exam. Fri Nov. 10 16:00 - 17:20 (extendable)
open book / electronic.
5 Questions. (normal questions. → 20 pts./
3 bonus → 10 pts.
strictly hard
QR decomp.

ZJU-UIUC INSTITUTE
Zhejiang University/University of Illinois at Urbana-Champaign Institute

CS450: Numerical Analysis

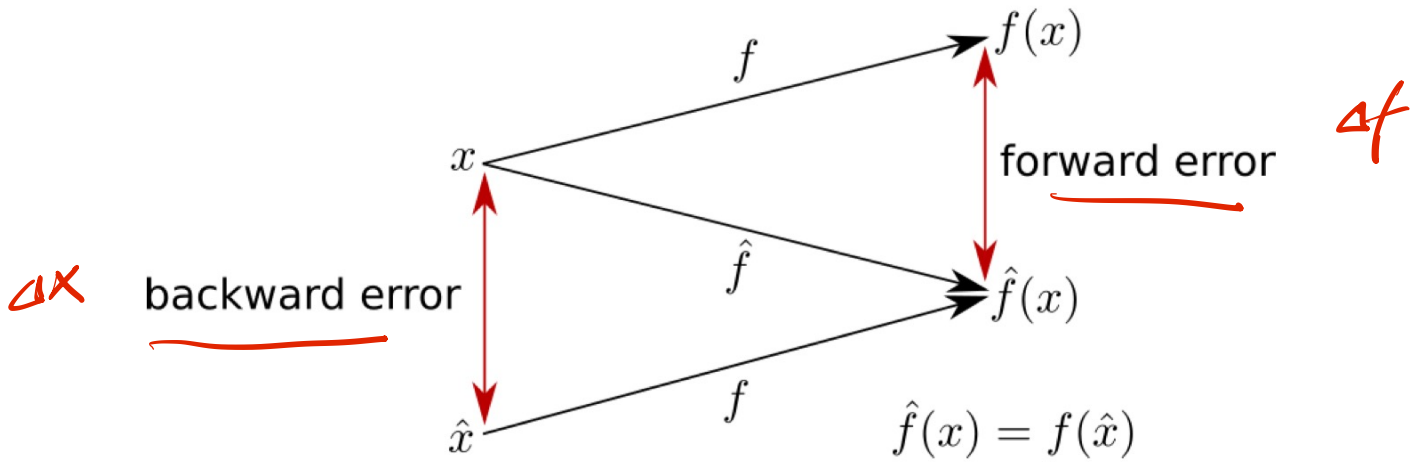
Howard Hao Yang

Assistant Professor, ZJU-UIUC Institute

08/11/2023

Approximation and Errors

- Forward and backward error
  - Suppose we want to compute $f(x)$, where $f: R \to R$, but obtain an approximate value $\hat{f}(x) = f(\hat{x})$
  - Forward error: $\Delta f = \hat{f}(x) - f(x)$
  - Backward error: $\Delta x = \hat{x} - x$



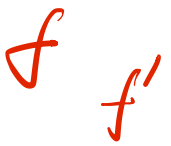
Approximation and Errors

- Forward and backward error
  - What is forward error?
    - The computational error of an algorithm/machine
    - Essentially, this is what we really want for an algorithm, but usually hard to obtain…
  - What is backward error?
    - Backward error enables us to measure computational error with respect to data propagation error

Conditioning

- Absolute condition number
  - The absolute condition number is a property of the problem, measuring its sensitivity to perturbations in input
  - How much a small change in the input leads to changes in the output
  - Formally, defined by the ratio of absolute errors at output and input

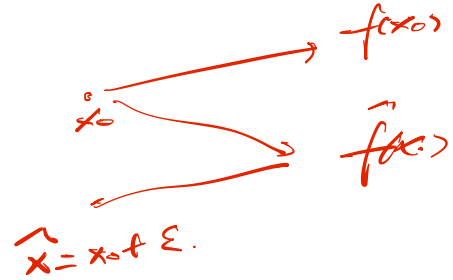
$$\kappa_{abs}(f) = \lim_{\delta \to 0} \max_{\|\Delta x\| < \delta} \frac{\|\Delta f\|}{\|\Delta x\|}$$



  - Linking the forward and backward error

$$\text{forward\_error} = \kappa(f) \times \text{backward\_error}$$

Condition Number



- Working example
  - Suppose we want to evaluate $f(x_0) = e^{\sqrt{x_0}}$ at $x_0 = 4$, but computation results in $\hat{f}(x_0) = f(x_0 + \epsilon)$ where $\epsilon = 10^{-4}$. Give a bound to the forward error.

