Review

if columns of X are linearly independent,

then $\underset{w}{\operatorname{argmin}} \|Xw - y\|_2^2 = \underbrace{(X^T X)^{-1} X^T}\, y$

pseudoinverse of X

$X^\dagger$

$XX^\dagger$ = projection matrix onto span(cols(X))

if $X = U\Sigma V^T$ (SVD of X), then

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

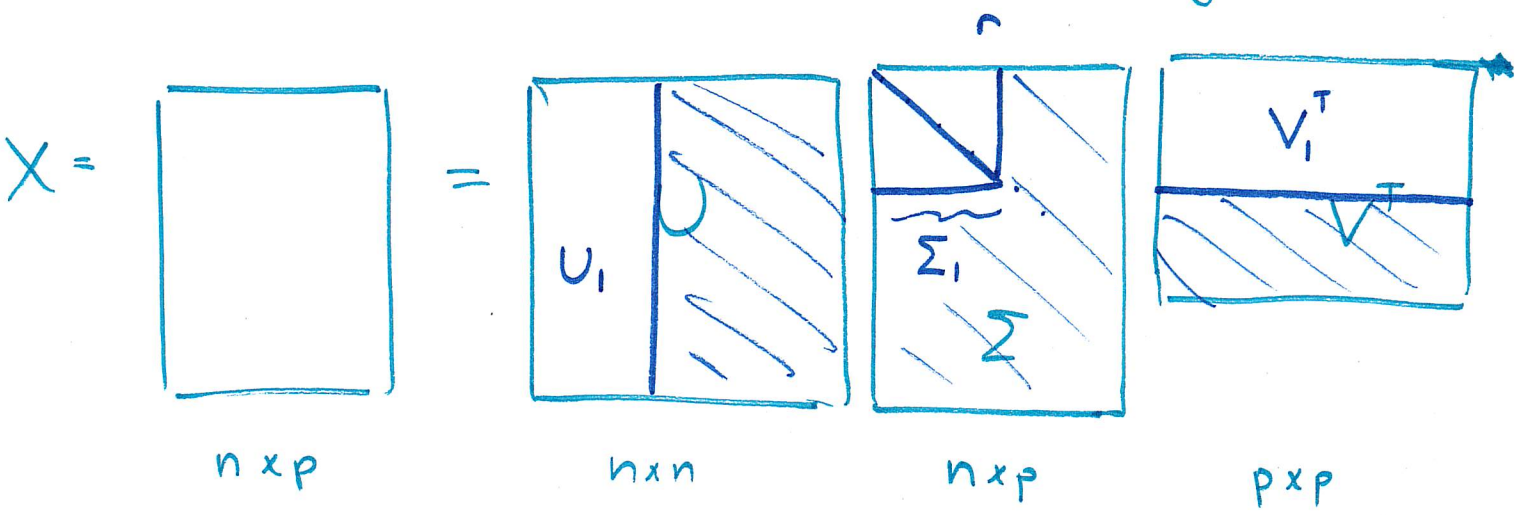
$= V^T \Sigma^{-2} U^T U \Sigma V^T =$

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

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

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

Now if cols of X might not be linearly independent



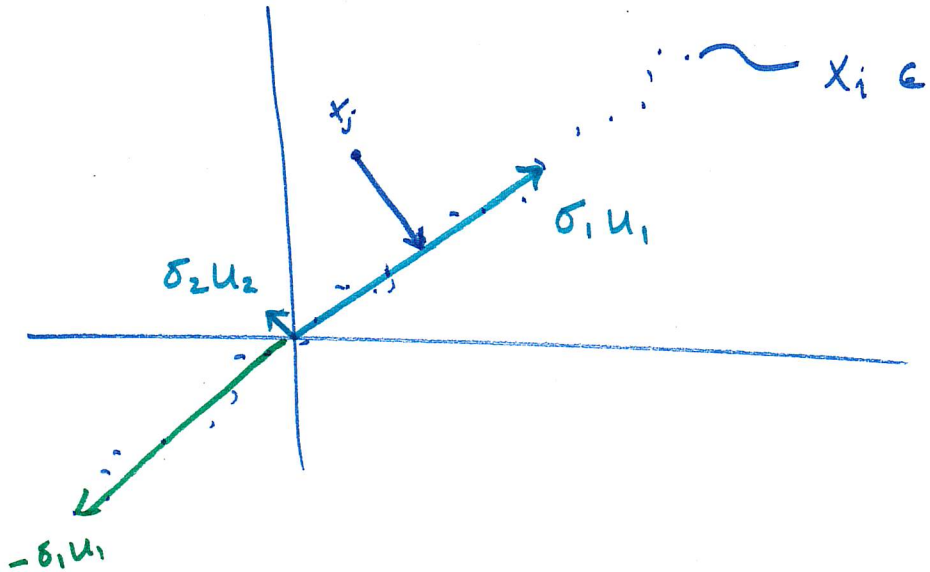
if X has r L.I. cols, then Σ has r nonzero singular values.

