Computations with Large Matrices

- Ax=b in its many variations
  - Ordinary elimination might compute an accurate x, or maybe not
  - Too many equations (m>n) and no solution → least^2
  - Square matrix might be singular A^{-1} ?
  - Solution might be impossible to compute (A is extremely ill-conditioned or simply too large)
  - In deep learning we have too many solutions: we want one that will generalize well to unseen test data
- Separate sources of difficulty
- Identify the problem
- Suggest a course of action

0 ~ 7

0 1 2 3 ~ 7

Good Problems

- Every matrix A has a pseudoinverse A^+
  - Inverse for every matrix, but this might not help
- Elimination will succeed (with row changes, A₩b)
  - Square and invertible, reasonable size, not large condition number
- (m>n=r) normal equations to find the least squares solution
  - columns of A are independent and not too ill-conditioned
  - b is probably not in the column space of A

$A^+ = A^{-1}$ if invertible

A\b

$Ax = b$

Difficult Problems

- (m<n) many solutions (underdetermined)
- Columns of A may be in bad condition → QR
  - Too large condition number
  - x is not well determined → orthogonalize the columns by a Gram-Schmidt or Householder algorithm
- A may be nearly singular
  - A^TA will have a very larger inverse, Gram-Schmit may fail
  - Different approach to add a penalty term → make A^TA more positive (common in inverse problem)
- A is way too big (no elimination)
  - Random sampling of the columns
  - Results are never certain, but the probability of going wrong is low

Numerical Linear Algebra

By Trefethen and Bau

$$\mathbf{Ax} = \mathbf{b}$$
$$\mathbf{Ax} = \lambda\mathbf{x}$$
$$\mathbf{Sq} = \lambda\mathbf{q}$$
$$\mathbf{Av} = \sigma\mathbf{u}$$

- Fundamentals
  - Reaching the SVD and Eckart-Young
- QR Factorization and Least Squares
  - All 3 ways: A^{+}, $(A^{T}A)^{-1}A^{T}$, QR
- Conditioning and Stability
  - Condition numbers, backward stability, perturbations
- Systems of Equations
  - Direct elimination: PA=LU, Cholesky's $S=A^{T}A$
- Eigenvalues ✓
  - Reduction to tridiagonal-Hessenberg-bidiagonal; QR with shifts
- Iterative Methods
  - Arnoldi, Lanczos, GMRES, conjugate gradients, Krylov

- Krylov Subspaces and Arnoldi Iteration
- Eigenvalues from Arnoldi
- Linear Systems by Arnoldi and GMRES
- Symmetric Matrices: Arnoldi becomes Lanczos
- Eigenvalues of Tridiagonal T by QR Iteration
- Computing the SVD
- Conjugate Gradient for Sx=b
- Preconditioning for Ax=b

Least Squares: Four Ways

- SVD of A leads to its pseudoinverse $A^+ \rightarrow \hat{\mathbf{x}} = \mathbf{A}^+\mathbf{b}$
- $\mathbf{A}^T\mathbf{A}\hat{\mathbf{x}} = \mathbf{A}^T\mathbf{b}$ can be solved directly when A has independent columns
- Gram-Schmidt idea produces orthogonal columns in Q → A=QR

$$\mathbf{A}^T\mathbf{A}\hat{\mathbf{x}} = \mathbf{A}^T\mathbf{b} \rightarrow (\mathbf{QR})^T(\mathbf{QR})\hat{\mathbf{x}} = (\mathbf{QR})^T\mathbf{b} \rightarrow \mathbf{R}^T(\mathbf{Q}^T\mathbf{Q})\mathbf{R}\hat{\mathbf{x}} = \mathbf{R}^T\mathbf{Q}^T\mathbf{b}$$

$$\rightarrow \mathbf{R}\hat{\mathbf{x}} = \mathbf{Q}^T\mathbf{b} \ \ (\text{safe to solve and fast})$$

