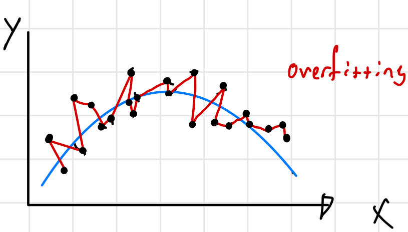
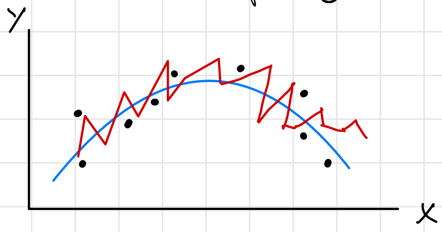
Lecture 6

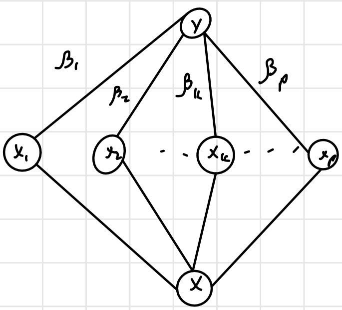
- Overfitting & Regularization & model complexity

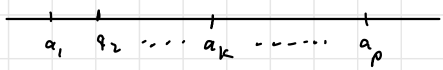


training data



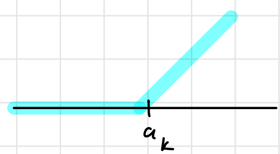
testing data





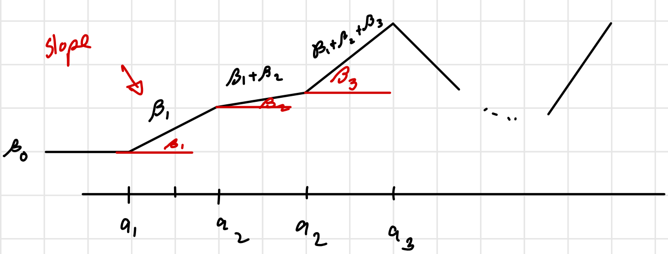
knots

$X_k = \max(0, x - a_k)$ design



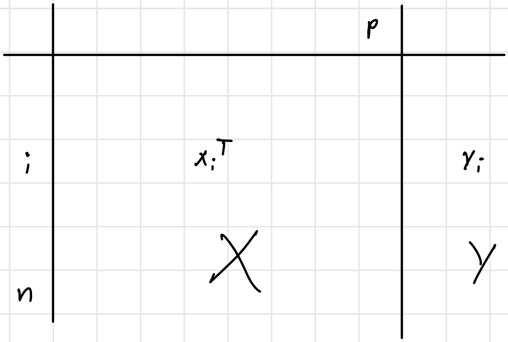
ReLU

$S = f(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \beta_p x_p$



$\beta_k \doteq$ change of slope
"curvature"

piecewise linear / spline



$p >> n$

Ridge Regression: $| y - X\beta |^2 + \lambda \underbrace{|\beta|^2}$

l_2 - regularization

we want the curvature to be small.

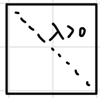
$$\text{Loss}(\beta) = (Y - X\beta)^T (Y - X\beta) + \lambda \beta^T \beta$$

$$= Y^T Y - \beta^T X^T Y - Y^T X \beta + \beta^T X^T X \beta + \lambda \beta^T \beta$$

$$= \text{const} - 2\langle X^T Y, \beta \rangle + \beta^T (X^T X + \lambda I_p) \beta$$

$$\text{Loss}'(\beta) = -2 X^T Y + 2 (X^T X + \lambda I_p) \beta = 0$$

$$\hat{\beta} = (X^T X + \lambda I)^{-1} X^T Y$$



ridge

Shrinkage estimator

- Note :

$$|\beta|_{\ell_2}^2 = \sum_{k=1}^{p} \beta_k^2 \quad \text{(We didn't penalize } \beta_0\text{)}$$

$$= \beta^T D \beta$$

$$\beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} \qquad D = \begin{pmatrix} 0 & & & \\ & \lambda & & \\ & & \ddots & \\ & & & \lambda \end{pmatrix}$$