$$|\Delta f| = |f(x_0+\varepsilon) - f(x_0)| \le K(f) \cdot |x_0 - \hat{x}|$$

$$= f'(x_0) \cdot \varepsilon \qquad x_0 = 4$$

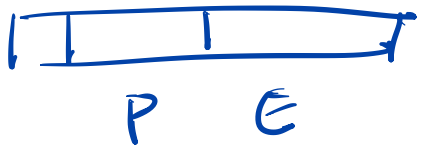
$$= \frac{1}{2} \cdot \frac{1}{\sqrt{x_0}} \cdot e^{\sqrt{x_0}} \cdot \varepsilon$$

$$= \frac{1}{2} \cdot \frac{1}{2} \cdot e^2 \times 10^{-4} \qquad e < 3$$

$$\le 2.3 \times 10^{-4}$$

Floating Point Numbers (cont'd)

- Floating point number system is characterized by four integers
  - Base/radix: β
  - Precision: p
  - Exponent range: $[L, U]$



- Number x is represented as

$$x = \pm\left(d_0 + \frac{d_1}{\beta} + \frac{d_2}{\beta^2} + \cdots + \frac{d_{p-1}}{\beta^{p-1}}\right) \times \beta^E$$

where $0 \le d_1 \le \beta - 1$, $i = 0, \ldots, p-1$, and $L \le E \le U$

  - Sign, exponent, and mantissa are stored in separate fixed-width fields

Rounding

- Not all real numbers can be represented exactly by floating-point systems
- The numbers representable are called machine numbers
- The numbers not representable are approximated by "nearby" floating-point numbers
- Two commonly used rounding rules
  - Chop: truncate base-β expansion of x after $(p-1)$-th precision digit; also called round toward zero
  - Round to nearest: $\mathrm{fl}(x)$ is nearest floating-point number to x
- Round to nearest is most accurate and is default in IEEE systems

Machine Precision

- Definition 1: The machine precision or machine epsilon is the smallest number ϵ such that $\mathrm{fl}(1+\epsilon) > 1$, formally,

$$\epsilon_{\mathrm{mach}} = \mathrm{argmin}_{\epsilon>0}\{\, \mathrm{fl}(1+\epsilon) \neq 1 \,\}$$

  - Under FP system: $1 + 0.00000000000000000000000000000000000001 = 1$
  - Under FP system: $1 + 0.001 = 1.001 \neq 1$
  - Which is larger, the machine epsilon or underflow level?

example: $\beta=10$, $p=3$, $E=2$ $\quad 0.01 \sim 9.99\times10^{2}$ ($\times10^{-2}$)

$1.00 \to 1.00\times10^{0}$
$0.00\beta \to 0.0\beta\times10^{-1}$ $\}\to$

$1.00 + 0.00\beta = 1.00\times10^{0} + 0.0\beta\times10^{-1}$
$= 1.00\times10^{0} + 0.00\beta\times10^{0}$
$= 1.00\times10^{0}$

Floating-Point Arithmetic

- Rule: When adding two numbers, the exponents align before mantissa adds

$$2.414 \times 10^3 + 1.2 \times 10^2 = 2.414 \times 10^3 + 0.12 \times 10^3 = 2.534 \times 10^3$$

- Floating-point addition and multiplication are commutative but not associative
  - If ϵ is slightly smaller than the machine precision ϵ_{mach}, then $(1+\epsilon)+\epsilon = 1$, but $1 + (\epsilon + \epsilon) > 1$
  - Q: How to add up $1 + \frac{1}{2}\epsilon_{\text{mach}} + \frac{1}{4}\epsilon_{\text{mach}} + \cdots + \frac{1}{2^n}\epsilon_{\text{mach}}$?

order matters

$\frac{1}{2^{n-1}}\epsilon_m + \frac{1}{2^n}\epsilon_m$

$n \to \infty$

$\frac{1}{2}\epsilon_{\text{mach}} + \frac{1}{4}\epsilon_{\text{mach}} = \frac{3}{4}\epsilon_{\text{mach}}$

$\approx \frac{1}{2} \cdot \frac{1}{1-\frac{1}{2}} \cdot \epsilon_{\text{mach}}$

Vector Norms

- What is a (vector) norm?
- A metric to measure the "length" of a vector, or "distance" between two vectors
  - An operator that $\|\cdot\|: R^d \rightarrow R_+$, and satisfies
    - $\|\boldsymbol{x}\| \geq 0$ and $\|\boldsymbol{x}\| = 0$ if and only if $\boldsymbol{x} = \mathbf{0}$
    - $\|a \cdot \boldsymbol{x}\| = |a| \cdot \|\boldsymbol{x}\|$ for any $a \in R$
    - $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$

Vector Norms (cont'd)

- Typical norms for vectors
  - The L-p norms
    - L-p norm: $\|\boldsymbol{\beta}\|_p = \left(\sum_{j=1}^{d} \beta_j^p\right)^{1/p}$, $p \geq 1$
    - $p = 1$, $\|\boldsymbol{\beta}\|_1 = \sum_j |\beta_j|$
    - $p = 2$, $\|\boldsymbol{\beta}\|_2 = \sqrt{\sum_j |\beta_j|^2}$
    - $p = \infty$, $\|\boldsymbol{\beta}\|_\infty = \max_j |\beta_j|$
    - The L-0 norm: Counts the number of non-zero entries, e.g., if $\boldsymbol{\beta} = (10, 0, 2, 0.01, 0, 1)^T$, then $\|\boldsymbol{\beta}\|_0 = 4$

