Lecture 14: LASSO

Least-Absolute Shrinkage and Selection Operator

Variable selection is like forcing some of my coefs $\hat{\beta}$s to be zero

e.g.

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots$$

force $\hat{\beta}_2 = 0$

$$= \hat{\beta}_0 + \hat{\beta}_1 X_1$$

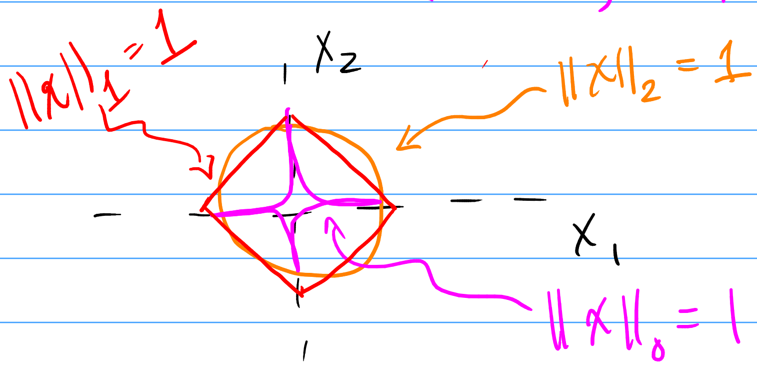
selected out X_2

Ideally I could solve the problem

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\, L(\beta) \text{ s.t. } \underbrace{\|\beta\|_0 \le t}_{\text{only use at most } t \text{ vars.}}$$

best subset selection problem

Problem: optimizing under a $\|\cdot\|_0$ constraint is often difficult/intractable b/c $\|\cdot\|_0$ isn't nice (convex, differentiable)



To make this work-able, work w/ 1-norm instead

$\|\cdot\|_1$ is the convex-relaxation of $\|\cdot\|_0$

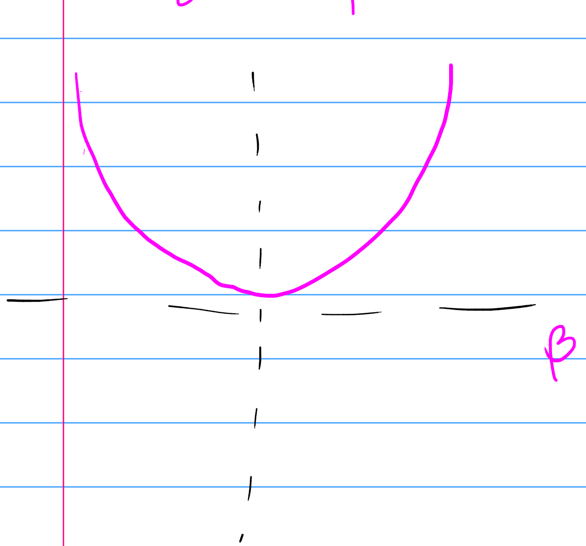
LASSO: Solves

(1) $$\hat{\beta}^{(LASSO)} = \underset{\beta}{\operatorname{argmin}}\, L(\beta) \text{ s.t. } \|\beta\|_1 \le t$$

(2) $$\hat{\beta}^{(LASSO)} = \underset{\beta}{\operatorname{argmin}}\, L(\beta) + \lambda \|\beta\|_1$$