So the more general version of ridge regression is

$$Loss'(\beta) : -2X^T Y + 2(X^T X + \underline{D})\beta = 0$$

- Suppose now : $Loss(\beta) = \frac{1}{2} |Y - X\beta|^2$

$$= \frac{1}{2} \sum_{i=1}^{n} (Y_i - x_i^T \beta_i)^2$$

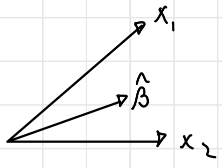
- Gradient Descent : $\beta_0 = 0$ ← starting point of gradient descent

$$Loss'(\beta) = -\sum_{i=1}^{n} e_i x_i$$

$$\beta_{t+1} = \beta_t - \eta_t L'(\beta_t)$$

$$\longrightarrow \underset{p\times 1}{\hat{\beta}} = \sum_{i=1}^{n} c_i \underset{p\times 1}{x_i}$$

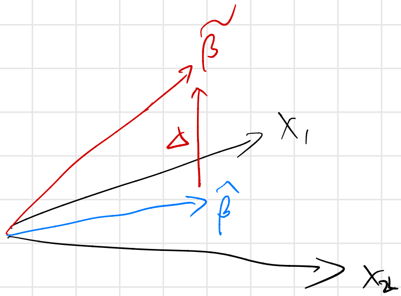
Representer



$L(\hat{\beta}) = 0$, but there can be other solutions: $L(\tilde{\beta}) = 0$

- This implies $x_i^T \hat{\beta} = y_i \quad \forall i$

$$x_i^T \tilde{\beta} = y_i \quad \forall i$$



Let $\tilde{\beta} = \hat{\beta} + \Delta$

Then $x_i^T \Delta = 0 \quad \forall i$

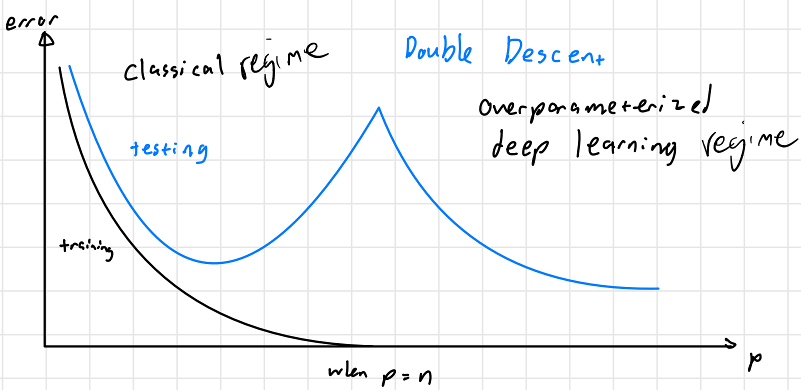
So $\Delta \perp \hat{\beta}$.

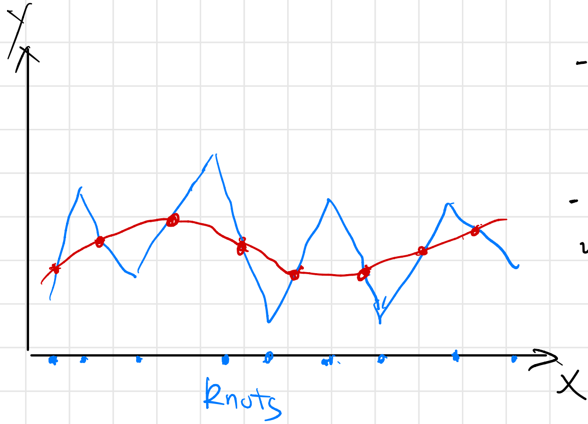
- Now $|\tilde{\beta}|^2 = |\hat{\beta}|^2 + |\Delta|^2$

therefore $\hat{\beta}$ min $|\beta|^2$ among all β s.t. $L(\beta) = 0$

- In other words gradient descent self-regularizes.