$\|\beta\|_p = (|\beta_1|^p + |\beta_2|^p + \cdots + |\beta_d|^p)^{1/p}$

$\leq (d \cdot (\max_j |\beta_j|)^p)^{1/p}$

$\leq d^{1/p} \cdot \max_j |\beta_j|$

$p \to \infty$

$d^{1/p} \to 1$

$p \to \infty$

Matrix Norms

- Given matrix $\boldsymbol{A}$ (square)
  - We want an operator that satisfies
    1. $\|\boldsymbol{A}\| \geq 0$ and $\|\boldsymbol{A}\| = 0$ if and only if $\boldsymbol{A} = \boldsymbol{0}$
    2. $\|a \cdot \boldsymbol{A}\| = |a| \cdot \|\boldsymbol{A}\|$ for any $a \in R$
    3. $\|\boldsymbol{A} + \boldsymbol{B}\| \leq \|\boldsymbol{A}\| + \|\boldsymbol{B}\|$
    4. $\|\boldsymbol{AB}\| \leq \|\boldsymbol{A}\| \cdot \|\boldsymbol{B}\|$ (a property in matrix norms)

$\|A\|_F = \left\| \begin{matrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{matrix} \right\|_2$

$\|A\|_{max} = \max_{ij} |a_{ij}|$

If $A = B = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$, $\|A\|_{max} = \|B\|_{max} = 1$

$AB = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}$ $\|AB\|_{max} = 2$

Matrix Norms (cont'd)

- Given a vector norm $\|\cdot\|$
  - The matrix norm induced by this vector norm is given as

$$\|\boldsymbol{A}\| = \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \max_{\|x\|=1} \|Ax\|$$

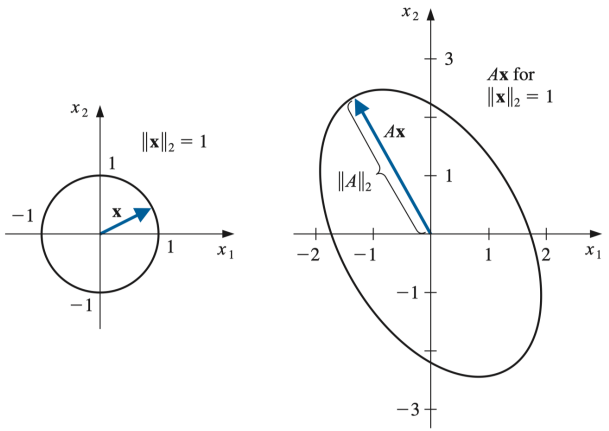
    - Different vector norms can induce different matrix norms for the same matrix
  - Examples
    - When the norm is L-1 norm, $\|\boldsymbol{A}\|_{\boldsymbol{1}} = \max_j \sum_{i=1}^{m} |a_{ij}|$: the maximum absolute column sum of the matrix
    - When the norm is L-∞ norm, $\|\boldsymbol{A}\|_{\infty} = \max_i \sum_{j=1}^{n} |a_{ij}|$: the maximum absolute row sum of the matrix

Matrix Norms (cont'd)

- When the norm is the Euclidean norm $\|\cdot\|_2$
  - The matrix norm induced by this vector norm is given as

$$\|\boldsymbol{A}\|_2 = \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} = \max_{\|\boldsymbol{x}\|_2=1} \|\boldsymbol{A}\boldsymbol{x}\|_2 \sim \lambda_{max}(A)$$

    - This is called the spectral norm of matrix $\boldsymbol{A}$



(Matrix) Norms

- Working example
  - If $\boldsymbol{A}$ is a square matrix, show that $\|\boldsymbol{A}\|_1 = \|\boldsymbol{A}^T\|_\infty$ and $\|\boldsymbol{A}\|_2^2 \leq \|\boldsymbol{A}\|_1 \|\boldsymbol{A}\|_\infty$?

