Ridge: $\|X\theta - Y\|_2^2 + \lambda\|\theta\|_2^2 \rightarrow \hat{\theta}_{ridge} = (X^TX + \lambda I)^{-1} X^TY$

Lasso: $\|X\theta - Y\|_2^2 + \lambda\|\theta\|_1 \rightarrow \hat{\theta}_{lasso} =$

$\|\theta\|_1 = \sum_{i=1}^{n} |\theta_i|$ → Induces sparsity → Easy to interpret.

→ In many cases, has few non-zero entries.

$\|\theta\|_2^2 = \sum_{i=1}^{n} \theta_i^2$

→ $d > n$ → OLS is not unique

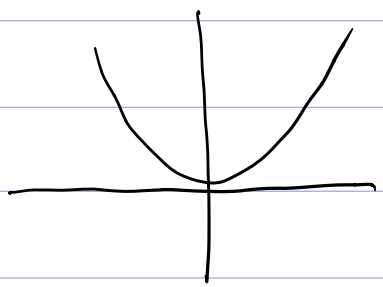
$Y = X\theta^* + \varepsilon$

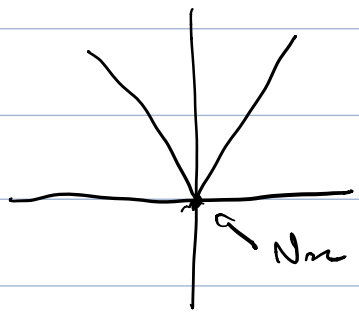
non-zero entries in $\theta^* = \|\theta^*\|_0 = s$ $\quad s \ll d, \ d \geq n.$

$\hat{\theta}_{OLS}$ $\hat{\theta}_{ridge}$ $\hat{\theta}_{lasso}$

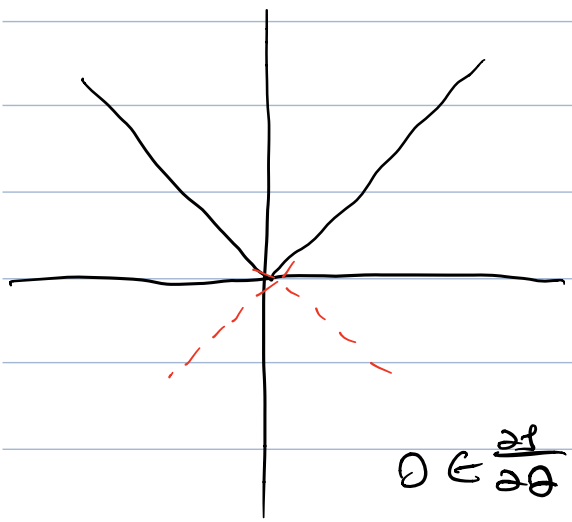
$\sum (y_i - X_i\theta)^2 + \lambda|\theta|$

$\underset{\theta}{\operatorname{argmin}} (y-\theta)^2 + \lambda|\theta|$





↖ Not differentiable at zero.



$$\frac{\partial|\theta|}{\partial\theta} \nearrow \text{Sub-Gradient} = \begin{cases} 1 & \theta > 0 \\ -1 & \theta < 0 \\ [-1,1] & \theta = 0 \end{cases}$$

$0 \in \frac{\partial f}{\partial \theta}$

Suppose $\theta^* > 0$

$2(\theta - y) + \lambda = 0$

$\Rightarrow \theta^* = y - \frac{\lambda}{2}$ → (If $y > \frac{\lambda}{2}$)

$\theta^* < 0$

$\Rightarrow 2(\theta - y) - \lambda = 0$

$\Rightarrow \theta^* = y + \frac{\lambda}{2} \quad (\text{if } y < -\frac{\lambda}{2})$

$\theta^* = 0$:

$\frac{\partial f}{\partial \theta} = 2(\theta - y) + \lambda\alpha = 0 \qquad \alpha \in [-1, 1].$

