Lec 6.1 Least squares Geometry

1. Recall:

- We're try to find $\underline{w}$ that minimize

$$\min_{\underline{w}} \| \underline{y} - \underbrace{\underline{X}\,\underline{w}}_{\hat{\underline{y}}} \|^2$$

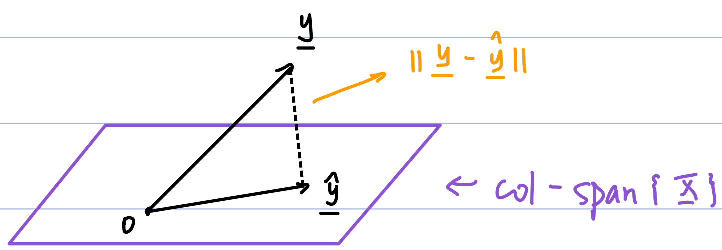
- LS solution : $\boxed{\underline{w}_{LS} = \underbrace{(\underline{X}^T \underline{X})^{-1} \underline{X}^T}_{\tilde{\underline{X}}^{+}} \underline{y}}$

2. Moving on:

↳ $\hat{\underline{y}} = \underline{X}\,\underline{w}$: linear comb. of col of $\underline{X}$.

⇒ $\hat{\underline{y}}$ is a vector in col-span of $\{\underline{X}\}$.

∴ To minimize $\|\underline{y} - \hat{\underline{y}}\|$ (distance between $\hat{\underline{y}}$ and $\underline{y}$)



- The best $\hat{\underline{y}}$ (i.e. $\hat{\underline{y}}_{LS}$) is the projection of $\underline{y}$ onto col-span $\{\underline{X}\}$.

⇒ Every col. of $\underline{X}$ is orthorgnoal to $\underline{y} - \hat{\underline{y}}$

↳ $\underline{a} \perp \underline{b} \Leftrightarrow \underline{a}^T \underline{b} = 0$

⇒ $\underline{X}^T (\underline{y} - \hat{\underline{y}}_{LS}) = 0$

⇒ $\underline{X}^T (\underline{y} - \underline{X}\,\underline{w}_{LS}) = 0$

⇒ $\underline{X}^T \underline{X}\,\underline{w}_{LS} = \underline{X}^T \underline{y}$

$\boxed{\Rightarrow \underline{w}_{LS} = (\underline{X}^T \underline{X})^{-1} \underline{X}^T \underline{y}}$ // Same result, but from Geometric view.

Lec 6.2 Regularized linear regression / Least squares

1. Regularized version of Lin. regr:

↳ $$\min_{\underline{w}} \underbrace{\|\bar{X}\,\underline{w} - \underline{y}\|^2}_{\text{simple version}} + \underbrace{\lambda\|\underline{w}\|^2}_{\text{penalty func. (against large } \|\underline{w}\|)}$$

if $\underline{w}$ is large, penalty ↑

↳ **Motivition:**

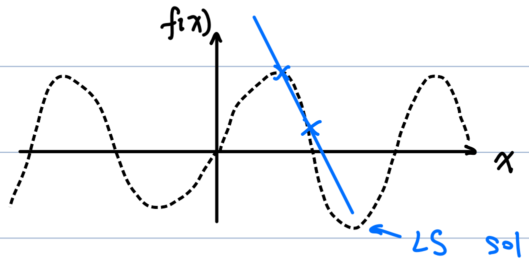
Want $\|\underline{w}\|$ small to avoid over-fitting

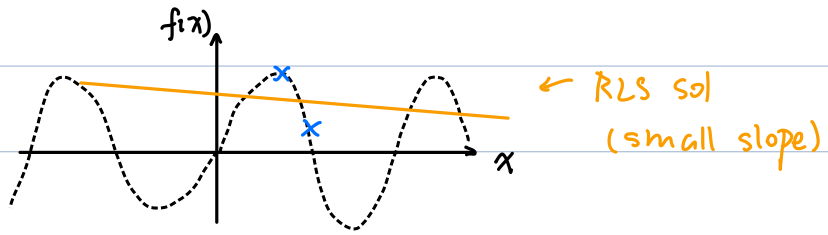
due to ① noisy data ② Not enough data

2. Text book example (can skip):

↳ Target func: $f(x) = \sin(\pi x)$

Training set: 2 points (randomly sampled)





↳ $f(\underline{w}) = \|\bar{X}\,\underline{w} - y\|^2 + \lambda\|\underline{w}\|^2$

$$\frac{\partial f}{\partial w_k} = \frac{\partial}{\partial w_k}\left(\|\bar{X}\,\underline{w} - \underline{y}\|^2 + \lambda \sum_{j=0}^{d} w_j^2\right)$$ // break into one

k-th item

$$= 2\left[\bar{X}^T(\bar{X}\,\underline{w} - \underline{y})\right]_k + 2\lambda w_k$$

$$= 2\left[\bar{X}^T(\bar{X}\,\underline{w} - \underline{y}) + \lambda w\right]_k$$