$\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2 \longleftrightarrow \max\ x^TA^TAx \longrightarrow \mathcal{L}(x,\lambda)$

$\text{s.t.}\ x^Tx=1 \qquad = x^TA^TAx - \lambda(x^Tx-1)$

$\frac{\partial \mathcal{L}(x,\lambda)}{\partial x} = 0 \Rightarrow A^TAx - \lambda x = 0 \Rightarrow A^TAx = \lambda x \longrightarrow \lambda = \lambda_{max}(A^TA) = \|A\|_2^2$

$x^* = \arg\max_{\|x\|_2=1} \|Ax\|_2$

$A^TAx^* = \|A\|_2^2 \cdot x^* \Rightarrow \|\|A\|_2^2 \cdot x^*\|_1 = \|A^TAx^*\|_1$

$\leq \|A^T\|_1 \cdot \|Ax^*\|_1$

$\leq \|A\|_\infty \cdot \|A\|_1 \cdot \|x^*\|_1$

$\|A\|_2^2 \cdot \|x^*\|_1$

Matrix Condition Number

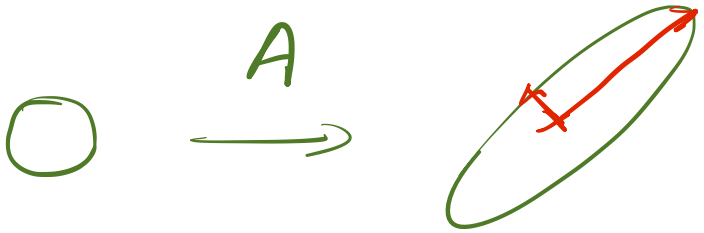
- Given a matrix $\boldsymbol{A}$, the condition number is

$$\kappa(\boldsymbol{A}) = \|\boldsymbol{A}\| \cdot \|\boldsymbol{A}^{-1}\|$$

- Definition: a $n \times n$ matrix $\boldsymbol{A}$ is said to be nonsingular/invertible if $\boldsymbol{A}^{-1}$ exists such that $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$; otherwise, it is singular
- Large $\kappa(\boldsymbol{A})$ implies the matrix is nearly singular

(Matrix) Condition Number

- Working example



  - If $\boldsymbol{A}$ is a square matrix, show that $\kappa(\boldsymbol{A}) = \|\boldsymbol{A}\| \cdot \|\boldsymbol{A^{-1}}\| = \left(\max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{Ax}\|}{\|\boldsymbol{x}\|}\right)\left(\min_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{Ax}\|}{\|\boldsymbol{x}\|}\right)^{-1}$

$$\|A^{-1}\| = \max_{x\neq 0} \frac{\|A^{-1}x\|}{\|x\|} \qquad z = A^{-1}x \iff x = Az$$

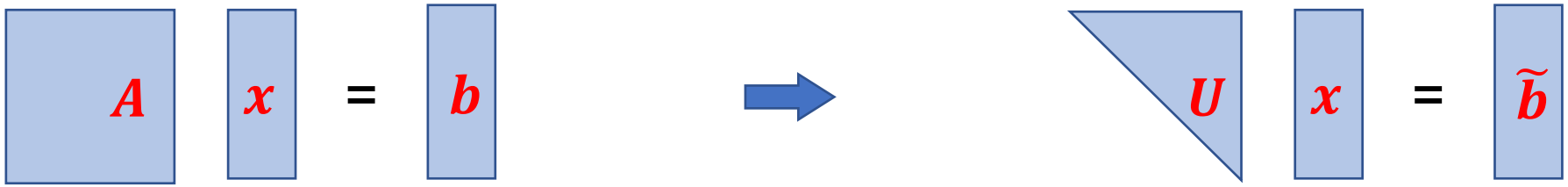
$$= \max_{z\neq 0} \frac{\|z\|}{\|Az\|}$$

$$= \frac{1}{\min_{x\neq 0} \frac{\|Ax\|}{\|x\|}}$$

$$\kappa(A) = \frac{\max_{\|x\|=1} \|Ax\|}{\min_{\|x\|=1} \|Ax\|}$$

Solving Linear Systems

- For a square matrix $\boldsymbol{A}$, how to (systematically) solve for $\boldsymbol{Ax} = \boldsymbol{b}$?
  - Transform it into one whose solution is the same but easier to compute
  - Specifically, eliminate x_1 from $n-1$ equations to get a smaller system $\boldsymbol{A_2 x} = \boldsymbol{b_2}$ of size $n-1$
  - Eventually, reaching the 1 by 1 system $A_n x_n = b_n$ which we know $x_n = b_n / A_n$
  - Working backwards produces x_{n-1}, x_{n-2}, …, and eventually x_2 and x_1



Triangular Linear Systems

- What type of linear system is easy to solve?
  - Systems that form triangular matrices

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= a_{1,n+1} \\ a_{22}x_2 + \cdots + a_{2n}x_n &= a_{2,n+1} \\ \ddots \quad \vdots \qquad &\quad \vdots \\ a_{nn}x_n &= a_{n,n+1} \end{aligned}$$

  - Back-substitution

$$x_n = \frac{a_{n,n+1}}{a_{nn}},$$

$$x_i = \frac{a_{i,n+1} - \sum_{j=i+1}^{n} a_{ij}x_j}{a_{ii}}, \; i = n-1, \ldots, 1$$