- Minimize $\|\mathbf{b} - \mathbf{A}\mathbf{x}\|^2 + \delta^2\|\mathbf{x}\|^2$ → that penalty changes the normal equations to $(\mathbf{A}^T\mathbf{A} + \delta^2\mathbf{I})\mathbf{x}_\delta = \mathbf{A}^T\mathbf{b}$
  - Now the matrix is invertible and $\mathbf{x}_\delta$ goes to $\hat{\mathbf{x}}$ as $\delta \rightarrow 0$

A^TA and A^TCA

weight

covariance

- Samples in applied mathematics
  - Stiffness matrix in mechanical engineering
  - Conductance matrix in circuit theory
  - (weighted) graph Laplacian in graph theory
  - Gram matrix (inner products of columns of A) in mathematics
- Characteristics
  - Symmetry: attractive

A^TA

  - Size may be a problem
  - Condition number: square of the condition number of A

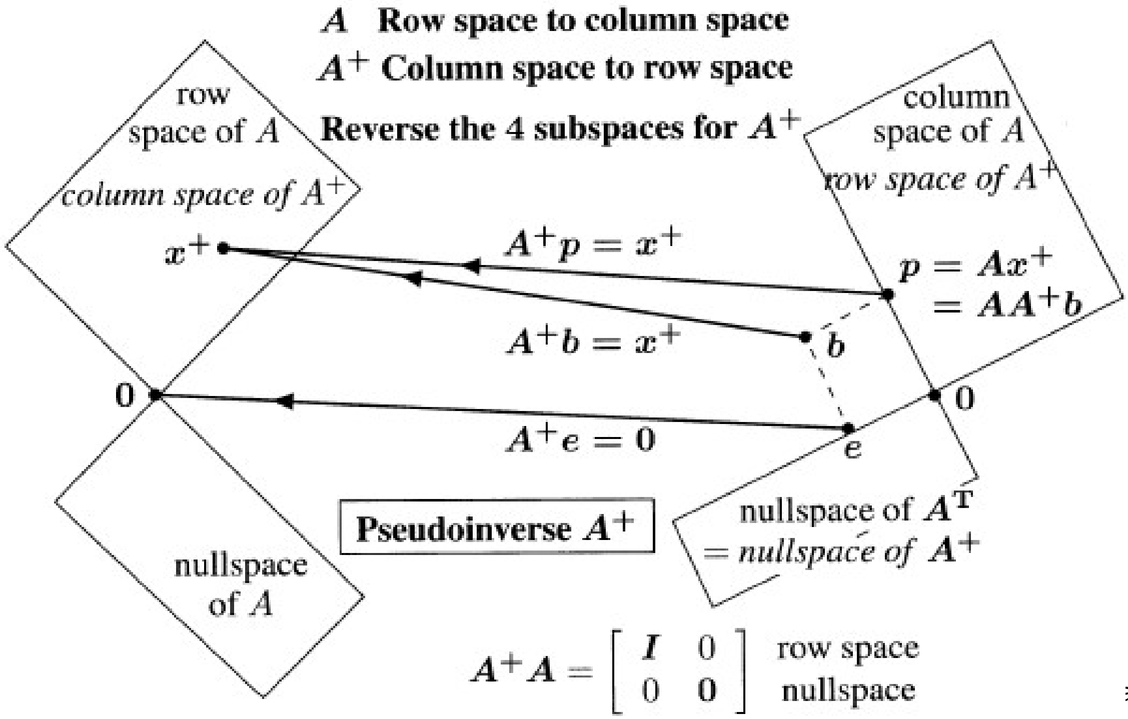
2

  - In large problems, expensive and often dangerous to compute
- We try not to compute them
  - Orthogonal matrices and triangular matrices are good ones

Q R

A^+ is the pseudoinverse of A

- Suitable "pseudoinverse" when A has no inverse
  - Rule 1: if A has independent columns, then $A^+=(A^TA)^{-1}A^T$
  - Rule 2: if A has independent rows, then $A^+=A^T(AA^T)^{-1}$
  - Rule 3: A diagonal matrix Σ is inverted where possible, otherwise, Σ^+ has zeros