$\nabla f(\underline{w}) = 2[\bar{X}^T(\bar{X}\underline{w} - \underline{y}) + \lambda w]$

We want $\nabla f(\underline{w}) = \underline{0}$

$\therefore (\bar{X}^T\bar{X} + \lambda I)\,\underline{w}_{RLS} = \bar{X}^T\underline{y}$

$$\underline{w}_{RLS} = (\bar{X}^T\bar{X} + \lambda I)^{-1}\bar{X}^T\underline{y}$$

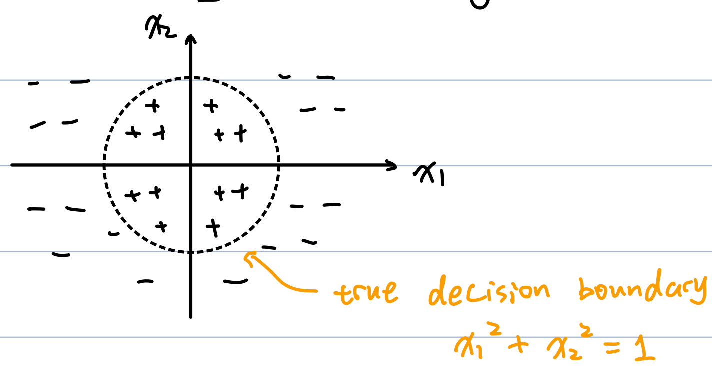
Note: ① If $\lambda = 0$, then RLS is same as LS form

② RLS side benifit more numerically stable solution.

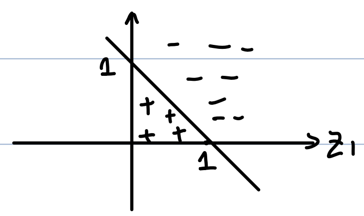
Lec 6.3 Non-linear transformation

1. linear func may not be enough:

e.g.



↳ We transfer this into $z_1 = x_1^2$, $z_2 = x_2^2$



$\Rightarrow z_1 + z_2 = 1$

↳ Suppose $h(\underline{z}) = \text{sign}(z_1 + z_2 - 1)$

Then our solution $g(\underline{x}) = \text{sign}(x_1^2 + x_2^2 - 1)$

∴ We can transfer to another space, and use Lin. Regression

2. General non-linear regression:

↳ Let $\underline{z} = \Phi(\underline{x})$ be a non-linear transform

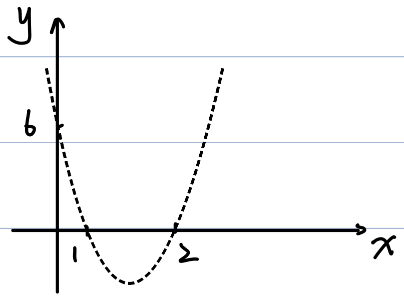
$\underline{h}(\underline{z}) = \underline{w}^T \underline{z}$ be a linear classifer in z space

Then $\underset{\uparrow}{g}(\underline{x}) = h(\Phi(\underline{x}))$ is a non-linear classifier in $\underline{x}$ space

classifier in terms of x

// $\Phi(\cdot)$ is called a feature transform

3. e.g. Quadratic Regression:



$\underline{z} = (z_0 = 1,\ z_1 = x,\ z_2 = x^2)$ // $d=2$

$\therefore\ y = \underline{w}^T \underline{z}$

$= w_0 + w_1 z_1 + w_2 z_2$

$= w_0 + w_1 x_1 + w_2 x_2^2$

$\because\ \underline{z_1} = (1, 0, 0)$, $y_1 = 6$

$\underline{z_2} = (1, 1, 1)$, $y_2 = 0$

$\underline{z_3} = (\underset{\substack{\uparrow \\ 1}}{1}, \underset{\substack{\uparrow \\ x}}{2}, \underset{\substack{\uparrow \\ x^2}}{4})$, $y_3 = 0$

$\therefore\ Z = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{bmatrix}$, $\underline{y} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$

$\therefore\ \underline{w}_{LS} = \underbrace{(Z^T Z)^{-1} Z^T}_{\text{pesudo inverse of } Z} \underline{y}$

$\because\ Z$ is in full rank $\quad \therefore\ Z^{+} = Z^{-1}$

$\therefore\ \underline{w}_{LS} = (Z^T Z)^{-1} Z^T \underline{y} = Z^{-1} \underline{y} = \begin{bmatrix} 6 \\ -9 \\ 3 \end{bmatrix}$

$\therefore\ y = 6 - 9 z_1 + 3 z_2 = 6 - 9x + 3x^2$