Last lecture

$$Y = \underset{N\times D}{X}\,\underset{D\times 1}{w} + \epsilon$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

$$\begin{array}{cc} & \longleftarrow \text{features} \longrightarrow \\ \underset{N\times D}{X} = & \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{bmatrix} \end{array} \text{sample}$$

N = all the sample

$$\epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix} \leftarrow \text{error residue}$$

side

$f(x) = b^T x$

$f'(x) = b$

b/c

if $f(x) \in \mathbb{R}$

$x \in \mathbb{R}^D$

$$f'(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_D} \end{bmatrix}$$

$$\underline{\underline{RSS}} = \sum_{i=1}^{N} (y_i - x_i^T w)^2$$

$$= \sum_{i=1}^{N} \epsilon_i^2 = \|\epsilon\|_2^2$$

want to find

$\hat{w}$ s.t $\hat{y}_i = \underline{\underline{x_i^T \hat{w}}}$.

minimizes $\sum_{i=1}^{N} (y_i - \hat{y}_i)^2$

$\Updownarrow$

or in Matrix notation

minimize $(Y - XW)^T (y - XW)$

$(y - X\hat{w})$

$\underline{\underline{\hat{y}}} = X\hat{w}$ s.t $(y - \hat{y})^T (y - \hat{y})$ is minimum

Let search for such $\hat{w}$.

Note the action of a generic $W \in \mathbb{R}^D$ on X

$$\bar{y} = \underset{N\times D}{X}\,\underset{D\times 1}{W}$$

$$= \begin{bmatrix} \tilde{x}_1 & \tilde{x}_2 & \cdots & \tilde{x}_D \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_D \end{bmatrix}$$

$\tilde{x}_j \in \mathbb{R}^D$ is a column of X

ie $\bar{y} = XW = w_1 \tilde{x}_1 + w_2 \tilde{x}_2 + \cdots w_D \tilde{x}_D \in \mathbb{R}^N$

weighted linear combination of the columns of the matrix X, where weights w_i comes from vector $W \in \mathbb{R}^D$

Hence as W varies it spans column space of matrix X

Our desired vector $\hat{y} = X\hat{w}$ will also be a point in this column space of X which minimize $(y - Xw)^T(y - Xw)$

i.e $w = \hat{w}$

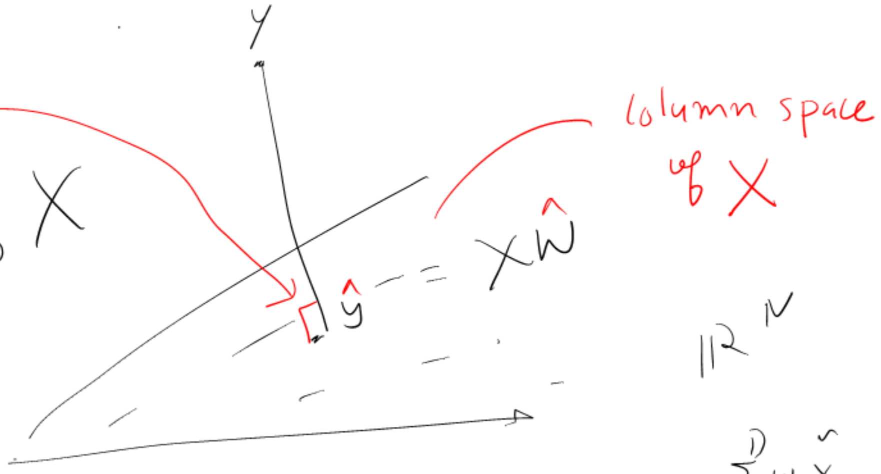
If $\hat{y} = X\hat{w}$ is such a point then

vector $y - \hat{y}$ has to be perpendicular to column space of X

OR

$y - \hat{y}$ has to be perpendicular to columns of X

OR



$\sum_{i=1}^{D} w_i x_i$

side: $a \perp b \in \mathbb{R}^N \iff a^T b = b^T a = \langle a, b \rangle = 0$

This is algebraic definition of two vector a, b being perpendicular is

$$x_j^T (y - \hat{y}) = 0 \quad \forall \text{(all)}\ j = 1 : D$$

$$x_j^T (y - X\hat{w}) = 0$$

$$X^T (y - X\hat{w}) = 0$$

$$X^T y - X^T X \hat{w} = 0$$

$$\Rightarrow \hat{w} = (X^T X)^{-1} X^T y$$

Ridge Regression

MLE can overfit. It tries to explain current evidence, Not good for noisy situations.

$P(D|W)$

we can force each w_i with prior belief i.e $w_i \sim \mathcal{N}(x|0,\tau)$

$P(W|D)$

then using MAP estimate of $W = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_D \end{bmatrix}$

equivalent to solving following optimization problem.

will try to post a derivation of this

$$\underset{W}{\arg\min} \Big(\underbrace{\sum_{i=1}^{N} (y_i - W^T x_i)}_{MLE} \Big) + \lambda \|W\|_2^2$$

$\Updownarrow$

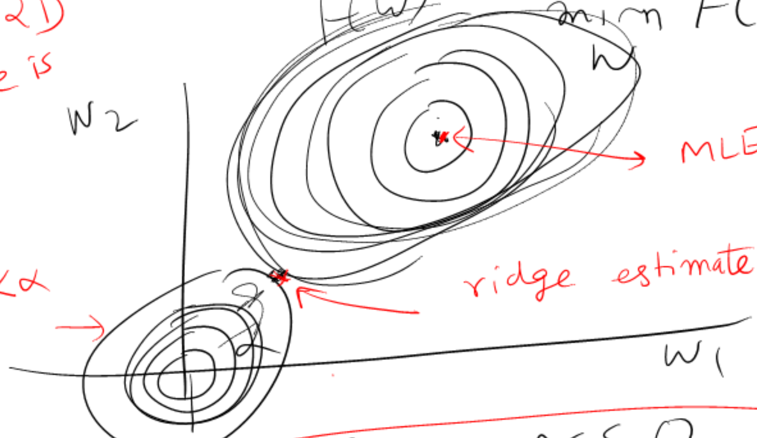
$$\arg\min \Big(\sum_{i=1}^{N} (y_i - W^T x_i) \Big) \quad s.t \quad \|W\|_2^2 < \alpha$$

$\Updownarrow$

$$\underset{W}{\min}\ F(W) \quad s.t \quad \|W\|_2^2 < \alpha$$

$$w_1^2 + w_2^2 < \alpha$$

In 2D picture is



Ridge Regression

LASSO

$$\min F(W) + \lambda |W|$$

$$W = |W|_1 + |W|_2 + \cdots |W|_k$$

$$\begin{bmatrix} 0 \\ \hat{W}_2 \end{bmatrix}$$

$$\mu_1 \|W\|^2 + \mu_2 |W|$$