^s \times 2^{\text{exponent}-\text{exponent bias}} \times 1.\text{mantissa}$$

	single (32bits)	double (64bits)
s	1 bit	1 bit
exponent	8 bits ($e = 8$)	11 bits ($e = 11$)
mantissa	23 bits	52 bits

- Normalization: the leading digit is 1.
  - Subnormal: when the exponent is the smallest number, the leading digit is allowed to be zero
- Exponent bias: exponent are shifted by $2^{e-1} - 1$.

IEEE 754–continue

- Special numbers
  - `Inf`: (Infinite) exponent=$2^e - 1$, mantissa=0.
  - `NaN`: (Not a Number) exponent=$2^e - 1$, mantissa$\neq 0$.
  - `Zeros`: exponent=0, mantissa=0.
- Representable ranges:

absolute values	single (32bits)	double (64bits)
Min. normal	2^{-126}	2^{-1022}
Min. subnormal	2^{-149}	2^{-1074}
Max. finite	$(1-2^{-24})2^{128}$	$(1-(2)^{-53})2^{1024}$

- Overflow and underflow

Multiple precision arithmetic

- Double-double approach: use two doubles for a number.
- Extend floating-point format: use more bits for exponent and mantissa.
- Software:
  - vpa (Matlab Symbolic Math Toolbox),
  - GNU Multi-Precision Library (c/c++),
  - ARPREC and MPFUN (Fortran),
  - Bignum and BigInteger (Java).

Errors

Source of errors

- From measurement/sampling.
- From modeling.
- From number representation.
- From algorithm.

Measuring errors LVF pp.16

- Let x^* be the exact value.
  - Absolute error: Error(x)=$|x - x^*|$.
  - Relative error: Rel Error(x)=$|x - x^*|/|x^*|$.

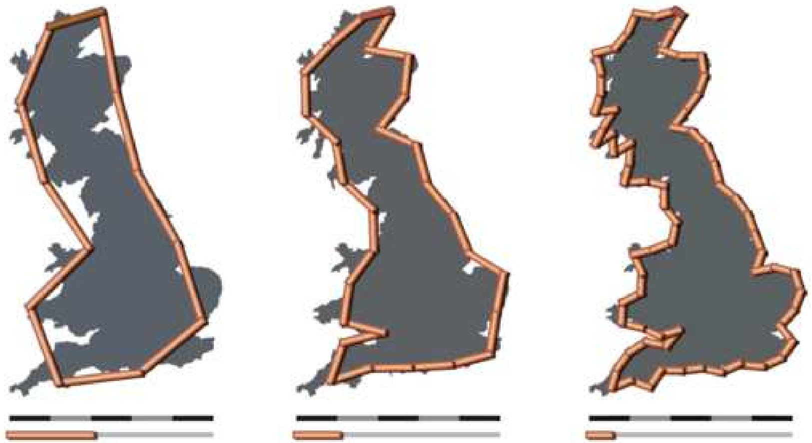
- *Significant digits*: The number x is said to have t significant digits if t is the largest nonnegative integer for which

$$\frac{|x - x^*|}{|x^*|} < 5 \times 10^{-t}.$$

- Big-Oh notation: $f(h) = O(g(h))$ if $f(h) \leq c|g(h)|$ for some positive constant c when $h \to 0$. LVF pp.21

Errors from modeling

- Benoit Mandelbrot, *How Long Is the Coast of Britain? Statistical Self-Similarity and Fractional Dimension*, Science Vol 156, 1967.