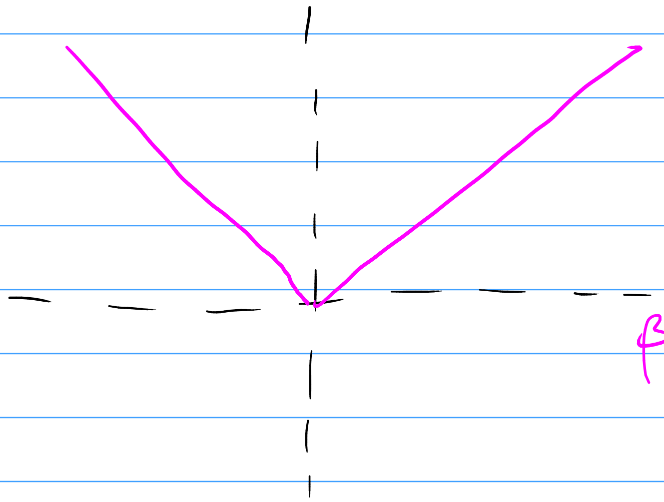
$\lambda \geq 0$

ridge: $\|\beta\|_2^2$



- convex
- differentiable

LASSO: $\|\beta\|_1$



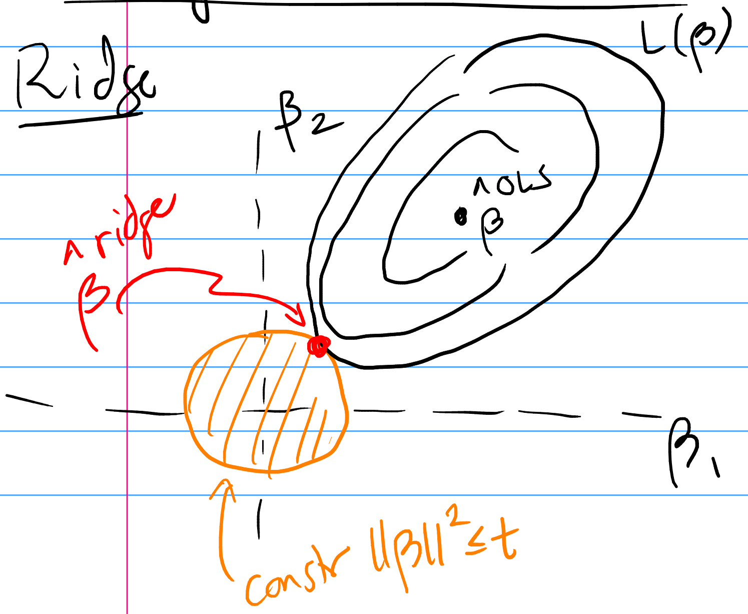
- convex
- not diff'able

because $\|\beta\|_1$ isn't diff'able, no closed form soln for $\hat{\beta}^{(LASSO)}$

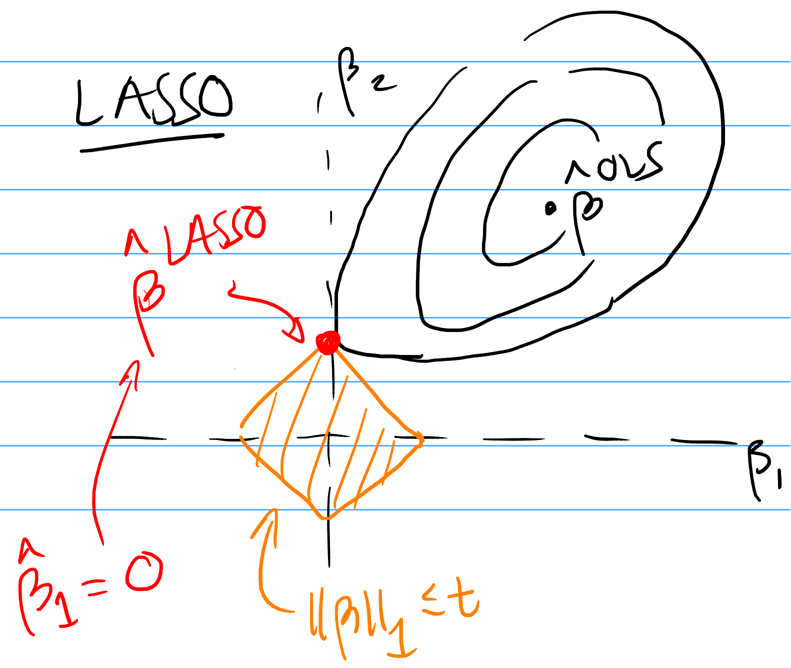
Need numerical methods to solve for $\hat{\beta}^{(LASSO)}$

Why use LASSO?

Ridge



LASSO



Often, $\hat{\beta}^{LASSO}$ is at vertex of constraint which zeroes out one or more $\hat{\beta}$ components - i.e. it selects out those vars.

Comparison: Assume X is orthogonal

1. Variable Selection (Hard-Thresholding)

$$\hat{\beta}_i^{HS} = \begin{cases} \hat{\beta}_i^{OLS} & \text{if } |\hat{\beta}_i^{OLS}| \geq t \\ 0 & \text{else} \end{cases}$$

