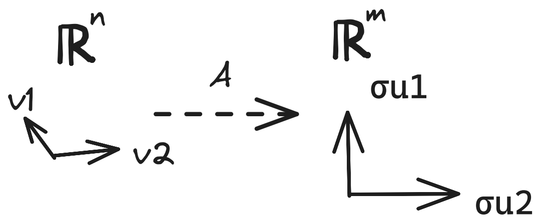
Lecture 7 Low Rank Approximation and Pseudo inverse

Note by Samion Suwito on 2/11/25

Last Time

A (Symmetric) = $A = U\Lambda U^\intercal = \sum \lambda_i u_i u_i^\top$ (weighted sum of projection on U where the weights are eigenvalues and since they're orthogonal it's an orthogonal basis)
$A \in \mathbb{R}^{m\times n} : A = U\Sigma V^\top = \sum \sigma_i u_i v_i^\top$



Spectral Decomposition

Solve: $u_1 = \arg\max_{u:\|u\|_2=1} u^\top Au$
$\lambda_1 = u_1^\top Au_1$
$A' = A - \lambda_1 u_1 u_1^\top$
Then repeat to get the next eigenvector value pair.

SVD

Solve: $v_1 = \arg\max_{v:\|v\|_2=1} \|Av\|_2$
$u_1 = \frac{Av_1}{\|Av_1\|_2}$
$\sigma_1 = \|Av_1\|_2$
$A' = A - \sigma_1 u_1 v_1^\top$
Then repeat to get the next values.

Low Rank Approximation

Given matrix $A \in \mathbb{R}^{m\times n}$, find a "low rank" "approximation" to A.
"Low" is subjective
"Approximation" definition used to determine how close things are to each other.

$$X = \begin{bmatrix} 5 & 4 & \cdots \\ 1 & 5 & \cdots \\ 2 & 3 & \cdots \\ \vdots & \vdots & \cdots \\ 5 & 1 & \cdots \end{bmatrix}$$

Each column is a user profile of a Netflix. Each row corresponds to a movie in Netflix's database. For each movie, there's a number (secret sauce but let's say star) to how much you like a movie.

This matrix has a bunch of empty values for the movies not watched. There a typically patterns to what certain users like to watch e.g. sci-fi movie wathcers might like action. Then a lot of the columns may look similar to each other. This means there will be a *low rank*.

Netflix Challenge

Find a way to fill the blanks in a way that can beat the predictions of Netflix's algorithm. The winners used Low Rank approximation essentially.

$$\min_{A_k \in \mathbb{R}^{m \times n}} \|A - A_k\|_F^2 = \sum_{i=k+1}^{r} \sigma_i^2$$

S.t. rank(A_k) =k, increasing k will therefore decrease error

We use frobinias norm since we can relate it to SVD.
Lecture 4 Orthogonality and QR Decomp > Matrix Norms
Optimal solution is:

$$A_k^* = \sum_{i=1}^{k} \sigma_i u_i v_i^\top$$

from the SVD of A. Experiment to find k. As k increases the approximation gets better and better

Example: Distance to rank deficiency

Take a rank 0 matrix and add random noise to each entry, now the matrix is rank n despite still looking like rank 0.

How far away is it from a matrix of lower rank. Suppose $A \in \mathbb{R}^{m \times n}$ has full column rank (if it has full row then transpose it).

Question: How big is the smallest petarbation $\delta A \in \mathbb{R}^{m \times n}$ s.t. $A + \delta A$ is rank deficient?
$\min \|\delta A\|_F$ such that rank $A + \delta A = n - 1$
Which is σ_n

Can be written as
$\min \|\delta A + A - A\|$ s.t. rank $\delta A + a$ is n-1
We can think of $\delta A + A$ as a rank n-1 approximation
Which is the same as

$$\min_{B:\text{rank}(B)=n-1} \|B - A\|_F$$

$$\implies B^* = \sum \sigma_i u_i v_i^\top$$

Therefore $\delta A^* = B^* - A = -\sigma_n u_n v_n^\top$

PCA, **(Image) Compression**, **Denoising** are low rank approximation problems.

Matrix Pseudo inverse:

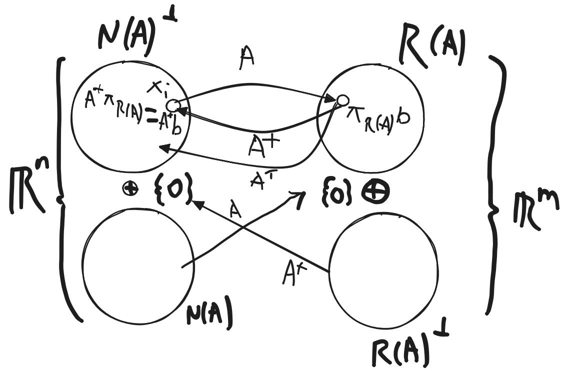
For $A \in \mathbb{R}^{m \times n}$, the Moore-Penrose Pseudo inverse $A^+ \in \mathbb{R}^{n \times m}$ is unique matrix satisfying:

1. $AA^+A = A$
2. $A^+AA^+ = A^+$
3. $(AA^+)^\top = AA^+$
4. $(A^+A)^\top = A^+A$
   Suitable proxy for an inverse.

Claim: $A^+ = V_r \Sigma_r^{-1} U_r^\top$ where $A = U_r \Sigma_r V_r^\top$ = compact SVD
$A^+ = V \Sigma^+ U^\top$ where

$$\Sigma^+ = \begin{bmatrix} \Sigma_r^{-1} & 0 \\ 0 & 0 \end{bmatrix}$$

Intuition: Change of basis picture



What happens if I multiply $A^+\sigma_i u_i = \sigma_i V_r \Sigma_r^{-1} U_r^\top u_i = v_i$
$R(A)^\perp = N(A^+)$ And $R(A^+) = N(A)^\perp$

Examples

1. If A square, invertible: $A^+ = A^{-1}$.
2. If A has full column rank then $A^+ = (A^\top A)^{-1} A^\top =$ left inverse of A
3. If A has full row rank then $A^+ = A^\top (A A^\top)^{-1} =$ right inverse.
   Remark: More information can be gained from SVD:

$$A = [U_r \;\; U_\perp] \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix} [V_r \;\; V_\perp]^\top$$

4. $U_r U_r^\top$ is projection onto $R(A)$ (sum of dyads)
5. $V_r V_r^\top$ is projection onto $N(A)^\perp$
6. $U_\perp U_\perp^\top = I - U_r U_r^\top$ is projection onto $R(A)^\perp$
7. $V_\perp V_\perp^\top = I - V_r V_r^\top$ is projection onto $N(A)$

$x = \pi_{N(A)^\perp}(x) + \pi_{N(A)}(x)$ (Applies for any subspace)
We know $V_r V_r^\top = \pi_{N(A)^\perp}$
$I = \pi_{N(A)^\perp} + \pi_{N(A)} = V_r V_r^\top + \pi_{N(A)}$ Therefore
$I - V_r V_r^\top = \pi_{N(A)}$
$I = VV^\top = V_r V_r^\top + \text{cross terms} + V_\perp V_\perp^\top$ but since V is orthogonal the cross terms are 0. Therefore
$V_\perp V_\perp^\top = \pi_{N(A)}$

Least Squares

Ex. For $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$
Find x of smallest l_2 norm that minimises

$$\min_x \|Ax - b\|_2^2$$

Important $A^\top$ sends $R(A)$ to $N(A)^\perp$ but not to the exact same point. Whereas A^+ does send it to the exact same point.