CS371 course note

Chenxuan Wei

Jan 2022

Contents

1 Floating point

1.1 Source of Error

1. Errors in input
   - Measurement error
   - rounding error
2. Error as a result of calculation
   - Truncation error: taylor series
   - Rounding error in elementary steps of algorithm
3. definition 1,1 Round error
   - Absolute error = $|x - \overline{x}|$ where $\overline{x}$ is an approximation of x
   - Relative Error = $\frac{|x-\overline{x}|}{|x|}$
4. Truncation Error: Taylor series
   $R_n(x) = \frac{f^{(n+1)}(E_n(x))}{(n+1)!}(x-a)^{n+1}$

1.2 Floating Point numbers and Operation

1. Floating point representation

   - Base: base of the number system b_f
   - The mantissa: contains the normalized value of the number m_f
   - exponent: which defines the offset from normalization, e_f

   F = $0.x_1 \dots x_m \times b^{y_1 \cdots y_e}$
   whrer $1 \leq x_1 \leq b-1$, $0 \leq x_i \leq b-1$

2. Compare to fixed point

   - Fixed point
     - Values are evenly spaced
     - Really small or large value can't be represented
     - To represent real number, choose the cloest computer value
   - Floating
     - not evenly spaced, smaller value are closer
     - Greater range value can be represente
     - Required to rounding

3. how to represent number in a certainsystem Assume $x = 0.x_1 \dots x_n$, F[b, m, e]

   (a) Normallizaed x to make it like $0.x_1 \dots x_n$ where $x_1 \neq 0$
   (b) then $x = x \times b^{a=thenumberyoumoveyourdicimalpoint}$
   (c) and $\overline{x} = 0.x_1 \dots x_m \times b^{a_b withedigit}$

   the x we get called $fl(x)$

4. New terms

   - chopping: remove all digits > m
   - rounding: round the digites

5. Single precision numbers
   F[b = 2, m = 23, e = 7]
   $s_m b_1 \dots b_n s_e e_1 \dots e_m$
   where s_i are sign bits
   n = 23, m = 7

6. Double Precision Numbers
   F[b = 2, m = 52, e = 10]

7. Machine epsilon

   - Definition
     ϵ_{mach} is the smallest number $e > 0$ such $fl(1 + e) > 1$
   - Proposition
     $\epsilon_{mach} = b^{1-m}$ if chopping is used
     $\epsilon_{mach} = \frac{1}{2}b^{1-m}$ if rounding is used
   - Theorem
     For any floating point system, under chopping
     $|\delta_x| = |\frac{x-fl(x)}{x}| \leq \epsilon_{mach}$
     hence for
     single precision $|\delta_x| \leq 0.24 \times 10^{-6} \rightarrow 6$ or 7 digits accuracy
     double precision $|\delta_x| \leq 0.24 \times 10^{-15} \rightarrow 15$ to 16 digits of accuracy

8. Floating Point Operations

   - Addition
     $a \oplus b = fl(fl(a) + fl(b))$
   - Proposition
     $a \oplus b = (fl(a) + fl(b))(1 + n)$
     where $|n| \leq \epsilon_{mach}$
     also $= (a(1 + n_1) + b(1 + n_2))(1 + n)$
     and this operation is not associative

1.3 Condition of a Mathematical Problem

1. well-conditioned
   We say a problem P is well0conditioned with respected to the absolute error
   if small changes $\delta\vec{x}$ in $\vec{x}$ result in small changes $\delta\vec{z}$ in $\vec{z}$
   we say P is ill-conditioned if result in large changes of z

2. condition number

   - absolute
     $\kappa_A = \frac{||\delta\vec{z}||}{||\delta\vec{x}||}$
   - relative
     $\kappa_R = \frac{\frac{||\delta\vec{z}||}{||\vec{v}||}}{\frac{||\delta\vec{x}||}{||\vec{x}||}}$

   If they are between 0.1 and 10, we consider them to be small $\rightarrow$ well-conditioned

