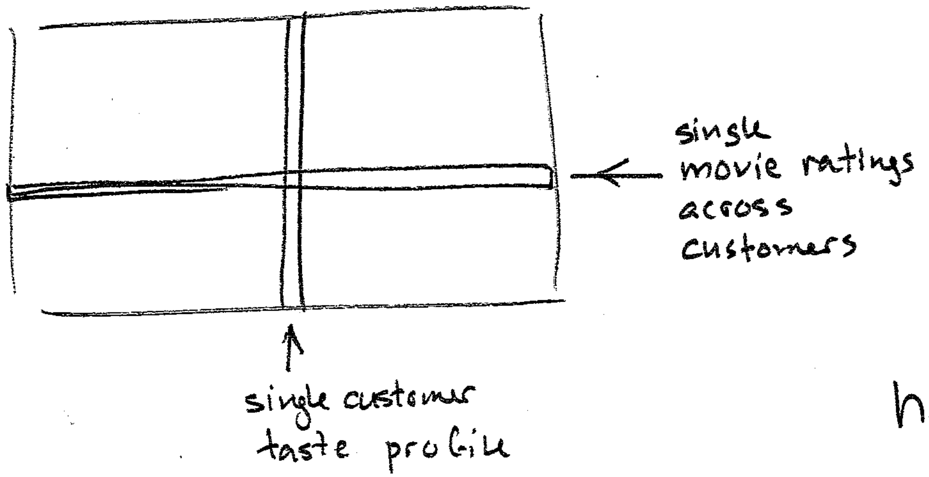
Lecture 11 : The Singular Value Decomposition

Ex: Netflix matrix



$$X = \underbrace{T}_{\text{representative taste profiles}} \; W \leftarrow \text{weights}$$

how many rep. taste profiles are needed?

if $X \in \mathbb{R}^{n \times p}$ and full rank
($n < p$, $\mathrm{rank}(X) = n$)

$\Rightarrow$ naively, need n profiles

Netflix : $p = 1M$, $n = 10k$

storage $\approx np \cdot 4$ bytes

$\approx 10^{11} \cdot 4$ bytes $= 400$ GB.

if we use only $r \ll n$ taste profiles, $X \approx TW$

then $T \sim n \times r$, $W \sim r \times p$, $X \approx$ rank-r

if $r = 10$, storage $\approx (nr + pr)\, 4$ bytes $\approx (10^7 + 10^4)\, 4$ bytes ≈ 44 MB

factor of 9000 less memory/storage

Every matrix $X \in \mathbb{R}^{n \times p}$ can be factorized as

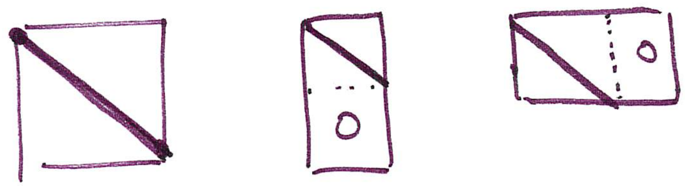
$$X = U \Sigma V^T$$

- $U \in \mathbb{R}^{n \times n}$, orthogonal $(U^T U = U U^T = I)$
  "left singular vectors" of X
- $\Sigma \in \mathbb{R}^{n \times p}$, diagonal

  diagonal elements $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min(n,p)}$

  these are called the singular values of X

$n = p$ $\quad$ $n > p$ $\quad$ $n < p$



- $V \in \mathbb{R}^{p \times p}$, orthogonal $(V^T V = V V^T = I)$
  right singular vectors of X

Interpretation:

U is an ~~orthogonal~~ basis for the columns of X
orthonormal

ΣV^T = basis coefficients.

Netflix example:

i^{th} col of U = basis vector in $\mathbb{R}^n$ = i^{th} representative customer taste profile (vector of normalized movie ratings)

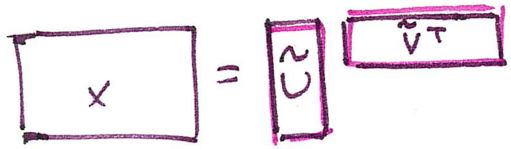
j^{th} col of V^T (row of V) = relative importance of each rep. taste profile to predicting customer j's preferences.

i^{th} row of $V^T = i^{th}$ col of V = vector of users' affinities to the i^{th} representative profile

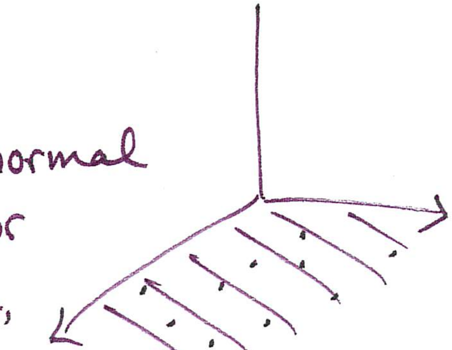
What if X is low rank? (rank $r \lesssim \min(n,p)$)?

