CS 370

Numerical Computation

Zhisu Wang
Taught by Jeff Orchard
University of Waterloo
Spring 2023

Contents

1 Floating-Point Numbers

Norm Form: $\pm 0.\underbrace{d_1 d_2 \ldots}_{\text{mantissa}} \times \underbrace{\beta}_{\text{base}}{}^p$, p integer exponent.

Limits: Density (finite number of digits) and Range (finite number of integers for exponent)

FPNS: A four-tuple (t, β, L, U) such that any nonzero values have the form

$$0.d_1 d_2 \ldots d_t \times \beta^p$$

where $d_1 \neq 0, d_i \in \{0, \ldots, \beta - 1\}, L \leq p \leq U$.

Approximated Value: $\mu(x) = fl(x) = \bar{x}$, the rounded value of x in the given FPNS. $|x - fl(x)|$ is called the round-off error.

Absolute Error: $|x - \bar{x}|$ (round-off error), **Relative Error:** $\frac{|x-\bar{x}|}{|x|}$

Machine Epsilon: E, maximum relative error such that $|\delta| \leq E$ for any $\frac{\mu(x)-x}{x} = \delta$. The smallest number E in FPNS such that $\mu(1 + E) > 1$

Error Analysis: $\mu(x + y) = (x + y)(1 + \delta) = x \oplus y$

Cancellation Error: Round-off error when subtracting two large values with similar magnitude.

2 Linear Algebra

Google Page Rank: directed graph, where the chance of visiting each neighbor of v is $\frac{1}{\deg_{out}(v)}$. The probability matrix form is $P = \text{to} \overset{\text{from}}{\begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 1 \end{bmatrix}}$. R is # nodes and dim of P.

Terminal Page: If Randy visits a terminal page, the walk does not give good ranking of web page importance. The solution is Teleportation: jump to another random page.

Terminal Branch: If Randy loops in a cycle, it's hard to detect. The solution is at each page, jump to a random page with probability $(1-\alpha)$.

GPR Matrix: $M = \underbrace{\alpha P' + (1+\alpha)\frac{1}{R}ee^{\mathrm{T}}}_{\text{random jump}}$, $P' = \underbrace{P + \frac{1}{R}ed^{\mathrm{T}}}_{\text{teleport}}$. $d_i = 1$ if i is terminal else $= 0$.

Markov Transition Matrices: Q such that $0 \le Q_{ij} \le 1$, $\sum_j Q_{ij} = 1$

Steady State: $Mp = p \Rightarrow (I - M)p = 0$: find eigenvector of eigenvalue 1.

LU Factortization: $LU = PA$ (P is permutation matrix), $O(N^3)$ flops. Forward/Backward substitution $O(N^2)$. Example:

$$A = \begin{bmatrix} 2 & -1 & 3 \\ -4 & 6 & -5 \\ 6 & 13 & 16 \end{bmatrix} \underset{\substack{R3-\frac{6}{2}R1 \\ R2-\frac{-4}{2}R1}}{\rightarrow} \begin{bmatrix} 2 & -1 & 3 \\ 0 & 4 & 1 \\ 0 & 16 & 7 \end{bmatrix} \underset{R3-\frac{16}{4}R2}{\rightarrow} \begin{bmatrix} 2 & -1 & 3 \\ 0 & 4 & 1 \\ 0 & 0 & 3 \end{bmatrix} = U, \quad L = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 3 & 4 & 1 \end{bmatrix}$$

Induced Matrix Norm: $||A||_p = \max_{x \neq 0} \frac{||Ax||_p}{||x||_p} = \max_{||z||_p=1} ||Az||_p$, $||A||_\infty = \max_i(\sum_{j=1}^n |a_{ij}|)$, $||A||_1 = \max_j(\sum_{i=1}^n |a_{ij}|)$

Condition Number: $K(A) = ||A||||A^{-1}||$, $K(A) \ge 1$ since $1 = ||AA^{-1}||$. Also, $K_2(A) = ||A||_2||A^{-1}||_2 = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)} \overset{A^{\mathrm{T}}=A}{=} \frac{|\lambda_{\max}(A)|}{|\lambda_{\min}(A)|}$.

Pivoting: If pivot close to zero, calculation unstable. So find and swap row with largest $a_{jk}, j = k, \ldots, N$. Example with pivoting:

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 4 & 16 & 64 \\ 2 & 2 & 8 \end{bmatrix} \underset{R1\leftrightarrow R2}{\overset{\text{modify } P}{\rightarrow}} \begin{bmatrix} 4 & 16 & 64 \\ 1 & 1 & 1 \\ 2 & 2 & 8 \end{bmatrix} \underset{\substack{R2-\frac{1}{4}R1 \\ R3-\frac{2}{4}R1}}{\overset{\text{modify } L}{\rightarrow}} \begin{bmatrix} 4 & 16 & 64 \\ 0 & -3 & -15 \\ 0 & -6 & -24 \end{bmatrix} \underset{R2\leftrightarrow R3}{\overset{\text{modify } P}{\rightarrow}} \begin{bmatrix} 4 & 16 & 64 \\ 0 & -6 & -24 \\ 0 & -3 & -15 \end{bmatrix}$$

$$\underset{R3-\frac{-3}{-6}R2}{\overset{\text{modify } L}{\rightarrow}} \begin{bmatrix} 4 & 16 & 64 \\ 0 & -6 & -24 \\ 0 & 0 & -3 \end{bmatrix}, \quad \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{4} & 1 & 0 \\ \frac{1}{2} & \frac{1}{2} & 1 \end{bmatrix}}_{L} \underbrace{\begin{bmatrix} 4 & 16 & 64 \\ 0 & -6 & -24 \\ 0 & 0 & -3 \end{bmatrix}}_{U} = \underbrace{\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}}_{P} \underbrace{\begin{bmatrix} 1 & 1 & 1 \\ 4 & 16 & 64 \\ 2 & 2 & 8 \end{bmatrix}}_{A}$$