Let Σ^+ : take reciprocal of non-zero elements of Σ, then transpose result.

e.g. $\Sigma = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \longrightarrow \Sigma^+ = \begin{bmatrix} 1/4 & 0 & 0 & 0 \\ 0 & 1/3 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$

then $\boxed{X^\dagger = V \Sigma^+ U^T}$ = pseudoinverse of X

$$X^\dagger = V_1 \Sigma_1^{-1} U_1^T$$



$$P_U x_j = U(U^T U)^{-1} U^T x_j$$
$$= U U^T x_j$$

if $U = u_1$,

$$P_{u_1} x_j = (u_1^T x_j) u_1$$

if $X = \begin{bmatrix} 1 & & \\ & 2 & \\ & & 3 \\ 0 & & \end{bmatrix}$, $U = I$, $V = I$, $\Sigma = \begin{bmatrix} 1 & & & \\ & 2 & & \\ & & 3 & \\ & & & 4 \end{bmatrix}$

σ_i's are increasing

if $x_i \in$ subspace w/ basis U,
then $x_i = U a_i$ for some a_i

if wanted σ_i's decreasing,

$$\Sigma = \begin{bmatrix} 4 & & \\ & 3 & \\ & & 2 \\ & & & 1 \end{bmatrix}$$

$$\underset{X}{\begin{bmatrix} 1 & & \\ & 2 & 0 \\ & 3 & \\ 0 & & 4 \end{bmatrix}} = \underbrace{\underset{U}{\begin{bmatrix} 0 & & 1 \\ & 1 & \\ 1 & & 0 \end{bmatrix}} \begin{bmatrix} 4 & & \\ & 3 & \\ & & 2 \\ & & & 1 \end{bmatrix}} \underset{V^T}{\begin{bmatrix} & & \\ & & \\ & & \\ & & \end{bmatrix}}$$

$$=$$

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

Lecture 20: Stochastic Gradient Descent

$$\underset{w}{\text{minimize}} \quad \ell(w) + \lambda r(w)$$

$\ell(w)$: loss, measures fit to data

$r(w)$: regularization function

$r(w) = \|w\|_2^2$ (ridge)

$r(w) = \|w\|_1$ (lasso)

e.g. $\ell(w) = \|y - Xw\|_2^2 = \sum_{i=1}^n (y_i - x_i^T w)^2$, $f_i(w) = (y_i - x_i^T w)^2$

$$\nabla \ell(w) = -2X^T(y - Xw) = \sum_{i=1}^n -2(y_i - x_i^T w) x_i$$

e.g. $\ell(w) = \sum_{i=1}^n (1 - y_i x_i^T w)_+$ (hinge loss), $f_i(w) = (1 - y_i x_i^T w)_+$

$$\nabla \ell(w) = \sum_{i=1}^n I_{\{1 > y_i x_i^T w\}} (-y_i x_i)$$

in either case

let $f(w) = \ell(w) + \lambda r(w)$

goal: $\min_w f(w)$

write $f(w) = \sum_{i=1}^{n} f_i(w)$

before: Gradient Descent

$$\hat{w}^{(k+1)} = \hat{w}^{(k)} - \frac{\tau}{2} \nabla f(\hat{w}^{(k)})$$

(e.g., w/ squared error loss:

$$\hat{w}^{(k+1)} = \hat{w}^{(k)} - \tau X^T (X\hat{w}^{(k)} - y)$$)

today: Stochastic Gradient Descent (SGD)

@ iteration k, choose $i_k \in \{1, \dots, n\}$

$$\hat{w}^{(k+1)} = \hat{w}^{(k)} - \frac{\tau}{2} \nabla f_{i_k}(\hat{w}^{(k)})$$

SGD

- each iteration is easier / faster to compute
- need more iterations.

How to choose i_k?

A. Cyclical (incremental gradient descent)

$i_k = k \bmod n$

if $n = 3$, $i_k = 1, 2, 3, 1, 2, 3, 1, 2, 3$

B. random permutations

every n rounds, reshuffle training data

i_k's $= 1, 3, 2, 3, 1, 2, 2, 1, 3 \ldots$

C. Stochastic gradient descent

$i_k \sim \text{unif}(1, \ldots, n)$

i_k's $= 1, 3, 3, 2, 3, 1, 2, 2, 2, 1, 3$

$\mathbb{E}[\nabla f_i(w)] = \nabla f(w)/n$

Ex 1 : $f(w) = \|y - Xw\|_2^2 + \lambda \|w\|_2^2$

$$= \sum_{i=1}^{n} (y_i - x_i^T w)^2 + \underline{\lambda \|w\|_2^2}$$

$$\Rightarrow f_i(w) = (y_i - x_i^T w)^2 + \frac{\lambda}{n} \|w\|_2^2$$

$$\Rightarrow \sum_{i=1}^{n} f_i(w) = f(w)$$

$$\nabla f_i(w) = -2(y_i - x_i^T w) x_i + 2\frac{\lambda}{n} w$$

SGD :

$$\hat{w}^{(k+1)} = \hat{w}^{(k)} - \frac{\tau}{2} \left(-2(y_i - x_i^T \hat{w}^{(k)}) x_i + 2\frac{\lambda}{n} \hat{w}^{(k)}\right)$$

$$= \hat{w}^{(k)} + \tau (y_i - x_i^T \hat{w}^{(k)}) x_i - \frac{\tau\lambda}{n} \hat{w}^{(k)}$$