cols of X lie in a subspace

$\Rightarrow$ can write $X = \tilde{U}\tilde{V}^T$, where $\tilde{U}$ = orthonormal basis for subspace, $\tilde{V}$ = basis coefficients



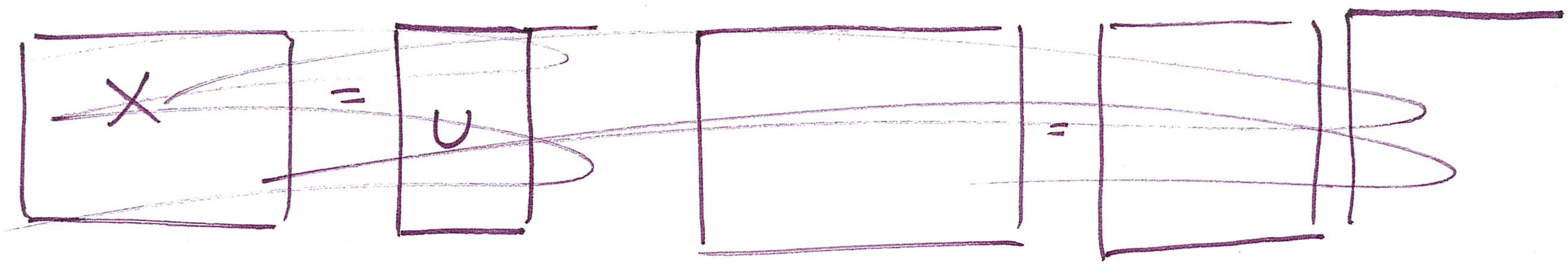
$n \times p$ $n \times r$ $r \times p$



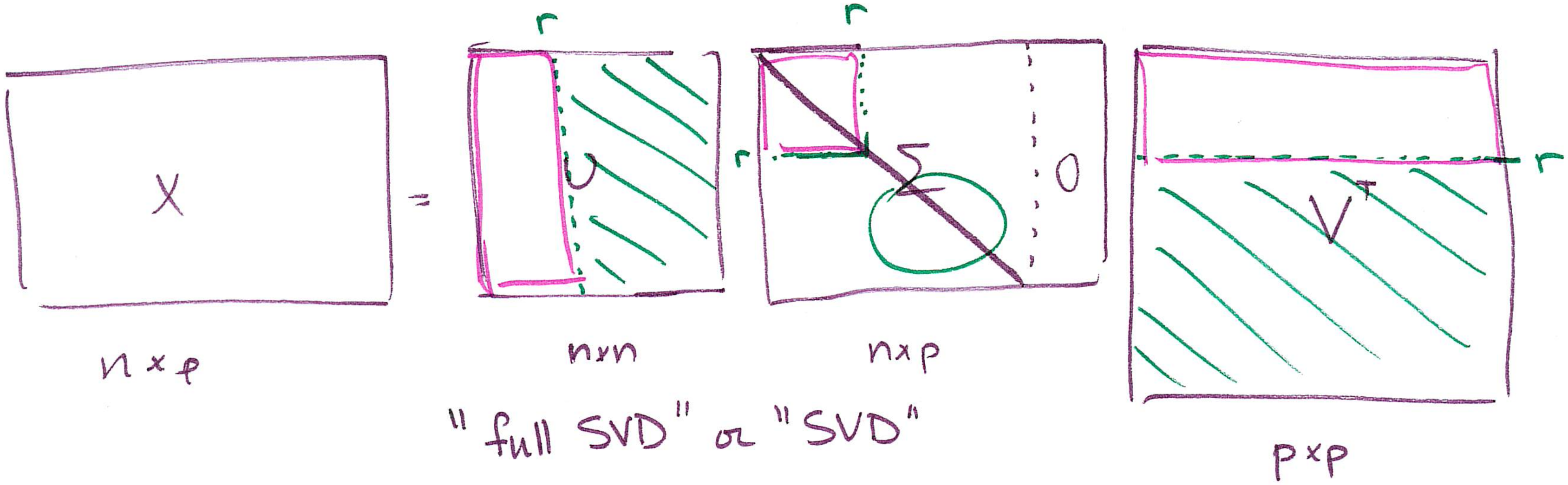
Then $X = U \Sigma V^T$ can be thought of in 2 ways:

(A) matrices U, Σ, V^T same as before, but

only $\sigma_1, \ldots, \sigma_r > 0$, and $\sigma_{r+1}, \ldots, \sigma_{\min(n,p)} = 0$