3 Interpolation

3.1 Monomial Form: Vandermonde System

Let $p(x)$ be written

$$p(x) = c_1 + c_2 x + c_3 x^2 + \dots c_n x^{n-1}$$

The **vandermode system** is

$$\begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{n-1} \\ \vdots & & & & \\ 1 & x_n & x_n^2 & \cdots & x_n^{n-1} \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

The simplified notation is $X\vec{c} = \vec{y}$.
Disadvantages of Monial Form:

1. Need to solve a linear system

2. The matrix entries become large as n gets bigger

   (a) matrix X becomes nearly singular (not invertible)

   (b) difficult to solve accurately

3.2 Lagrange Form

For (x_i, y_i), $i = 1, \dots, n$. $L_i(x) = 0$ at x_j, $j \neq i$.

$$L_i(x) = \frac{(x - x_1)(x - x_2)\cdots(x - x_{i-1})(x - x_{i+1})\cdots(x - x_n)}{(x_i - x_1)(x_i - x_2)\cdots(x_i - x_{i-1})(x_i - x_{i+1})\cdots(x - x_n)}$$

Disadvantages of polynomial interpolation:

1. The interpolant often does not follow the data in a ”reasonable” way.

2. The Vandermonde matrix can beome difficult to invert as you include more points

3. Especially if two points have similar x-values.

This is not an industry standard.

3.3 Piecewise Polynomial Interpolation

The entire piecewise polynomial is denoted

$$S(x) = \begin{cases} p_1(x) \text{ for } x_1 \leq x < x_2 \\ p_2(x) \text{ for } x_2 \leq x < x_3 \\ \vdots \\ p_{n-1}(x) \text{ for } x_{n-1} \leq x \leq x_{x_n} \end{cases}$$

A piecewise polynomial interpolator is

1. An **interpolant** $\implies S(x_i) = y_i,\ i = 1, \ldots, n.$
2. A **polynomial** on each subinterval $[x_i, x_{i+1}]$
3. **continuous** on the whole interval $[x_1, x_n]$

By definition it is not differentiable (smooth). We need to use higher order polynomials to achieve smoothness.

3.4 Cubic Spline Interpolation

$S(x)$ is called a **cubic spline** if

1. $S(x)$ is an interpolant
2. $S(x)$ is piecewise cubic
3. $S(x)$ is twice differentiable ($S'(x)$ and $S''(x)$ are both continuous on (x_1, x_n))

There are 3 constraints in order for a function to be cubic split

1. Interpolant constraint:
$$p_i(x_i) = y_i$$
$$p_i(x_{i+1}) = y_{i+1}$$
for $i = 1, \ldots, n-1$.

2. Differentiability constraint:
$$p_i'(x_{i+1}) = p_{i+1}'(x_{i+1})$$
for $i = 1, \ldots, n-2$.

3. Twice Differentiability constraint

$$p_i''(x_{i+1}) = p_{i+1}''(x_{i+1})$$

   for $i = 1, \dots, n-2$.

3.5 Alternative Representation of Cubic Spline

Monomial form is

$$p_i(x) = c_1^{(i)} + c_2^{(i)}x + c_3^{(i)}x^2 + c_4^{(i)}x^3$$

Instead, we'll use

$$p_i(x) = a_i \frac{(x_{i+1}-x)^3}{6h_i} + a_{i+1}\frac{(x-x_i)^3}{6h_i} + b_i(x_{i+1}-x) - c_i(x-x_i)$$

where $h_i = x_{i+1} - x_i$, $i = 1, \dots, n-1$.

$$p_i'(x) = -a_i \frac{(x_{i+1}-x)^2}{2h_i} + a_{i+1}\frac{(x-x_i)^2}{2h_i} - b_i + c_i$$

$$p_i''(x) = a_i \frac{x_{i+1}-x}{h_i} + a_{i+1}\frac{x-x_i}{h_i}$$

where

$$b_i = \frac{y_i}{h_i} - a_i\frac{h_i}{6}$$

$$c_i = \frac{y_{i+1}}{h_i} + a_{i+1}\frac{h_i}{6}$$

We use differentiable constraint to solve $a's$ (we can add 2 more constraints to get a unique solution), and thus gives us all b and c.

4 Least Squares

The model for Least Squares problems can be written as

$$y = \begin{bmatrix} a_1 \,\big|\, a_2 \,\big|\, \cdots \,\big|\, a_5 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} + r \Leftrightarrow y = A\beta + r$$

Want $\hat{\beta}$ so $y - A\hat{\beta}$ is as small as possible.

LS Problem: $y = \begin{bmatrix} a_1 \,\big|\, a_2 \,\big|\, \cdots \,\big|\, a_m \end{bmatrix} \beta + r = A\beta + r$
Residual: $r = y - A\beta$
Total Square Error: $E(\beta) = \sum_{i=1}^{n} r_i^2 = r^{\mathrm{T}} r$
Minimizing: $\frac{\partial E}{\partial \beta_i}(\beta) = \lim_{e_i \to 0} \frac{E(\beta + e_i) - E(\beta)}{||e_i||}, e_i = [0, \ldots, \delta, \ldots, 0]^{\mathrm{T}}$
Normal Equations: $A^{\mathrm{T}} A \beta = A^{\mathrm{T}} y$
Solution: $\beta = A^{+}$, where $A^{+} = (A^{\mathrm{T}} A)^{-1} A^{\mathrm{T}}$