$= -2y + \lambda\alpha \qquad \alpha \in [-1, 1].$

$0 = -2y + \lambda\alpha \Rightarrow \alpha = \frac{2y}{\lambda}$

$\frac{2y}{\lambda} \in [-1, 1]$

$\Leftrightarrow \quad y \in [-\frac{\lambda}{2}, \frac{\lambda}{2}]$

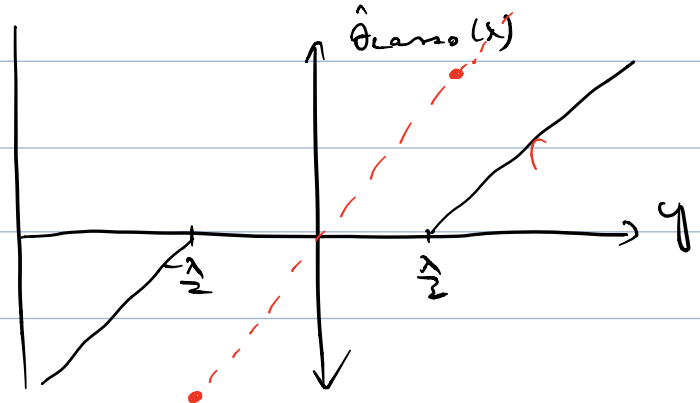
$\Rightarrow \quad \theta^* = 0$

$\underset{\theta}{\text{argmin}} (y-\theta)^2 + \lambda|\theta|$

$$\theta^* = \begin{cases} y - \frac{\lambda}{2} & \text{if } y > \lambda/2 \\ y + \frac{\lambda}{2} & \text{if } y < -\lambda/2 \\ 0 & \text{if } y \in [-\frac{\lambda}{2}, \frac{\lambda}{2}]. \end{cases}$$

Why?

$0 \in \frac{\partial f}{\partial \theta}\Big|_{\theta=\theta^*}$



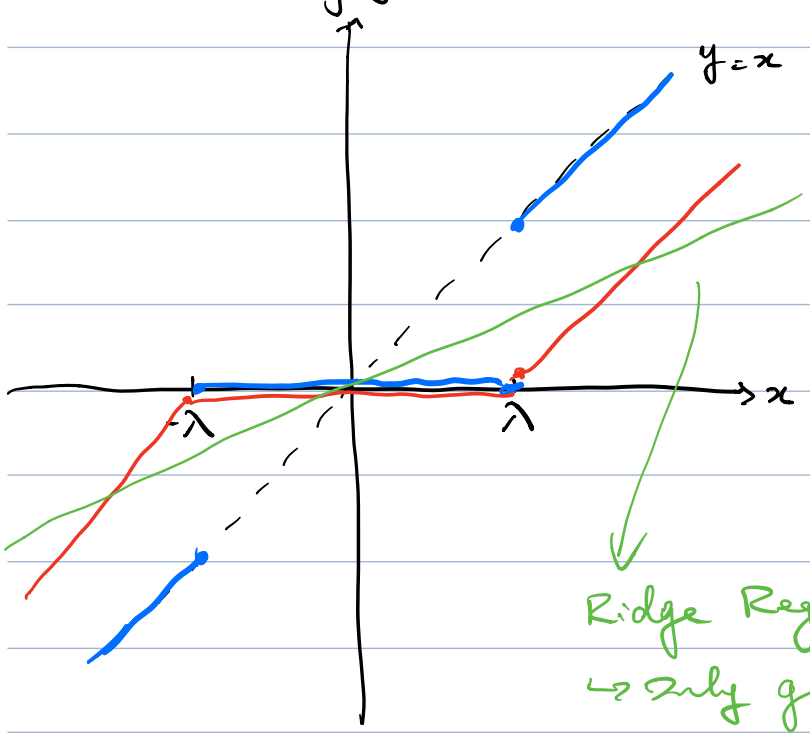
* Soft Thresholding

$\hat{\theta} = \text{sign}(y)\ (|y| - \frac{\lambda}{2})_+$

$(x)_+ = \begin{cases} x & x > 0 \\ 0 & o.w \end{cases}$

$\text{soft}(y, \frac{\lambda}{2})$

Soft Thresholding



Hard (x, λ)

Hard Thresholding

Soft (x, λ)

Soft Thresholding

- Continuous
- No sudden jump

Ridge Regression

↳ Only gives sparse solution as 0.