3. Vector Norms

   - Definition 1.7
     Let V be vector space, then $||*||$ si a vector norm on V $\iff$
     - $||\vec{v}|| = 0 \iff \vec{v} = \vec{0}$
     - $||\lambda\vec{v}|| = |\lambda| * ||\vec{v}|| \forall\vec{v} \in V, \forall\lambda \in R$
     - $||\vec{u} + \vec{v}|| \leq ||\vec{u}|| + ||\vec{v}|| \forall\vec{u}, \vec{v} \in V$
   - Definition 1.8: 2-norm
     $||\vec{x}||_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$
   - Definition 1.9: ∞-norm
     $||\vec{x}||_\infty = max_{1\leq i\leq n}(x_i)$
   - Definition 1.10: 1-norm
     $||\vec{x}||_1 = \sum_{i=1}^{n} |x_i|$
   - Theorem 1.2: Cauchy-Schwartz Inequality
     $|\vec{x} * \vec{y}| \leq ||\vec{x}||||\vec{y}||$

2 Root founding

2.1 Intro

1. Definition 2.1: double root
   we say x is double root of f(x) $\iff f(x) = 0$ and $f'(x) = 0$

2. Theorem 2.1 Intermidiate value theorem (IVT)
   if f(x) is countiuous on $[a, b]$ and $c \in [f(a), f(b)]$ then
   $\exists x* \in [a, b]$ such $f(x*) = c$

2.2 4 algorithms

1. Bisection Method

   - Theorem 2.2
     If f(x) is couninuous function on the interval $[a_0, b_0]$ such that $f(a_0) * f(b_0) \leq 0$
     then interval $[a_k, b_k]$ is defined as
     - $a_k =$
       * $a_{k-1} iff((a_{k-1} + b_{k-1})/2) * f(a_{k-1}) \leq 0$
       * $(a_{k-1} + b_{k-1})/2$ otherwise
     - $b_k =$
       * $a_{k-1} iff((a_{k-1} + b_{k-1})/2) * f(a_{k-1}) > 0$
       * $(a_{k-1} + b_{k-1})/2$ otherwise
   - Algorithm:
     Input: $f(x), [a, b], t$
     while $|b - a| > t$
     c = (a + b) /2
     if $f(a) * f(c) \leq 0$
     b = c
     else a = c
     return (a + b) / 2
   - Step to take n
     $n = \geq \frac{1}{log2} * log(\frac{|b-a|}{t})$

2. Fixed Point Interation

   - Definition
     We say $x*$ is a fixed point of g(x) if $g(x*) = x*$
   - Algorithm
     Input: g(x), x_0, t
     let i = 0
     repeat
     i = i+1, x[i] = g(x[i-1])
     until $|x[i] - x[i-1]| < t$
     return x[i]

3. Newton's method

- algorightm
  input f(x), f'(x), x_0, t
  i = 0, x[0] = x_0
  repeat
  i = i+1, if $f'(x[i-1]) \neq 0$, x[i] = x[i-1] - $\frac{f(x[i-1])}{f'(x[i-1])}$
  until $|x[i] - x[i-1]| < t$
  return x[i]

2.3 Rate of converge

1. error
   for sequence x_i, the error at iterations i is
   $e_i = x_i - x$

2. converges
   sequence x_i converges to x with order q $\iff$ x_i converges to x, $\lim_{i\to\infty} c_i = N$, and $|e_{i+1}| = c_i|e_i|^q$

2.4 Convergence Theory

1. contraction
   let g be a function, defined and continuous on bounded closed interval [a, b], g is contraction if
   $\exists L \in (0, 1)$ such that $|g(x) - g(y)| \leq L|x - y|, \forall x, y \in [a, b]$

2. Thoerem 2.3: contaction Mapping Thoerem
   if g is a contraction

   - g has a unique fixed point x in $[a, b]$
   - define $g(x_k) = x_{k+1}$, converges to x as $k \to \infty$ for any starting value

3. Corollary 2.1
   Assume $g(x) \in [a, b]$, $x \in [a, b] = g(x)$ be a fixed point, $g'(x)$ continuous on $i_\delta = [x - \delta, x + \delta]$
   Define sequence x_i by $x_{i+1} = g(x_i)$
   then

