Last time: Ridge Regression

OLS:

$$\hat{\beta}^{(ols)} = \underset{\beta}{\operatorname{argmin}}\, L(\beta)$$

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

$$= (X^T X)^{-1} X^T Y$$

Ridge: Shrinkage Estimator

(1) $$\hat{\beta}^{(ridge)} = \underset{\beta}{\operatorname{argmin}}\, L(\beta) + \lambda \|\beta\|_2^2$$

$\lambda \geq 0$

(2) $$\hat{\beta}^{(ridge)} = \underset{\beta}{\operatorname{argmin}}\, L(\beta) \quad \text{s.t.} \quad \|\beta\|_2^2 \leq t$$

corresp. t

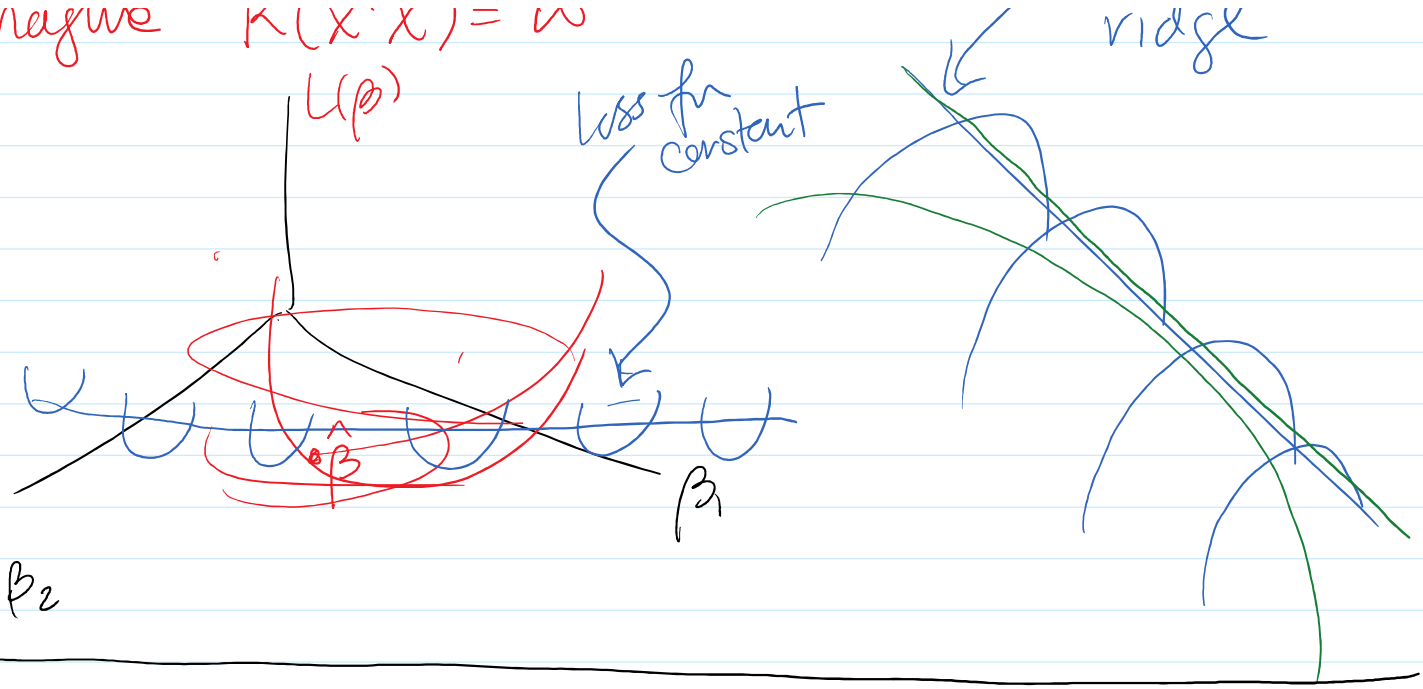
(3) $$\hat{\beta}^{(ridge)} = (X^T X + \lambda I)^{-1} X^T Y$$

Why called ridge regression

Imagine $K(X^T X) = b_0$

$L(\beta)$ ridge

Imagine $K(X, X) = w$



Why do variable selection

→ highly correlated vars (ridge)

→ interpretibility (look at this) (LASSO)