↳ Does not make it exactly zero.

↳ Only reduces magnitude.

$f(\theta) = (y - \theta)^2 + \lambda|\theta|$

$0 \in \frac{\partial f}{\partial \theta}$

$f(\theta) = \sum_{i=1}^{n} (x_i^T \theta - y_i)^2 + \lambda \sum_{j=1}^{d} (\theta_j)$

$= \|X\theta - Y\|_2^2 + \lambda \|\theta\|_1$

$\theta_i \rightarrow$ Variable

Fix other θ_j's

$\rightarrow$ Reduces to one-dimensional Lasso $\rightarrow$ closed form solution as shown before.

Coordinate-descent [like iterative algorithm]

Each iteration step: new θ_1

1. $R(\beta) = \|X\theta - Y\|_2^2$

$R(\beta) = R(\beta^{ols}) + (\beta^{ols} - \beta)^T \frac{X^TX}{n} (\beta^{ols} - \beta)$

$R(\hat{\beta})$

LHS: $\frac{\|X\beta - X\hat{\beta} + X\hat{\beta} - Y\|_2^2}{n} = \frac{\|X\beta - X\hat{\beta}\|_2^2}{n} + \frac{\|X\beta - X\hat{\beta}\|_2^2}{n}$

$+ 2 \frac{(X\beta - X\hat{\beta})^T (X\hat{\beta} - Y)}{n}$

$v^T A v = (-v)^T A (-v)$

$= \|X(\beta - \hat{\beta})\|$

$= (-\beta + \hat{\beta})^T X^T X (-\beta + \hat{\beta})$

$\hookrightarrow (\beta - \hat{\beta})^T X^T (X\hat{\beta} - Y)$ $\quad X^TY = X^TY$

0

$X^T(X\beta - Y) = 0$

$R(\beta) \geq R(\hat{\beta}_{ols}) \quad \forall \beta$

$\rightarrow \hat{\beta}_{ols}$ is the minimizer of training error

$R(\beta) = R(\hat{\beta}) + c$

$\Leftrightarrow \{ (\beta - \hat{\beta})^T \frac{X^TX}{n} (\beta - \hat{\beta}) = c \}$

Quadratic Form:

$(v - \hat{v})^T \cdot A (v - \hat{v}) = c$

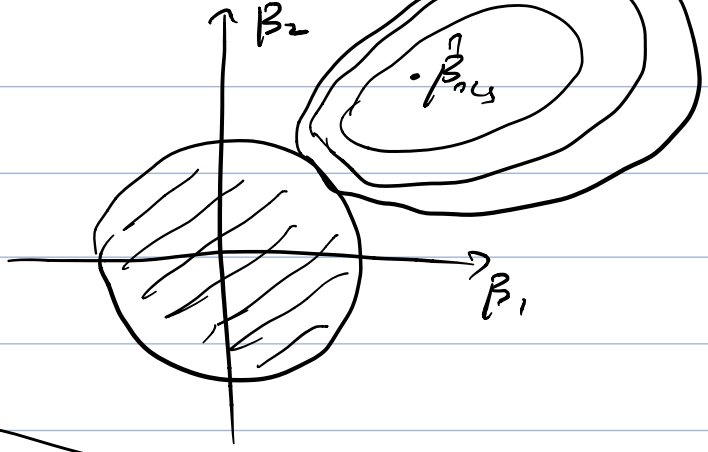
$\frac{X^TX}{n} = I \Rightarrow \|\beta - \hat{\beta}\|_2^2 = c$