2. Ridge:

$$\hat{\beta}_i^{ridge} = \frac{\hat{\beta}_i^{OLS}}{1+\lambda}$$

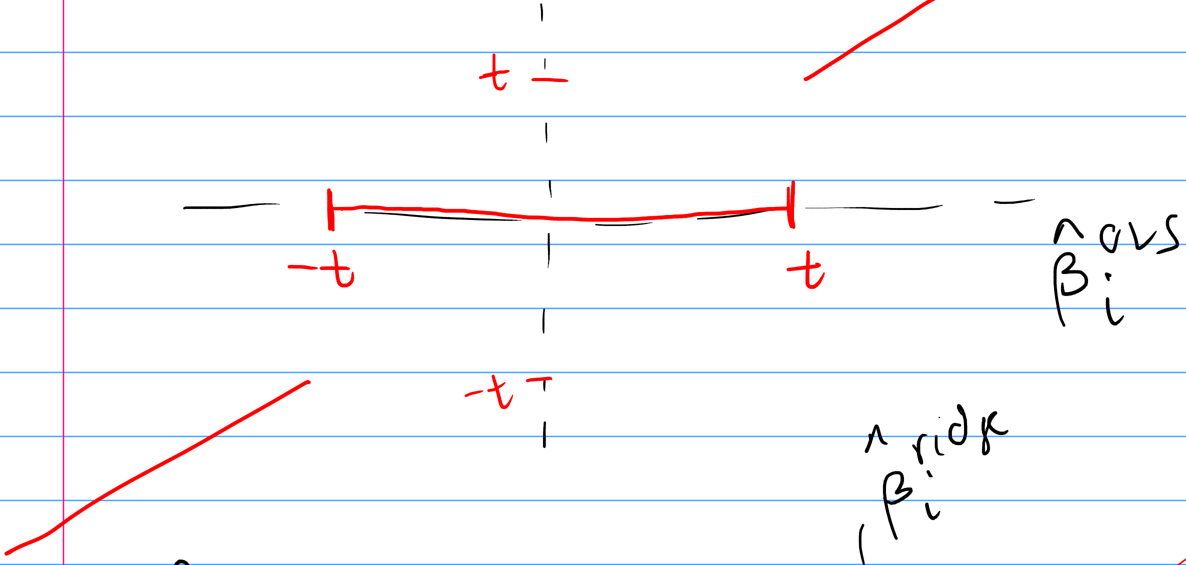
(proportional shrinkage)

3. LASSO: (soft thresholding)

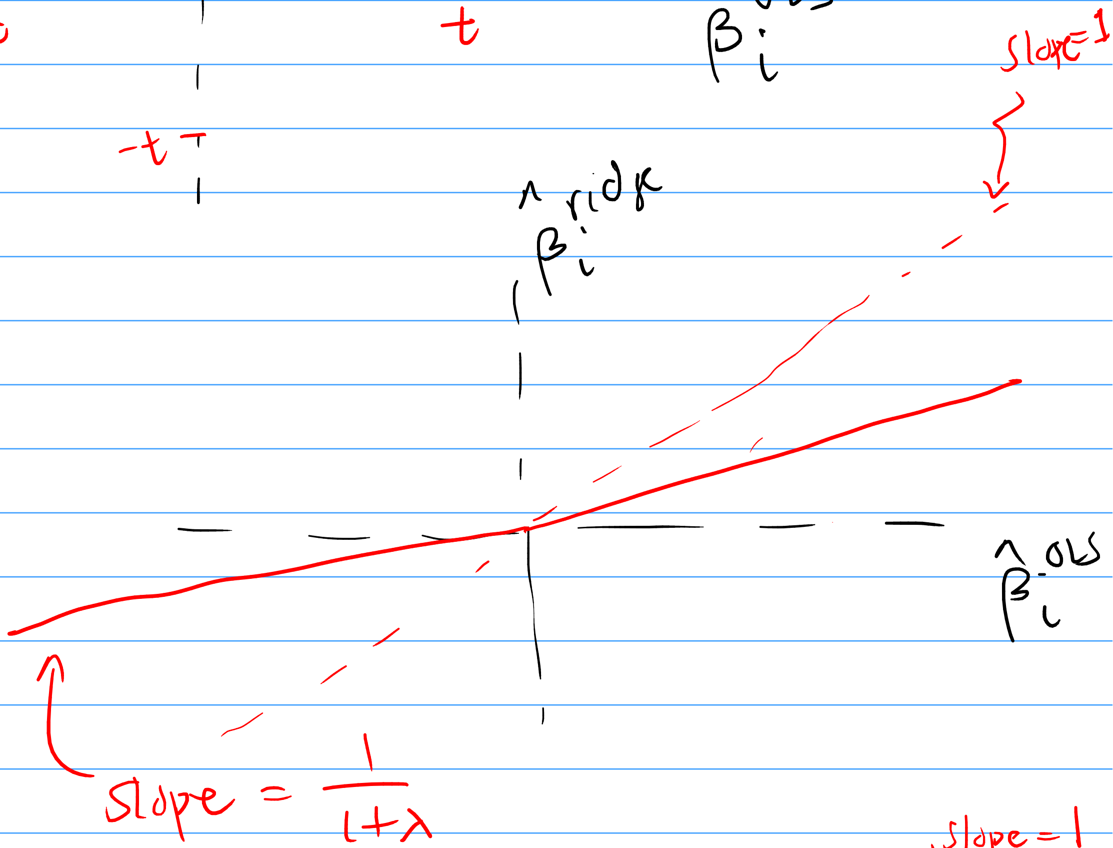
$$\hat{\beta}_i^{LASSO} = \begin{cases} \text{sign}(\hat{\beta}_i^{OLS})\left[|\hat{\beta}_i^{OLS}| - \lambda\right] & \text{if } |\hat{\beta}_i^{OLS}| \geq \lambda \\ 0 & \text{else} \end{cases}$$

$$= \text{sign}(\hat{\beta}_i^{OLS})\left[|\hat{\beta}_i^{OLS}| - \lambda\right]_+$$