   - if $g(x) < 1$, then $\exists e$ such that x_i converge to x for $|x_0 - x| < e$
   - if $g(x) > 1$, then x_i diverges for any start value x_0

4. Theorem

3 Numerical Linear Algebra

3.1 Introduction

1. Theorem 3.1
   Existence and Uniqueness consider of $Ax = b$

   - $det(A) \neq 0 \iff x = A^{-1}b$ is unique solution of Ax = b
   - $det(A) = 0$ range(A) = column space of A
     - if $b \in range(A)$, then there are infinitely many solution
     - if $b \notin range(A)$, then there are no solution

3.2 Gaussian Elimination

1. Definition 3.1

   - upper-trianglar
     if $a_{ij} = 0, \forall i > j$
   - lower-triangular
     if $a_{ij} = 0 \forall i < j$

2. Inversion Property
   L_i can be obtained from M_i by swapping the signs of the off-diagonal elements

3. Combination Property
   $L = \prod_{i=1}^{n-1} L_i$

4. Definition 3.2
   L is cllaed a lower triangular matrix with unit diagonal $\iff$ L is defined like above with 1 on the diagonal

5. LU decomposition

   - get A_i and M_i, U by gassiun elimination
   - converge to L_i by inversion property
   - get L by combination property
   - check A = LU
   - Solve Ly = b
   - Solve Ux = y

6. Definition 3.3
   permutation matatrix is obtained from I_n by change some rows

7. Theorem 3.2
   there is always a P to make $PA = LU$

8. corollary 3.1
   if A is non singular, then Ax = b can be solve by apply PA = LU

9. Determinants
   $det(A) = \sum_{j=1}^{n}(-1)^{i+j} a_{ij} det(A_{ij})$

10. Propositions

    - det(BC) = det(B) $*$ det(C)
    - $U \in R^{n*n}$ upper t or lower t $\rightarrow$ det(U) = $\prod_{i=1}^{n} u_{ii}$
    - P = -1 if even row change, 1 if odd row changes

11. Proposition 3.2
    $det(A) \neq 0 \iff PA = LU, u_{ii} \neq 0 \forall i \in [1, n]$

3.3 Condition and stability

1. Definition 3.5
   $||A||_p = max_{||x|| \neq 0} \frac{||Ax||_p}{||x||_p}$

2. Proposition 3.3
   $||Ax||_p \leq ||A||_p ||x||_p$

3. proposition 3.4

   - $||A||_1 = max_{1 \leq j \leq n} \sum_{i=1}^{n} |a_{ij}|$
   - $||A||_\infty = max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}|$
   - $||A||_2 = max_{1 \leq i \leq n} \lambda_i^{\frac{1}{2}}$, where λ_i is eigenvalue of $A^T A$

4. Propostion 3.5
   $||A + B||_p \leq ||A||_p + ||B||_p$

5. Proposition 3.6

   - $||A||_p \geq 0, ||A||_p = 0 \iff A = 0$
   - $||aA||_p = |a| ||A||_p$
   - $||A + B||_p \leq ||A||_p + ||B||_p$

6. Definition 3.6
   Condition number of a matrix A is $\kappa(A) = ||A|| ||A^{-1}||$

3.4 Interative Methods for solving Ax = b

1. Sparse matrix
   A is that $\iff$ number of no zero elements in A is much smaller than n^2

2. strictly diagonally dominant
   A is that $\iff$ $|a_{ii}| > \sum_{j=1,j\neq i}^{n} |a_{ij}|$

3. Proposition 3.8
   the above one is always non-singular

4. Jacobi
   $x_i^{new} = \frac{1}{a_{ii}}(b_i - \sum_{j=1,j\neq i}^{n} a_{ij}x_j^{old})$
   $x^{new} = A_D^{-1}(b - (A_L + A_R)x^{old})$

5. Gauss-Seidel
   $x_i^{new} = \frac{1}{a_{ii}}(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{old} - \sum_{j=i+1}^{n} a_{ij}x_j^{old})$
   $x^{new} = A_D^{-1}(b - (A_L x^{new} + A_R x^{old}))$