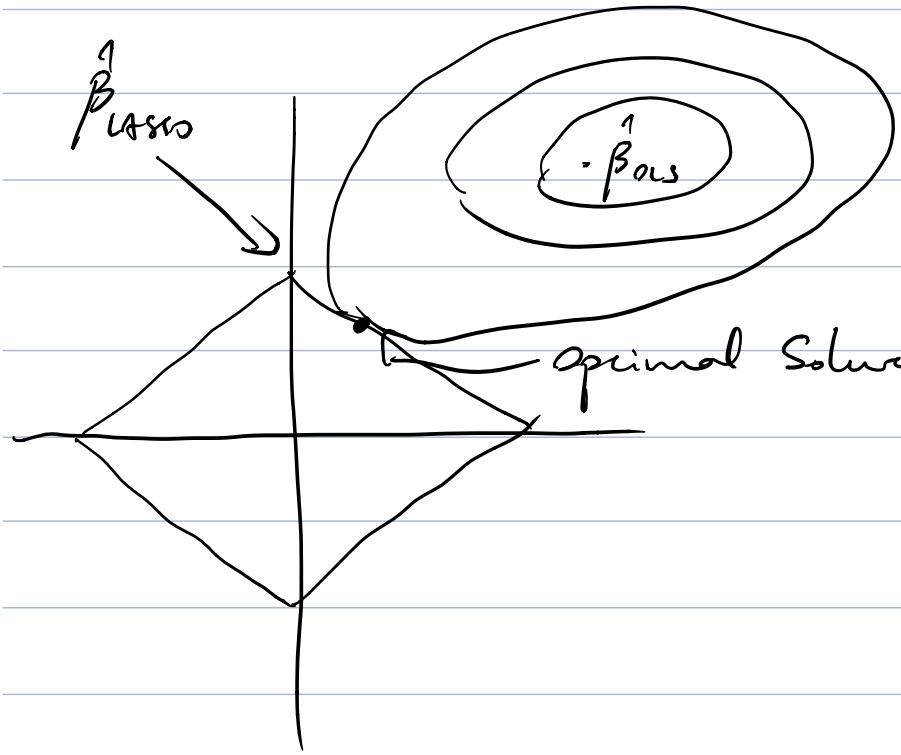
$v^TAv = \frac{v^T(A + A^T)v}{2}$

$|\beta_1| + |\beta_2| \le \mu$

$\|X\beta - Y\|_2^2 + \lambda \|\beta\|_2^2$

$\min \|X\beta - Y\|_2^2$ s.t. $\|\beta\|_1 \le \mu$





4. $\|X\beta - Y\|_2^2 + \lambda \|\beta\|_1$

Show $\beta = 0$ is the optimal Lasso estimator if $\lambda \ge \|X\theta - y\|_\infty$

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

$\lambda \to \infty \Rightarrow \hat{\beta}_\lambda = 0$

$\lambda > \|X\theta - Y\|_\infty \Rightarrow \hat{\beta}_{(\text{LASSO})} = 0$

$f(0) < f(\beta)$ for any $\beta \neq 0$,

Then $\beta = 0$ is the optimal estimator.

RHS

$\|Y\|_2^2 < \underbrace{\|X\beta - Y\|_2^2 + \lambda \|\beta\|_2}$

$\underbrace{\|X\beta\|_2^2 + \|Y\|_2^2 - 2\beta^T X^T Y + \lambda \|\beta\|_1}_{\ge 0} = RHS$