Elementary Elimination Matrices

- More generally, we can annihilate all entries below k-th in a n-dimensional vector $\boldsymbol{a}$ by transformation

$$\boldsymbol{M}_k \boldsymbol{a} = \begin{bmatrix} 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & -m_{k+1} & 1 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & -m_n & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_k \\ a_{k+1} \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} a_1 \\ \vdots \\ a_k \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

where $m_i = \frac{a_i}{a_k}$, $i = k+1, \ldots, n$

- Divisor a_k, called pivot, must be nonzero

Elementary Elimination Matrices (cont'd)

Example: For a vector $\boldsymbol{a} = [2, 4, -2]^T$, the elementary elimination matrices $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ are (recall: $m_i = \frac{a_i}{a_k}$, $i = k+1, \ldots, n$)

$$\boldsymbol{M}_1\boldsymbol{a} = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 4 \\ -2 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}$$

$$\boldsymbol{M}_2\boldsymbol{a} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1/2 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 4 \\ -2 \end{bmatrix} = \begin{bmatrix} 2 \\ 4 \\ 0 \end{bmatrix}$$

Elementary Elimination Matrices (cont'd)

- Matrix $\boldsymbol{M}_k$, called elementary elimination matrix, adds multiple of row k to each subsequent row, with multipliers m_i chosen so that result is zero
- $\boldsymbol{M}_k$ is unit lower triangular and nonsingular
- $\boldsymbol{M}_k = \boldsymbol{I} - \boldsymbol{m}_k \boldsymbol{e}_k^T$, where $\boldsymbol{m}_k = [0, \ldots, 0, m_{k+1}, \ldots, m_n]^T$ and $\boldsymbol{e}_k$ is the k-th column of identity matrix
- $\boldsymbol{M}_k^{-1} = \boldsymbol{I} + \boldsymbol{m}_k \boldsymbol{e}_k^T$, which means $\boldsymbol{M}_k^{-1} = \boldsymbol{L}_k$ is the same as $\boldsymbol{M}_k$ except signs of multipliers are reversed

Elementary Elimination Matrices (cont'd)

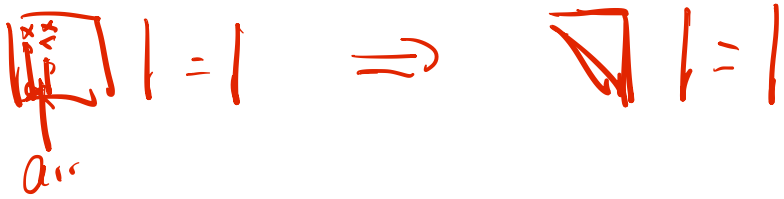
ZJU-UIUC INSTITUTE

- If $\boldsymbol{M}_j$, $j > k$, is another elementary elimination matrix, with vectors of multi pliers $\boldsymbol{m}_j$, then

$$\begin{aligned}\boldsymbol{M}_k\boldsymbol{M}_j &= \left(\boldsymbol{I} - \boldsymbol{m}_k\boldsymbol{e}_k^T\right)\left(\boldsymbol{I} - \boldsymbol{m}_j\boldsymbol{e}_j^T\right)\\ &= \boldsymbol{I} - \boldsymbol{m}_k\boldsymbol{e}_k^T - \boldsymbol{m}_j\boldsymbol{e}_j^T + \boldsymbol{m}_k\boldsymbol{e}_k^T\boldsymbol{m}_j\boldsymbol{e}_j^T\\ &= \boldsymbol{I} - \boldsymbol{m}_k\boldsymbol{e}_k^T - \boldsymbol{m}_j\boldsymbol{e}_j^T\end{aligned}$$

where means product is essentially "union", and similarly for product of inverses, $\boldsymbol{L}_k\boldsymbol{L}_j$

Gaussian Elimination



- To reduce a general linear system of equations $\boldsymbol{Ax} = \boldsymbol{b}$ into upper triangular form, do the following
  - Choose $\boldsymbol{M}_1$ with a_{11} as pivot, annihilate first column of $\boldsymbol{A}$ below first row: the system becomes $\boldsymbol{M}_1\boldsymbol{Ax} = \boldsymbol{M}_1\boldsymbol{b}$, but solution remains unchanged
  - Next choose $\boldsymbol{M}_2$ with a_{22} as pivot, annihilate second column of $\boldsymbol{M}_1\boldsymbol{A}$ below second row: the system becomes $\boldsymbol{M}_2\boldsymbol{M}_1\boldsymbol{Ax} = \boldsymbol{M}_2\boldsymbol{M}_1\boldsymbol{b}$, but the solution still unchanged
  - Process continues for each successive column until all subdiagonal entries have been zeroed, results in upper triangular linear system $\boldsymbol{M}_{n-1}\cdots\boldsymbol{M}_1\boldsymbol{Ax} = \boldsymbol{M}_{n-1}\cdots\boldsymbol{M}_1\boldsymbol{b}$, which can be solved via back-substitution
- This process is called Gaussian elimination

LU Factorization

- Denote by $\boldsymbol{M_k^{-1} = L_k}$, then $\boldsymbol{L = M^{-1} = M_1^{-1} \cdots M_{n-1}^{-1} = L_1 \cdots L_{n-1}}$ is lower triangular
- By design, $\boldsymbol{U = M_{n-1} \cdots M_1 A}$ is upper triangular
- Therefore, we can factorize matrix $\boldsymbol{A}$ into the product of a lower triangular matrix and an upper triangular matrix

$$\boldsymbol{A = LU}$$

- Thus, Gaussian elimination produces LU factorization of matrix into triangular factors
- Can every square matrix be factorized into an LU decomposition?