Least Square Solutions to Ax=b is $x^+=A^+b$

- $x^+=A^+b$ is the minimum norm least square solution
  - $x= x^+=A^+b$ makes $\|b-Ax\|^2$ as small as possible
  - If another $\underline{x}$ achieves that minimum then $\|x^+\|<\|\underline{x}\|$

$$\text{Example: What is the shortest least squares solution to } \underbrace{\begin{bmatrix} 3 & 0 \\ 0 & 0 \end{bmatrix}}_{\Sigma}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 6 \\ 8 \end{bmatrix}?$$

$Ax=b$

$x=A^{-1}b$

$x^+=A^+b$

- SVD solve the least squares problem in one step A^+b
  - Computational cost?

$= \begin{bmatrix} 1/3 & 0 \\ 0 & 0 \end{bmatrix} b$

$$\text{squared error } \underbrace{\|\mathbf{b}-\mathbf{Ax}\|^2}_{A^+} = \|\mathbf{b}-\mathbf{U\Sigma\Sigma}^T\mathbf{x}\|^2 = \|\mathbf{U}^T\mathbf{b}-\overset{\Sigma^+}{\underbrace{\mathbf{\Sigma V}^T\mathbf{x}}_{w}}\|^2$$

$= \begin{bmatrix} 2 \\ 0 \end{bmatrix}$

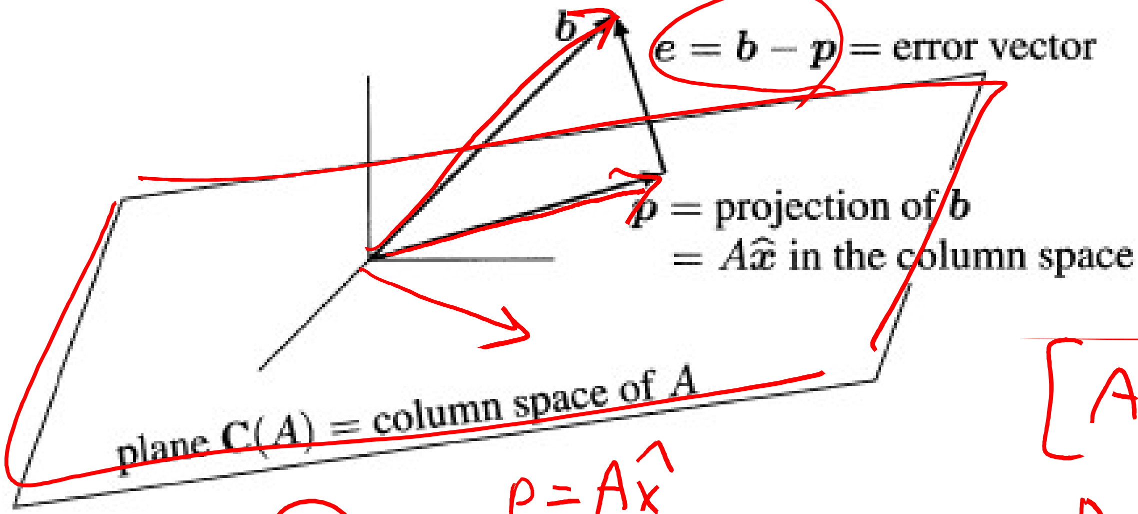
$A=U\Sigma V^T$

$A^{-1}=V\Sigma^{-1}U^{-1}$

$\to w=\Sigma^+U^Tb$

$w=V^Tx^+=\Sigma^+U^Tb \to x^+=V\Sigma^+U^Tb=A^+b$

Normal Equations



$p = A\hat{x}$

$[A][x]$

$x_1 a_1 + x_2 a_2 + \cdots$

$$\mathbf{e} = \mathbf{b} - \mathbf{p}\left(= \mathbf{b} - \mathbf{A}\hat{\mathbf{x}}\right) \perp \mathbf{A}\mathbf{x}$$