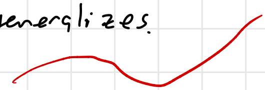
  - implicit regularization
  - benign overfitting



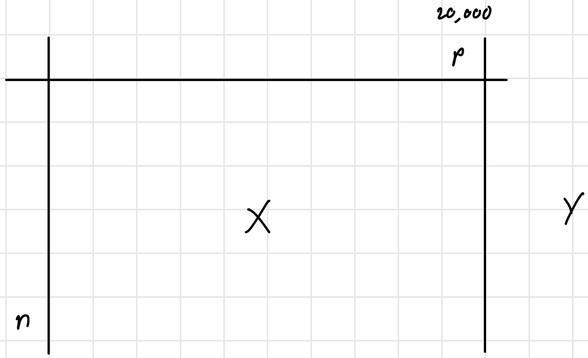


- small # of nodes

- with many nodes + GD we get a smooth overfit, so hopefully this generalizes.



- Gene :



"$\|\beta\|_{\ell_0}$"

Not convex or smooth { $Loss(\beta) = \frac{1}{2} | Y - X\beta |^2 + \lambda \#(\beta_j \neq 0)$ remove irrelivant variables

relax ↓ ↓ relax (approximation)

Convex $loss(\beta) = \frac{1}{2} | Y - X\beta |^2 + \lambda |\beta|_{\ell_1}$

$|\beta|_{\ell_1} = \sum_{k=1}^{p} |\beta_k|$

- Lasso : Least absolute shrinkage selection operator.

$$\text{Loss}(\beta) = \frac{1}{2} \left| Y - \sum_{j=1}^{p} x_j \beta_j \right|^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

$$y - \sum_{j \neq k} x_j \beta_j \quad - x_k \beta_k$$

- Coordinate Descent : $\hat{y}$

Each iteration :

for k in $1:p$

$$\hat{\beta}_k = \min_{\beta_k} |\hat{y} - x_k \beta_k|^2 + \lambda |\beta_k|$$

$$\hat{y} = y - \sum_{j \neq k} x_j \beta_j$$

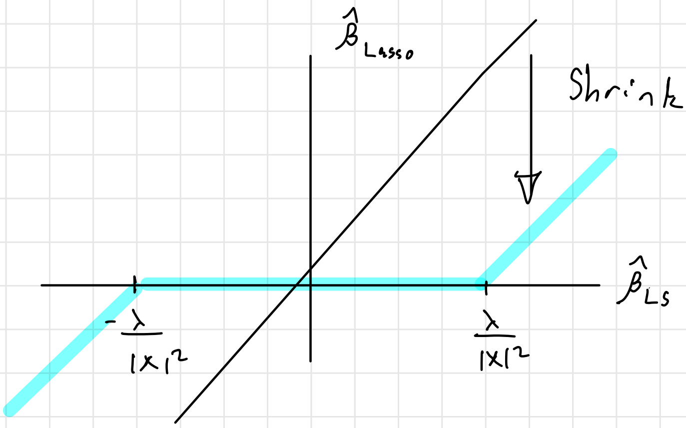
- One-dimensional Problem :

$$L(\beta) = |Y - X\beta|^2 + \lambda |\beta|$$

So $\hat{\beta}_{LS} = \dfrac{\langle x, y \rangle}{|x|^2}$ $\quad (\lambda = 0)$

$$\hat{\beta}_{Lasso} = \text{sign}(\hat{\beta}_{LS}) \max\left(0, \left|\hat{\beta}_{LS}\right| - \frac{\lambda}{|X|^2}\right)$$

soft-thresholding , selection



$H_0 : \beta = 0$

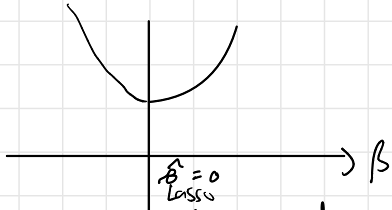
$H_1 : \beta \neq 0$

- Consider:

$$L'(\beta) = \begin{cases} -\langle x, y\rangle + |x|^2 \beta + \lambda & \text{if } \beta \geq 0 \\ -\langle x, y\rangle + |x|^2 \beta - \lambda & \text{if } \beta < 0 \end{cases}$$