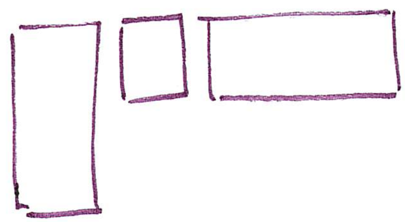
$n \times p$ $n \times n$



"full SVD" or "SVD"

B use smaller matrices:

$$X = \underbrace{U}_{n\times r}\underbrace{\Sigma}_{r\times r}\underbrace{V^T}_{r\times p} =$$



"economy" or "skinny" SVD

U ≈ subspace for customer taste profiles

V = subspace for "movie profile" (for fixed movie, ratings across customers)

Important facts / properties

- $X = U\Sigma V^T \Rightarrow X^T = (V^T)^T \Sigma^T U^T$

  $= V \Sigma U^T$

  $\Rightarrow$ X and X^T have same singular values

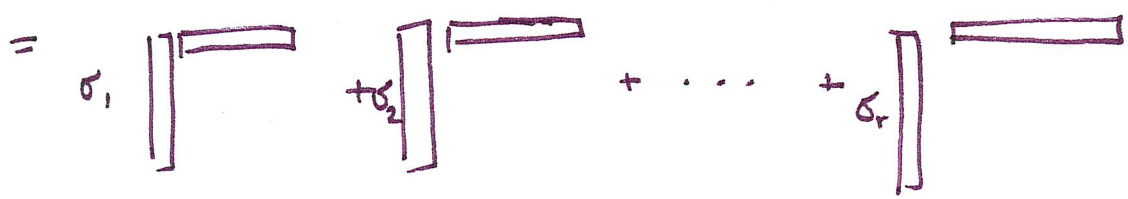
$$\boxed{\begin{array}{l} (AB)^T = B^T A^T \\ (AB)^{-1} = B^{-1} A^{-1} \\ \text{if } A, B \text{ invertible} \end{array}}$$

- r = # nonzero singular values = rank(X)

$$X = \begin{bmatrix} 1 & & 0 \\ & 1 & \\ 0 & & 10^{-4} \end{bmatrix} \Rightarrow U = I,\ V = I,\ \Sigma = \begin{bmatrix} 1 & & 0 \\ & 1 & \\ 0 & & 10^{-4} \end{bmatrix}$$

- $X = \sum_{i=1}^{r} \sigma_i (u_i v_i^T)$ $\quad u_i$ = i^{th} col of $U \in \mathbb{R}^n$

  $\quad v_i$ = i^{th} col of $V \in \mathbb{R}^p$



- If X is square and some singular values $= 0$ then X is not invertible and is called "singular"

- if $X = U\Sigma V^T$ and is square and <u>non</u>-singular <u>(all $\sigma_i > 0$)</u>, then

$$\underbrace{(U\Sigma V^T)}_{X}\underbrace{(V\Sigma^{-1}U^T)}_{X^{-1}} = U\Sigma V^T V \Sigma^{-1} U^T \qquad \text{(full SVD)}$$

$$= U\Sigma\Sigma^{-1}U^T$$

$$= UU^T$$

$$= I \qquad \Longrightarrow X^{-1} = V\Sigma^{-1}U^T$$

- $$\underbrace{(X^TX)^{-1}X^T}_{\text{pseudo-inverse of } X} = (V\Sigma^T \underline{U^T U} \Sigma V^T)^{-1} V\Sigma U^T$$