4 Interpolation

4.1 Polynomial Interpolation

1. Interpolation polynomial $y_n(x)$
   Given $n + 1$ discrete data points (x_i, f_i), $i \in [0, n]$ with $x_i \neq x_j, i \neq j$
   the polynomial is the degree n polynomial:
   $y_n(x) = \sum_{i=1}^{n} a_i x^i$
   such $y_n(x_i) = f_i$

2. Determinant
   2x2 = ad-bc

3. Vandermonde Matrix

   - Definition V
     a (n+1) x (n+1) row looks like
     $1, x_i, x_i^2, \ldots, x_0^n$, where $i \in [0, n]$
   - determinant
     $det(V) = \prod_{0 \leq i \leq j \leq n}(x_j - x_i)$

4. Theorem 5.1
   $y_n(x)$ exists and is unique

5. Largrange Form

   - n+1 Largrange polynomials for set of point (x_i, f_i)
     that is $l_i(x_j)$ satisified
     - if $i = j$, 1
     - 0, otherwise

     which $l_i(x) = \prod_{j=0,j\neq i}^{n} \frac{x-x_j}{x_i-x_j}$
     with $y_n(x) = \sum_{i=0}^{n} l_i(x) f_i$
   - The lagrange Basis
     $P_n(x) = \{y_n(x) | y_n(x)$ is a polynomial of degree $\leq n\}$

6. Hermite Interpolation

   - Definition $y(x)$
     given $\{(x_i, f_i, f_i')\}$ the hermite interpolating polynomial is the polynomial $y(x)$ of degree 2n+1 which satisified
     $y(x_i) = f_i$, $y'(x_i) = f_i'$

4.2 Piecewise polynomial Interpolation

1. Spline Interpolation
   there are 4 condition, $y(x)$ is a degree k spline $\iff$

   - $y(x)$ is a piecewise polynomial of degree k in each intercal I_i
     define $y_i(x)$ as the restricion of y(x) to $[x_{i-1}, x_i]$
   - Interpolation condition
     $y_i(x_{i-1}) = f_{i-1}$, $y_i(x_i) = f_i$
   - smoothness condition
     $y_j^{(k)}(x_j) = y_{j+1}^{(k)}(x_j)$ for k-1 times, from first derivitive to k-1 derivitive
   - extra boundary condition

2. Extra for cubic spine

   - "free boundary"
     $y_1''(x_0) = 0, y_n''(x_n) = 0$, this is a natural cubic spline
   - "clamped boundary"
     $y_1'(x_0) = f_0', y_n'(x_n) = f_n'$
   - "periodic boundary"
     if $f_0 = f_n$
     $y_1'(x_0) = y_n'(x_n), y_1''(x_0) = y_n''(x_n)$

5 Integration

5.1 Intergration of Interpolation polynomial

1. midpoint rule (y(x) degree 0)
   $I_0 = \int_a^b f(\frac{a+b}{2})dx = (b-a)f(\frac{a+b}{2})$

2. Trapezoid rule (y(x) degree 1)
   $I_1 = (b-a)\frac{1}{2}[f(a) + f(b)]$

3. Simpson Rule: (y(x) degree 2)
   $I_2 = \frac{b-a}{6}(f_0 + 4f_1 + f_2)$

4. Error formula

   - midpoint e = $\frac{(b-a)^3}{24} f''(\xi_0)$, dp = 1
   - trapezoid e = $-\frac{(b-a)^3}{12} f''(\xi_1)$, dp = 1
   - Simpson e = $-\frac{(b-a)^5}{2880} f^{(4)}(\xi_2)$, dp = 3

5.2 Composite Integration

1. Composite Trapezoid rule
   $I_i = h\frac{f(x_{i-1})+f(x_i)}{2}$
   $I = \frac{h}{2}[f_0 + \sum_{i=1}^{n-1} 2f_i + f_n]$
   $T_{loc,i} = \sum_{i=1}^{n} I_i = -\frac{1}{12}(x_i - x_{i-1})^3 f''(\xi_i)$