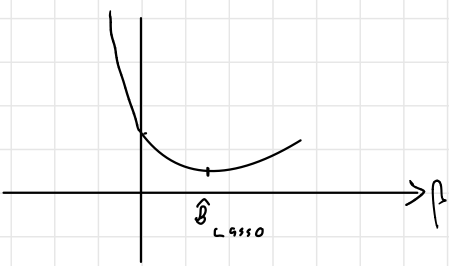
$$L'_{\text{left}}(0) = -\langle x, y\rangle - \lambda$$

$$L'_{\text{right}}(0) = -\langle x, y\rangle + \lambda$$



happens when $\langle x, y\rangle \in [-\lambda, \lambda]$

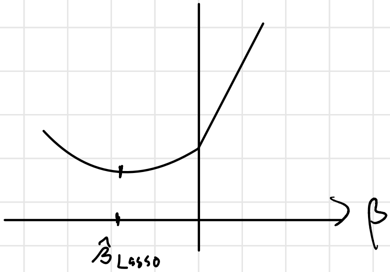
If $L'_{\text{left}} < 0$ & $L'_{\text{right}} < 0$



$$-\langle x, y\rangle + |x|^2 \beta + \lambda = 0$$

$$\hat{\beta}_{Lasso} = \hat{\beta}_{LS} - \frac{\lambda}{|x|^2}$$

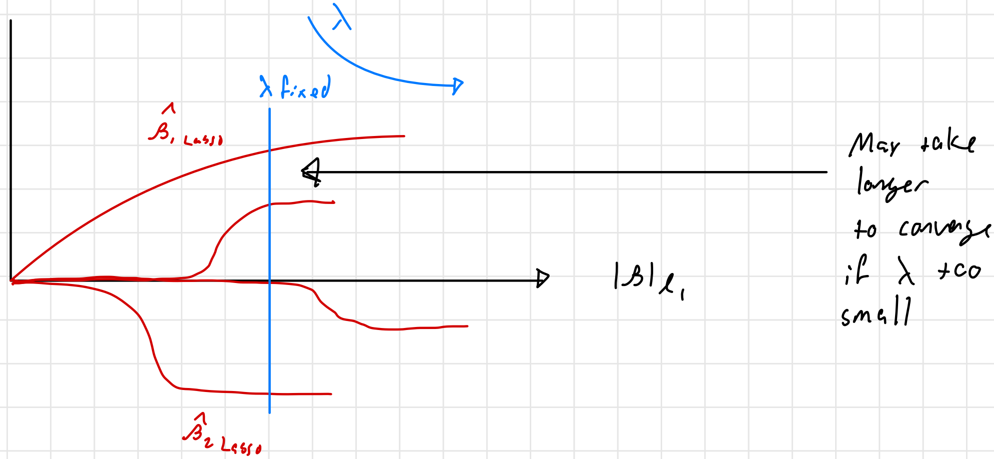
If $L'_{\text{left}} > 0$ & $L'_{\text{right}} > 0$



$$-\langle x, y\rangle + |x|^2 \beta - \lambda = 0$$

$$\hat{\beta}_{Lasso} = \hat{\beta}_{LS} + \frac{\lambda}{|x|^2}$$

- Solution Path:
  - Start with $\lambda = \max_j |\langle x_j, y \rangle|$
  - then gradually reduce λ.
  - $\hat{\beta}_{j\,Lasso} = \max\left(0, \hat{\beta}_j^{LS} - \frac{\lambda}{|x_j|^2}\right)$



Forward Selection, related to boosting