LU and Linear System of Equations

- Working example
  - How would you solve a partitioned linear system of the form

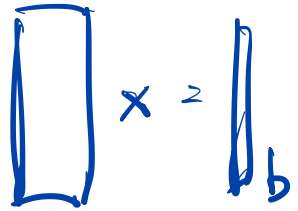
$$\begin{bmatrix} \boldsymbol{L}_1 & \boldsymbol{O} \\ \boldsymbol{B} & \boldsymbol{L}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} = \begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{c} \end{bmatrix}$$

  where $\boldsymbol{L}_1$ and $\boldsymbol{L}_2$ are nonsingular lower triangular matrices?

① $L_1 x = b$

② $L_2 y = c - Bx$

Linear Least Square



- Suppose we want to solve a linear system $\boldsymbol{Ax} = \boldsymbol{b}$, with $\boldsymbol{A}$ and $\boldsymbol{b}$ being known, what does this mean?
  - What does $\boldsymbol{Ax}$ mean?
  - What does $\boldsymbol{Ax} = \boldsymbol{b}$ mean?
- What happens if $\boldsymbol{A}$ is not a square matrix?
- Instead of solving $\boldsymbol{Ax} = \boldsymbol{b}$, we aim to find $\boldsymbol{x}$ to minimize $\|\boldsymbol{Ax} - \boldsymbol{b}\|_2$
  - Generally, a perfect fitting may not be possible, and we look for an approximation
  - In data science, this is called Linear Regression

Solving the Linear Least Square

- How to find $\boldsymbol{x}^* = \text{argmin}_{\boldsymbol{x}} \|\boldsymbol{Ax} - \boldsymbol{b}\|_2$ where $\boldsymbol{A} \in R^{m \times n}$?
  - The residual error can be written as

$$Err(\boldsymbol{x}) = (\boldsymbol{Ax} - \boldsymbol{b})^T (\boldsymbol{Ax} - \boldsymbol{b})$$

  - Let $\frac{\partial Err(\boldsymbol{x})}{\partial \boldsymbol{x}} = 2\boldsymbol{A}^T(\boldsymbol{Ax} - \boldsymbol{b}) = \boldsymbol{0}$, which is equivalent to $\left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\right)\boldsymbol{x} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{b}$
  - If $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ is nonsingular, we obtain the solution as

$$\boldsymbol{x} = \left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{b}$$

  - This is known as the normal equation

Singularity Issues

$Ax=b$

- Working example
  - Show that the following matrix $\boldsymbol{A}$ is singular; if $\boldsymbol{b} = [2\ 4\ 6]^T$, how many solutions are there in $\boldsymbol{Ax} = \boldsymbol{b}$?

① singularity

$a_2 = a_1 + a_3$

$$\boldsymbol{A} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ 1 & 3 & 2 \end{bmatrix} \quad x = \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix}$$

②

$$\alpha \begin{bmatrix} 0 \\ 2 \\ 0 \end{bmatrix} + (1-\alpha) \begin{bmatrix} 2 \\ 0 \\ 2 \end{bmatrix}$$

$\alpha \in [0, 1]$

Solving the Normal Equation

- How to solve $\boldsymbol{A^T A x = A^T b}$?
- Remember the LU factorization? Since $\boldsymbol{A^T A}$ is a square matrix, we can do LU factorization to it and then use back substitution?
- However, consider the following

$$\boldsymbol{A} = \begin{bmatrix} 1 & 1 \\ \epsilon & 0 \\ 0 & \epsilon \end{bmatrix}, \quad \text{where } 0 < \epsilon < \sqrt{\epsilon_{\text{mach}}}$$

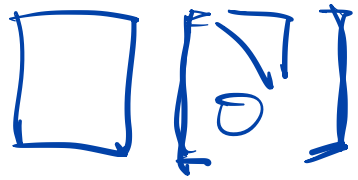
$$\rightarrow \quad \boldsymbol{A^T A} = \begin{bmatrix} 1+\epsilon^2 & 1 \\ 1 & 1+\epsilon^2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

- We want something more robust…

Wisdom from the Past

- When solving a linear system, we decompose the square matrix $\boldsymbol{A}$ into an LU form—we want something similar when $\boldsymbol{A}$ becomes a rectangular matrix
- For a rectangular matrix $\boldsymbol{A} \in R^{m \times n}$ with $m > n$, it can be decomposed as

$$\boldsymbol{A} = \boldsymbol{Q} \begin{bmatrix} \boldsymbol{R} \\ \boldsymbol{O} \end{bmatrix}$$



where $\boldsymbol{Q}$ is an $m \times m$ orthogonal matrix and $\boldsymbol{R}$ is an upper triangular matrix

- This is known as the QR factorization

Goodness of Triangles

- Working example
  - What is the Euclidean norm of the minimum residual vector for the following linear least square problem? What is the solution vector for this problem?