Images from Wikipeida

Errors from inexact representation LVF pp.17

- Machine epsilon (*eps*): the difference between 1 and the smallest exactly representable number greater than one.
  - Single: $2^{-24} = 5.96 \times 10^{-8}$
  - Double: $2^{-53} = 1.1 \times 10^{-16}$
- Rounding
  - Round to nearest even
  - Example: 1/3 (0.01010101→0.0101010) and 1/5 (0.00110011→0.0011010)
  - Roundoff error: 1/3+1/6

Errors from floating-point arithmetic LVF pp.17

- Cancellation: loss of significance
- Example: compute $2^1 \times 0.100 - 2^0 \times 0.111$
  - Alignment: $2^1 \times 0.100 - 2^1 \times 0.011$
  - Result: $2^1 \times 0.001 = 2^{-1} \times 0.100$
  - But the exact result is $2^{-2} \times 0.100$.
  - Relative error: $\dfrac{|2^{-1} \times 0.100 - 2^{-2} \times 0.100|}{|2^{-2} \times 0.100|} = 1.$

Errors from numerical algorithm LVF pp.18

- Example: Solve $ax^2 + bx + c = 0$ using $x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$.
  - Ex. $x^2 + 100x + 1 = 0$
    - * $\sqrt{100^2 - 4} = \sqrt{9996}$ is rounded to 100.

      $\Rightarrow x_1 = (-100+100)/2 = 0, x_2 = (-100-100)/2 = -100.$
    - * Actual solution $x_1 = -0.01, x_2 = -99.99$

      $$\Rightarrow \text{RE}(x_1) = 1, \text{RE}(x_2) = 10^{-4}.$$
  - Use $x = \frac{-2c}{b + \sqrt{b^2 - 4ac}}$ for x_1. ($x_1 = -0.01$)

Forward and backward error analysis

- What we concern is the errors in the solutions (output).

- Example: evaluate $f(x)$.
  - Let $y = f(x)$ and $\widehat{y}$ be the computed result.
  - Forward error: error of the output: $|y - \widehat{y}|$
  - Backward error: given the computed output $\widehat{y}$, backward error is the smallest $|\Delta x|$ such that $f(x + \Delta x) = \widehat{y}$.
  - Example: Evaluate $f(x) = \sqrt{x}$ at $x = 1/36$.

Condition number

- Let x be an input and $f(x)$ be its output. $\tilde{x}$ is a perturbed x.
- Condition number $= \dfrac{|f(\tilde{x}) - f(x)|/|f(x)|}{|\tilde{x} - x|/|x|}$
- If $f(x)$ is continuously differentiable around x, the condition number is

$$\left|\frac{xf'(x)}{f(x)}\right|$$

- ex: $f(x) = x^{-1}$ for $x > 0$.

Last year's notes (http://www.cs.nthu.edu.tw/∼cchen)

Performance Issues

Computational efforts LVF pp.30

- flops: floating-point operations
- Example: polynomial evaluation $P(x) = a_n x^n + \cdots a_1 x + a_0$
  - Direct method: evaluate $a_i x^i$ one by one

$$\text{flops} = \left(\sum_{k=1}^{n} k\right) + n = \frac{n(n+3)}{2}$$

  - Horner's algorithm: evaluate

$$P(x) = (\ldots((a_n x) + a_{n-1})x + \cdots a_1)x + a_0$$

$$\text{flops} = \sum_{k=1}^{n} 2 = 2n$$

Convergence LVF pp.22-23

- A sequence $\{y_k\}$ converges to y^* iff $\lim_{k\to\infty} |y_k - y^*| = 0$
- If there existing some $\lambda > 0, p > 0$ such that

$$\lim_{k\to\infty} \frac{|y_k - y^*|}{|y_{k-1} - y^*|^p} = \lambda,$$

the sequence $\{y_k\}$ is called converging to y^* with order p. The number λ is called the asymptotic error constant.

Sublinear convergence	$p = 1, \lambda = 1$	$y_k = 1/k$
Linear convergence	$p = 1, \lambda < 1$	$y_k = 2^{-k}$
Superlinear convergence	$p > 1$	$y_k = k^{-k}$
Quadratic convergence	$p = 2$	$y_k = 2^{-2^k}$

Stopping criteria LVF pp.25

- In practice, we do not know y^*.

- For equations, $f(x) = b$, we can measure the *residual*

$$|f(x_k) - b|$$

- Some other stopping criteria (none guarantees convergence)
  - $|y_k - y_{k-1}|/|y_k| <$ tol
  - $k >$ maxIter
  - $|x_k - x_{k-1}| <$ tol *

*maxIter and tol are pre-specified constants.