2. Composite Simpson rule
   $I_i = \frac{h}{6}(f_{i-1} + 4f_{i-\frac{1}{2}} + f_i)$
   $I = \sum_{i=1}^{n} I_i$

3. theroem 6.2
   global truncation error for the simpson rule is $O(h^4)$

5.3 Gaussian Integration

1. Gaussian Integration
   $= \frac{b-a}{2}[f((\frac{b-a}{2})(-\frac{1}{\sqrt{3}}) + (\frac{b+a}{2})) + f((\frac{b-a}{2})(\frac{1}{\sqrt{3}}) + (\frac{b+a}{2}))]$

6 Discrete Fourier Methods

6.1 Introduction

1. Complex number
   $\sqrt{-1} = i$, $z = a + bi$

2. Terms

   - Complex conjugate
     $\overline{z} = a - ib$
   - Real part
     $Re(z) = a$
   - Imaginary part
     $Im(z) = b$
   - Modulus
     $r = |z| = \sqrt{a^2 + b^2}$
   - Phase angle
     $\theta = arctan(\frac{b}{a})$

3. Another from(Euler formulas)
   $z = r * e^{i\theta} = r * (cos(\theta) + i * sin(\theta))$

6.2 Fourier series

1. Fourise Series
   $g(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty}[a_k cos(k * \frac{2\pi x}{b-a}) + b_k sin(k * \frac{2\pi x}{b-a})]$
   $a_k = \frac{2}{b-a}\int_a^b f(x) * cos(k * \frac{2\pi x}{b-a})dx$
   $b_k = \frac{2}{b-a}\int_a^b f(x) * sin(k * \frac{2\pi x}{b-a})dx$

2. Proposition 4.1

   - f(t) even, $b_k = 0$
   - f(t) odd, $a_k = 0$

3. Theorem 4.1 Fundamental Convergence Theorem for Fourier Series
   $V = \{f(x)|\sqrt{\int_a^b f(x)dx} < \infty\}$
   then all $f(x) \in V$ there exists a_0, a_k, b_k such that
   $g(x)$ converge to $f(x)$ for $n \to \infty$ in the sense that $\sqrt{\int_a^b (f(x) - g(x))} = 0$

4. Complex Fourier Serise
   $h(t) = \sum_{k=-\infty}^{\infty} c_k e^{ikt}$
   $c_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(t)e^{-ikt}dt$

5. Relationship
   $c_k = \frac{1}{2}(a_k - ib_k)$

6. Proposition 4.2

   - $\overline{c_k} = c_{-k}$
   - $a_{-k} = a_k, b_{-k} = b_k$
   - $a_k = 2Re(c_k,\ b_k = -2Im(c_k)$
   - $b_0 = 0, c_0 = \frac{1}{2}a_0$

7. Theorem 4.2 $h(t) = g(t)$

6.3 Fourier Series and Orthogonal Basis

1. Basic definition

   - scalar produxt
     $\vec{x} * \vec{y} = x_1 y_1 + x_2 y_2 =< \vec{x}, \vec{y} >$

2. Orthogonal Basis
   Let $B = \{\vec{e_i}\}$ be the Orthogonal Basis $\iff$
   $< \vec{e_i}, \vec{e_j} >= c_{ij}$
   where c_{ij} is nonzero $\iff$ $i = j$

6.4 Discrete Fourier Transform

1. Nth root of unity
   $W_N^k = e^{\frac{2k\pi i}{N}}$
   with property $(W_N^k)^N = 1$
   $W_N^{N-k} = W_N^{-k}$

2. Direct trnasform
   $F[k] = \frac{1}{N}\sum_{n=0}^{N-1} f[n]W_N^{-kn}$

3. Inverse discrete Fourier transform
   $f[n] = \sum_{k=0}^{N-1} F[k]W_N^{kn}$

4. Fast transformation
   if $N = 2^m$, then

   - $g[n] = f[n] + f[n + N/2]$
   - $h[n] = f[n] + f[n + N/2]$
   - $F[2l] = \frac{1}{2}DFT\{g[n]\}$
   - $F[2l + 1] = \frac{1}{2}DFT\{h[n]\}$

   Have $O(N * log_2(N))$ runtime