$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \cong \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}$$

$A \quad x \quad b$

$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

$(x_1+x_2-2)^2+(x_2-1)^2+1$

$=0$, when $x_1=x_2=1$

$$\|r\|_2^2 = \|b-Ax\|_2^2$$

$$= \left\| \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} x_1 + \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} x_2 - \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix} \right\|_2^2 = \left\| \begin{bmatrix} x_1+x_2-2 \\ x_2-1 \\ 1 \end{bmatrix} \right\|_2^2$$

QR Factorization for LSQ

- If we have a QR factorization of matrix $\boldsymbol{A} \in R^{m \times n}$, i.e.,

$$\boldsymbol{A} = \boldsymbol{Q} \begin{bmatrix} \boldsymbol{R} \\ \boldsymbol{O} \end{bmatrix}$$

- Then, the equation $\boldsymbol{Ax} = \boldsymbol{b}$ reduces to

$$\boldsymbol{Q}^T \boldsymbol{Ax} = \begin{bmatrix} \boldsymbol{R} \\ \boldsymbol{O} \end{bmatrix} \boldsymbol{x} \cong \begin{bmatrix} \boldsymbol{c}_1 \\ \boldsymbol{c}_2 \end{bmatrix} = \boldsymbol{Q}^T \boldsymbol{b}$$

- The following holds

$$\begin{aligned} \|\boldsymbol{Ax} - \boldsymbol{b}\|_2^2 &= \|\boldsymbol{Q}^T(\boldsymbol{Ax} - \boldsymbol{b})\|_2^2 \\ &= \|\boldsymbol{Rx} - \boldsymbol{c}_1\|_2^2 + \|\boldsymbol{c}_2\|_2^2 \end{aligned}$$

QR Factorization for LSQ (Cont'd)

- If we have a QR factorization of matrix $\boldsymbol{A}$, i.e., $\boldsymbol{A} = \boldsymbol{QR}$, the following holds

$$\|\boldsymbol{Ax} - \boldsymbol{b}\|_2^2 = \|\boldsymbol{Rx} - \boldsymbol{c_1}\|_2^2 + \|\boldsymbol{c_2}\|_2^2$$

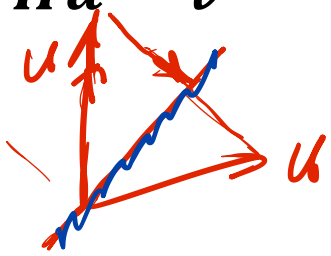
- The second term is the residual error, which we cannot do anything on it
- The first term, we can choose $\boldsymbol{x}$ to minimize it, and the solution is given by solving $\boldsymbol{Rx} - \boldsymbol{c_1} = \boldsymbol{0}$, which can be achieved via back substitution
- In this way, we avoid the cross-product matrix, saving us from round-off error issues
- The remaining question now is… how to perform QR factorization?

Householder Transformations (cont'd)

- Given a vector $\boldsymbol{x} \in R^n$ with $\boldsymbol{x}^T\boldsymbol{x} = 1$, the Householder transformation is

$$\boldsymbol{H} = \boldsymbol{I} - 2\boldsymbol{x}\boldsymbol{x}^T$$

- Reflection property: for any vector $\boldsymbol{a} \in R^n$, $\boldsymbol{H}\boldsymbol{a}$ reflects $\boldsymbol{a}$ by the hyperplane perpendicular to $\boldsymbol{x}$
- Let $n \geq 2$ and $\boldsymbol{u}, \boldsymbol{v} \in R^n$ be unit vectors (i.e., $\boldsymbol{u}^T\boldsymbol{u} = \boldsymbol{v}^T\boldsymbol{v} = 1$). Suppose $\boldsymbol{u} \neq \boldsymbol{v}$, let $\boldsymbol{x} = \frac{\boldsymbol{u}-\boldsymbol{v}}{\|\boldsymbol{u}-\boldsymbol{v}\|_2}$ and construct $\boldsymbol{H} = \boldsymbol{I} - 2\boldsymbol{x}\boldsymbol{x}^T$, then $\boldsymbol{H}\boldsymbol{u} = \boldsymbol{v}$