$$\Rightarrow (\mathbf{A}\mathbf{x})^T\left(\mathbf{b} - \mathbf{A}\hat{\mathbf{x}}\right) = \mathbf{x}^T\mathbf{A}^T\left(\mathbf{b} - \mathbf{A}\hat{\mathbf{x}}\right) = 0 \rightarrow \mathbf{A}^T\left(\mathbf{b} - \mathbf{A}\hat{\mathbf{x}}\right) = 0$$

$$\begin{cases} \mathbf{A}^T\mathbf{A}\hat{\mathbf{x}} = \mathbf{A}^T\mathbf{b} \;\; (\text{normal equation for } \hat{\mathbf{x}}) \\ \hat{\mathbf{x}} = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{b} \;\; (\text{least squares solution to } \mathbf{A}\mathbf{x} = \mathbf{b}) \\ \mathbf{p} = \mathbf{A}\hat{\mathbf{x}} = \mathbf{A}\left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{b} \;\; (\text{projection of } \mathbf{b} \text{ onto the colum space of } \mathbf{A}) \\ \mathbf{P} = \mathbf{A}\left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T \;\; (\text{projection matrix that multiplies } \mathbf{b} \text{ to give } \mathbf{p}) \end{cases}$$

Least Squares with a Penalty Term

- To prove that the limit is A^+ for every matrix A

$$\text{Minimize} \left\| \mathbf{b} - \mathbf{A}\mathbf{x} \right\|^2 + \delta^2 \left\| \mathbf{x} \right\|^2 \rightarrow \text{Solve} \left(\mathbf{A}^T \mathbf{A} + \delta^2 \mathbf{I} \right) \hat{\mathbf{x}} = \mathbf{A}^T \mathbf{b}$$

$\delta \rightarrow 0$

$\hat{x} = (A^T A + \delta I)^{-1} A^T b \rightarrow \hat{x} = A^+ b$

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{\Sigma}^T$$

$$\mathbf{A}^T \mathbf{A} + \delta^2 \mathbf{I} = \mathbf{V}\mathbf{\Sigma}^T \left(\mathbf{U}^T \mathbf{U} \right) \mathbf{\Sigma} \mathbf{V}^T + \delta^2 \mathbf{I} = \mathbf{V} \left(\mathbf{\Sigma}^T \mathbf{\Sigma} + \delta^2 \mathbf{I} \right) \mathbf{V}^T$$

$$\left(\mathbf{A}^T \mathbf{A} + \delta^2 \mathbf{I} \right)^{-1} \mathbf{A}^T = \mathbf{V} \left(\mathbf{\Sigma}^T \mathbf{\Sigma} + \delta^2 \mathbf{I} \right)^{-1} \left(\mathbf{V}^T \mathbf{V} \right) \mathbf{\Sigma}^T \mathbf{U}^T = \mathbf{V} \left[\left(\mathbf{\Sigma}^T \mathbf{\Sigma} + \delta^2 \mathbf{I} \right)^{-1} \mathbf{\Sigma}^T \right] \mathbf{U}^T$$

$$\lim_{\delta \rightarrow 0} \left(\mathbf{A}^T \mathbf{A} + \delta^2 \mathbf{I} \right)^{-1} \mathbf{A}^T = \lim_{\delta \rightarrow 0} \mathbf{V} \left[\left(\mathbf{\Sigma}^T \mathbf{\Sigma} + \delta^2 \mathbf{I} \right)^{-1} \mathbf{\Sigma}^T \right] \mathbf{U}^T = \mathbf{V} \mathbf{\Sigma}^+ \mathbf{U}^T = \mathbf{A}^+$$

Randomized Linear Algebra

$AB \to CR$

AS $\quad$ S^TB

- Matrix multiplication
  - Sampling matrix S
  - $C=AS$ and $R=S^TB \rightarrow CR=ASS^TB \sim AB$
  - It will not be true that SS^T is close to I
  - It will be true that the expected value of SS^T is I
- Random matrix multiplication with the correct mean AB
- Norm-squared sampling minimizes the variance
- Applications of random matrix multiplication
  - Interpolative approximation $A \sim CMR$
  - Application of A by a low rank matrix
  - Approximation of the SVD of A

P_j

$\dfrac{\|a_j\|\|b_j^T\|}{c}$