$\ge$

$$RHS \geq \|Y\|_2^2 + \lambda\|\beta\|_1 - 2\beta^T X^T Y$$

$$|\beta^T v| \leq \|\beta\|_1 \|v\|$$

$$LHS = \left|\sum_{i=1}^{d} \beta_i v_j\right| \qquad \text{Holder's Inequality}$$

$$\leq \sum_{i=1}^{d} |\beta_i||v_j| \qquad \text{Young's Inequality}$$

$$\leq \left(\sum_{i=1}^{d} |\beta_i|\right) \max_j (v_j) \qquad \text{Cauchy Schwarz.}$$

$$= \|\beta\|_1 \|v\|_\infty$$

$$RHS \geq \|Y\|_2^2 + \|\beta\|_1 (\lambda - 2\|X^T Y\|_\infty)$$

$$> \|Y\|_2^2$$

If $\lambda > 2\|X^T Y\|_\infty$

$\Rightarrow \hat{\beta}_\lambda(\text{Lasso}) = 0$

4 & 5

$$(X_1\theta_1 + X_2\theta_2 + X_2^T\theta_{-2} - Y)^2 + \lambda(\theta^2 + \theta_2^2) + \lambda\|\theta_{-2}\|_2^2$$

$$(X_1\theta_1 + X_2\theta_2 + X_2^T\theta_{-2} - y)^2 + \lambda(|\theta_1| + |\theta|_2) + \lambda\|\theta_{-2}\|_2^2$$

$$= (X(\theta_1 + \theta_2) + \tilde{X}^T\tilde{\theta} - y)^2 + \lambda(\theta_1^2 + \theta_2^2) + \lambda\|\tilde{\theta}\|_2^2$$

$$= (X(\theta_1 + \theta_2) + \tilde{X}^T\tilde{\theta} - y)^2 + \lambda(|\theta_1| + |\theta_2|) + \lambda\|\tilde{\theta}\|_2^2$$

$\Rightarrow \theta_1 = \frac{\theta^*}{2}$ $\theta_2 = \frac{\theta_1^*}{2}$ is the unique solution

$\Rightarrow \theta_1 = \theta_1^*$ $\theta_2 = 0$

$\theta_1 = \theta^*$ $\theta_2 = \theta_1^*$

& Many More

} Non-Unique but sparse solutions possible.

&

* Lasso can give sparse solutions even with collinear/correlated features but Ridge can't.

(P1) $\|X\beta - Y\|_2^2 + \lambda\|\beta\|$ $\longrightarrow$ $\hat{\beta}(\lambda)$ be the solution for problem 1

(P2) $\|X\beta - Y\|_2^2$ s.t. $\|\beta\|_1 \le \mu$ /

↪ then we want to find a μ s.t. for P_2, $\hat{\beta}(\lambda)$ is optimal

$\|X\hat{\beta} - Y\|_2^2 \le \|X\beta - Y\|_2^2$ for any $\|\beta_1\| \le \mu$.

Claim: $\mu = \|\hat{\beta}(\lambda)\|$

Proof: We know $\|\hat{\beta}(\lambda)\|$ is optimal for P_1

↪ $\|X\hat{\beta} - Y\|_2^2 + \lambda\|\hat{\beta}\|_2 \le \|X\beta - Y\|_2^2 + \lambda\|\beta\|_1 \quad \forall \beta \in$

for $\|\beta\|_1 \le \mu = \|\hat{\beta}\|$ $\le \|X\beta - Y\|_2^2 \quad \forall \|\beta_1\| \le \mu$

because $\tau_2 \ge 0$.

↑ Optimality Condition for P_2

Kernel Ridge Regression.

$\hat{\beta}_\lambda(\text{ridge}) = \underbrace{(X^TX + \lambda I_n)^{-1}}_{d \times d} X^TY$ preferable if $n \gg d$

$= X^T\underbrace{(XX^T + \lambda I_n)^{-1}}_{n \times n}Y$ preferable if $n \ll d$

$$XX^T = \begin{bmatrix} x_1^Tx_1 & x_1^Tx_2 & \cdots \\ & & \end{bmatrix} \qquad X = \begin{bmatrix} — x_1^T — \\ \vdots \\ — x_n^T — \end{bmatrix}$$

$$XX^T = \sum_{i,j}^{n} x_i x_j^T$$