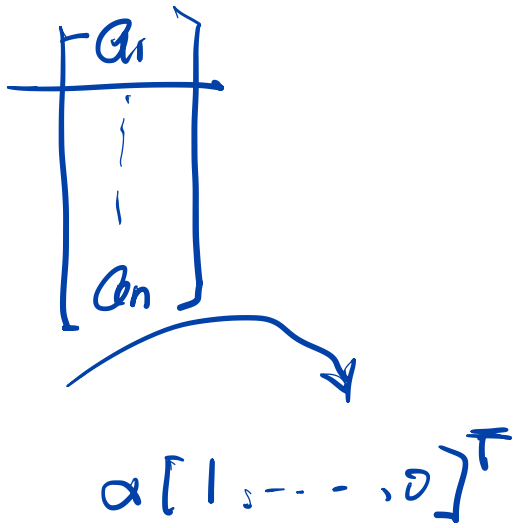


Householder Transformations (Cont'd)

- Suppose we have $\boldsymbol{a} \in R^n$, and want to annihilate all elements below the first entry while preserving the norm
- Can we leverage some ideas from the Householder transformation?
- Problem: find vector $\boldsymbol{x} \in R^n$, such that $\boldsymbol{x}^T\boldsymbol{x} = 1$ and

$$\boldsymbol{H}\boldsymbol{a} = (\boldsymbol{I} - 2\boldsymbol{x}\boldsymbol{x}^T)\boldsymbol{a} = \alpha\boldsymbol{e}_1$$

where $\boldsymbol{e}_1 = (1,0,\cdots,0)^T$ and $\alpha = \|\boldsymbol{a}\|_2$



- Solution to this is

$$\boldsymbol{x} = \frac{\boldsymbol{a}}{\alpha} \pm \boldsymbol{e}_1$$

Householder Transformation

- Working example
  - Can you identify an eigenvector of $\boldsymbol{H}$ that associates with the eigenvalue -1?; i.e., a vector such that $\boldsymbol{Hv} = -\boldsymbol{v}$?

$H = I - 2xx^T$

$Hx = (I - 2xx^T)x$

$= x - 2x\underbrace{x^Tx}_{=1} = x - 2x = -x$

$Hx = -x$

Householder Transformations (Cont'd)

- For a rectangular matrix $\boldsymbol{A} \in R^{m\times n}$ with $m > n$. Suppose $m = 6$, $\mathrm{n} = 5$ and we have computed the following

$$H_2 H_1 A = \begin{bmatrix} \times & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times \\ 0 & 0 & \times & \times & \times \\ 0 & 0 & \times & \times & \times \\ 0 & 0 & \times & \times & \times \\ 0 & 0 & \times & \times & \times \end{bmatrix}$$

- Concentrating on the highlighted entries, we determined a matrix $\tilde{H}_3 \in R^{4\times 4}$ such that

$$\tilde{H}_3 \begin{bmatrix} \times \\ \times \\ \times \\ \times \end{bmatrix} = \begin{bmatrix} \times \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Householder Transformations (Cont'd)

- Then, by forming $H_3 = \text{diag}(I_2, \widetilde{H}_3)$, we have

$$H_3 H_2 H_1 A = \begin{bmatrix} \times & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times \\ 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & \times & \times \\ 0 & 0 & 0 & \times & \times \\ 0 & 0 & 0 & \times & \times \end{bmatrix}$$

- More generally, for a given vector $\boldsymbol{a} = \begin{bmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \end{bmatrix}$, where a_1 is a (k-1)-vector, with $1 \leq k < m$. If we take the Householder vector to be $\boldsymbol{v} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{a}_2 \end{bmatrix} - \alpha \boldsymbol{e}_k$, where $\alpha = \mp \|\boldsymbol{a}_2\|_2$, then the resulting Householder transformation annihilates the last $m - k$ components of $\boldsymbol{a}$.

Householder Transformations (Cont'd)

ZJU-UIUC INSTITUTE
Zhejiang University/University of Illinois at Urbana-Champaign Institute

- For a rectangular matrix $\boldsymbol{A} \in R^{m \times n}$ with $m > n$. The QR factorization to this matrix can be written as

$$\boldsymbol{H}_n \cdots \boldsymbol{H}_1 \boldsymbol{A} = \begin{bmatrix} \boldsymbol{R} \\ \boldsymbol{O} \end{bmatrix}$$

- The product of successive Householder transformations $\boldsymbol{H_n} \cdots \boldsymbol{H_1}$ is itself an orthogonal matrix. Thus, if we take $\boldsymbol{Q^T} = \boldsymbol{H_n} \cdots \boldsymbol{H_1}$, or equivalently, $\boldsymbol{Q} = \boldsymbol{H_1} \cdots \boldsymbol{H_n}$, then

$$\boldsymbol{A} = \boldsymbol{Q} \begin{bmatrix} \boldsymbol{R} \\ \boldsymbol{O} \end{bmatrix}$$