$$= (V\Sigma^2V^T)^{-1}V\Sigma U^T$$

$$= (V^T)^{-1}\Sigma^{-2}\cancel{V^{-1}V}\Sigma U^T$$

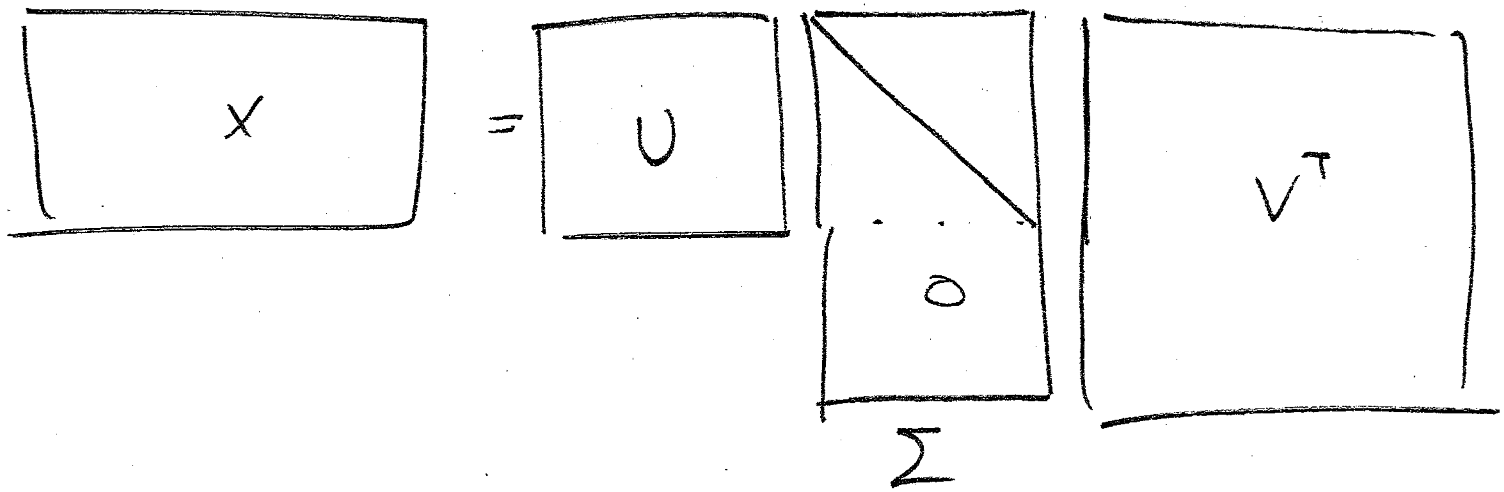
$$= V\Sigma^{-2}\Sigma U^T = V\Sigma^{-1}U^T$$

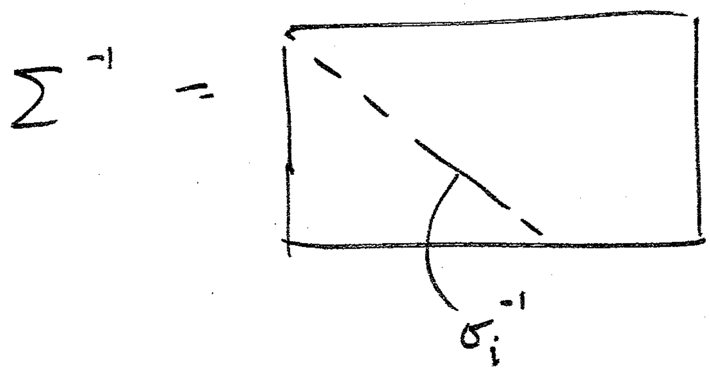
$\boxed{\text{important clarifications in lecture 12!}}$

V is orthogonal

$\Rightarrow V^TV = I$

$\Rightarrow V^T = V^{-1}$





• ~~if $X = U\Sigma V^T$, square, and symmetric : $X^T = X$~~

~~$= X^T = V\Sigma U^T \Rightarrow U = V \Rightarrow X = U\Sigma U^T$~~

- $$\|X\|_2 = \|X\|_{op} := \max_{a \neq 0} \frac{\|Xa\|_2}{\|a\|_2} \approx \text{"size of the operator } X\text{"}$$

$$= \sigma_1 = 1^{st} \text{ singular value}$$

$$\max_{a \neq 0} \frac{\|Xa\|_2}{\|a\|_2} = \max_a \frac{\|U\Sigma V^T a\|_2}{\|a\|_2}$$

b/c V is orthogonal,
$\|V^T a\|_2 = \|a\|_2$
$\|Ua\|_2 = \|a\|_2$

$X \in \mathbb{R}^{n \times p}$

$$= \max_a \frac{\|U\Sigma V^T a\|_2}{\|V^T a\|_2}$$

let $b := V^T a$
$\in \mathbb{R}^p$

$$= \max_{b \neq 0} \frac{\|U\Sigma b\|_2}{\|b\|_2} = \max_{b \neq 0} \frac{\|\Sigma b\|_2}{\|b\|_2} = \max_b \left(\frac{\sum_i (\sigma_i b_i)^2}{\sum_i b_i^2} \right)^{1/2}$$

e.g. $b_i = 1/\sqrt{p} \Rightarrow \|b\|_2 = 1 \Rightarrow \sum_i (\sigma_i b_i)^2 = \frac{1}{p} \sum \sigma_i^2$ = avg. sing val

e.g. $b_i = \begin{cases} 1 & \text{if } i=1 \\ 0 & \text{otherwise} \end{cases} \Rightarrow \|b\|_2 = 1 \Rightarrow \sum_i (\sigma_i b_i)^2 = \sigma_1^2$ = max sing val

$$\Rightarrow \max_{b \neq 0} \frac{\|\Sigma b\|_2}{\|b\|_2} = \sigma_1 \,, \quad \text{maximizing } b = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = V^T a$$

$\Rightarrow$ maximizing $\cancel{a}$ $= Vb = v_1 = 1^{st}$ col of V $= 1^{st}$ right singular vec.