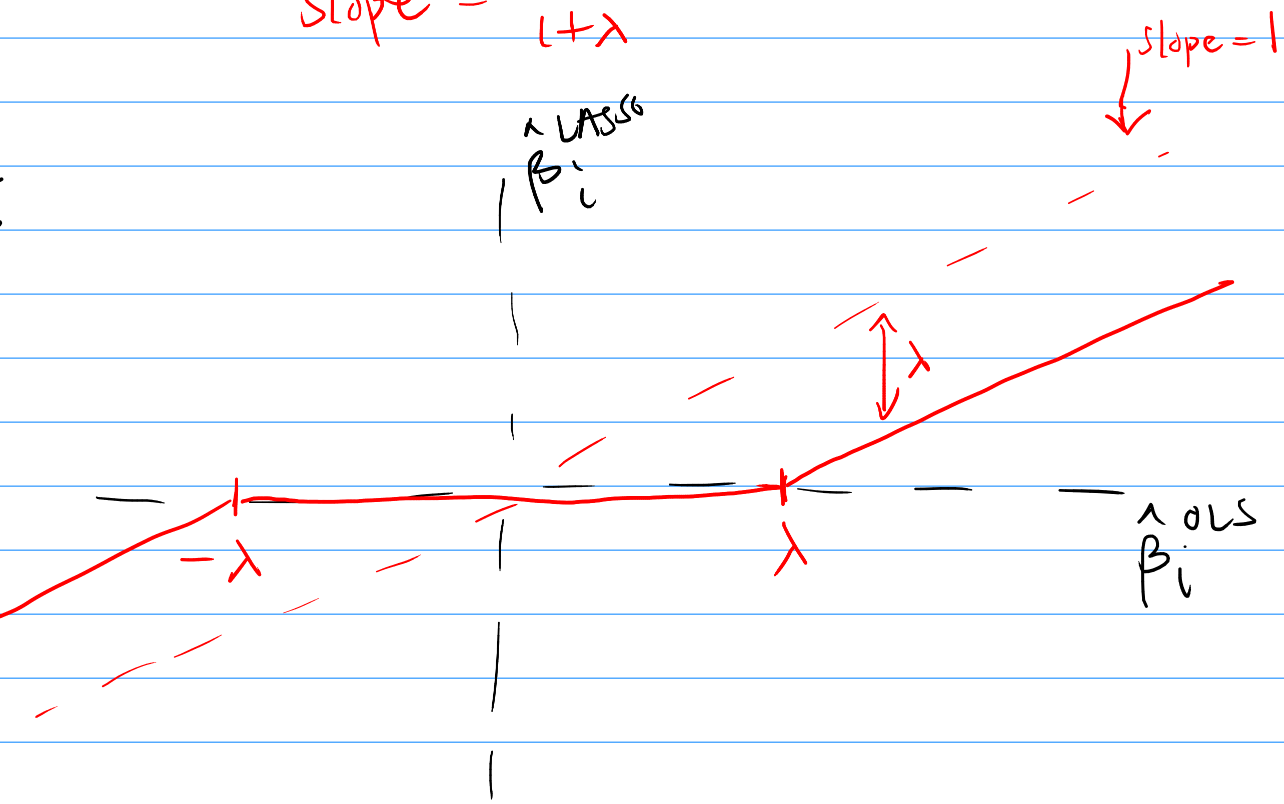
Hard Thresholding



Proportional Shrinkage



LASSO:



Elastic Net: L2 + L1 penalty

$$\hat{\beta}^{EN} = \underset{\beta}{\text{argmin}}\ L(\beta) + \lambda\left[\frac{(1-\alpha)}{2}\|\beta\|_2^2 + \alpha\|\beta\|_1\right]$$

$\lambda \geq 0$ = overall penalty strength

$\alpha \in [0, 1]$ = tradeoff btwn L1 and L2 penalties

$\alpha = 0 \Rightarrow$ Ridge
$\alpha = 1 \Rightarrow$ LASSO