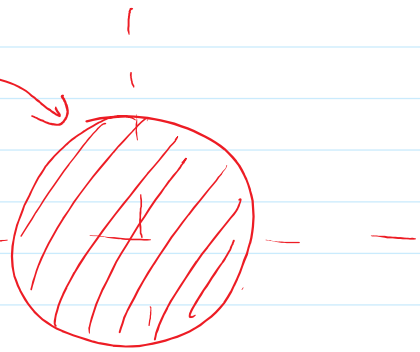
Aside: Norms

Ex. euclidean norm

$x \in \mathbb{R}^p$ then $\|x\| = \sqrt{\sum_{i=1}^{p} x_i^2}$

Consider

$\{y \mid \|y\| \leq 1\}$



Can generalize: Let $q \geq 0$ define the q-norm as

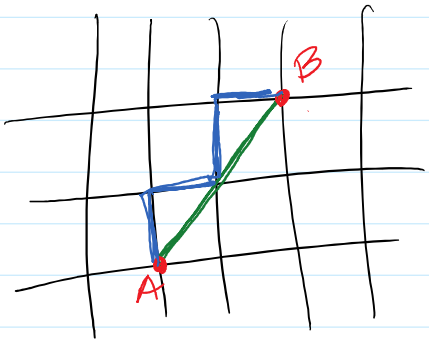
$$\|x\|_q = \left(\sum_{i=1}^{p} |x_i|^q \right)^{1/q}$$

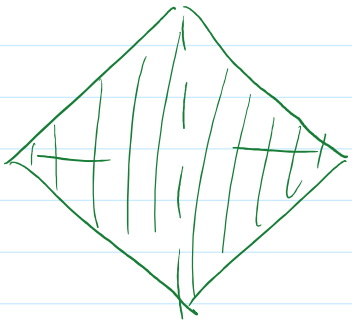
when $q = 2$ then $\|x\|_2$ is euclidean norm

$q = 1$ then $\|x\|_1$ is called the Manhattan norm (1-norm)

$$\{y \mid \|y\|_1 \leq 1\}$$

$$\|x\|_1 = \sum_{i=1}^{p} |x_i|$$



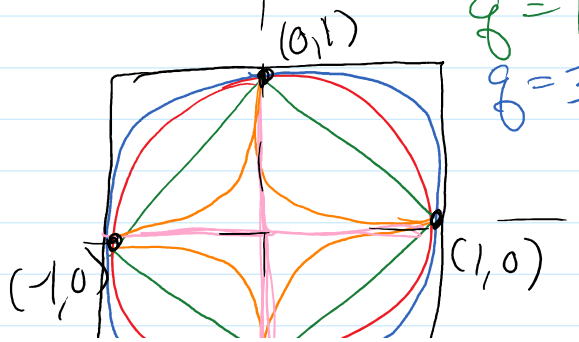


$d_1(A,B) = \|A-B\|_1$

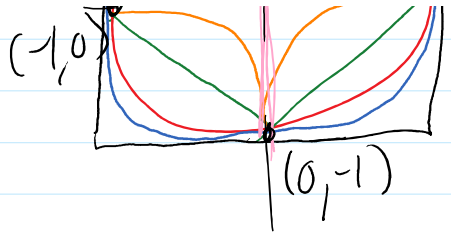
$d_2(A,B) = \|A-B\|_2$

$$\|x\|_\infty = \lim_{q \to \infty} \|x\|_q$$

$$= \max |x_i|$$



$= \max_i |x_i|$



$q = \infty$

$\|x\|_0 = \lim_{q \to 0} \|x\|_q = \#\{x_i \neq 0\}$

$q = 1/2$

↑ # of elements of x non-zero

$\|(0,1)\|_0 = 1 \quad \|(1/2, 1/2)\| = 2$ etc.

$\|(0,0)\| = 0$

LASSO: Least-Absolute Shrinkage and Selection Operator

Variable selection is like zeroing out entries of β

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

↑ free to zero = X_2 drop from model

Want:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\, L(\beta) \quad \text{s.t.} \quad \|\beta\|_0 \le t$$

$\|\beta\|_0 \le t$: # of non-zero elements of $\beta \le t$

Unfortunately, this $\|\cdot\|_0$ not convex

this is intractible problem.

Can't do this directly.

Relax the problem (convex relaxation)

using $q = 1$

LASSO:

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

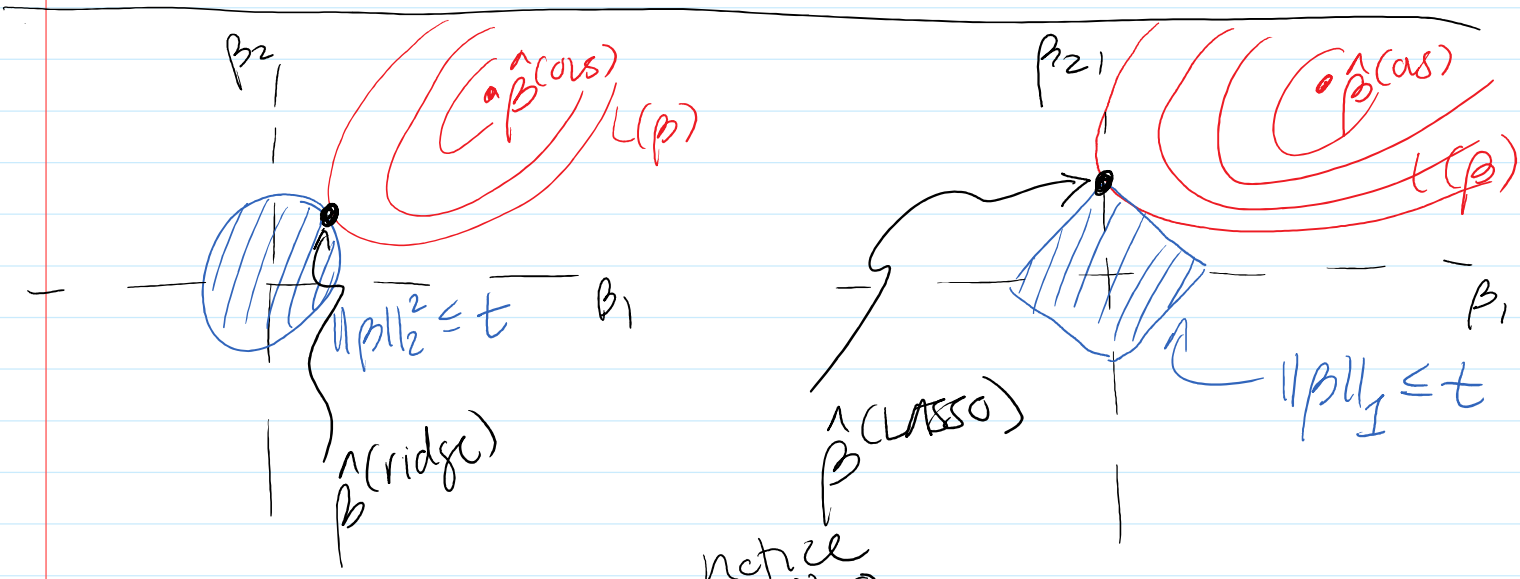
(2) $\hat{\beta}^{(LASSO)} = \underset{\beta}{\operatorname{argmin}}\, L(\beta) + \lambda \|\beta\|_1$

(3) amazing fact: $\|\beta\|_1$ continuous, convex, not differentiable
So no analytic solution

think of:

ridge: x^2 v. LASSO: $|x|$

<u>punchline</u>: no simple formula for $\hat{\beta}^{(LASSO)}$
need to numerically solve w/ convex optim.



Ridge LASSO

Ridge $\hat{\beta}_i^{(LASSO)} = 0$ LASSO

Comparison: Assume X is orthogonal (vars uncorrelate)

(1) Subset selection (Hard thresholding)

$$\hat{\beta}_i^{(HS)} = \hat{\beta}_i^{(ols)} \mathbb{1}(|\hat{\beta}_i^{(ols)}| \geq t) = \begin{cases} \hat{\beta}_i & \text{if large enough} \\ 0 & \text{else} \end{cases}$$

(2) Ridge:

$$\hat{\beta}^{(ridge)} = \frac{\hat{\beta}_i^{(ols)}}{1+\lambda}$$ (proportional shrinkage)

(3) LASSO:

soft thresholding

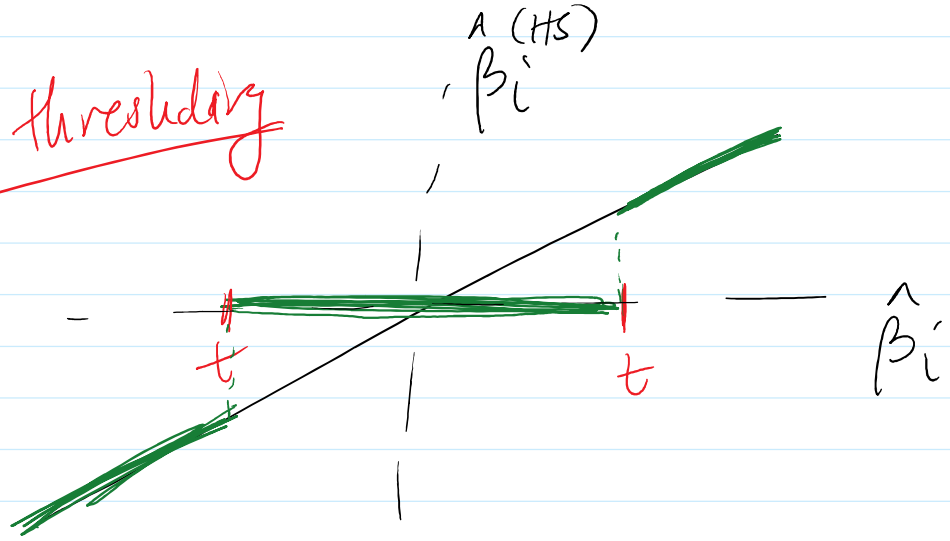
$$\hat{\beta}_i^{(LASSO)} = \text{Sign}(\hat{\beta}_i^{(ols)}) \max(|\hat{\beta}_i^{(ols)}| - \lambda, 0)$$

$$= \begin{cases} \hat{\beta}^{(ols)} - \lambda & \text{if this is} \geq 0 \end{cases}$$

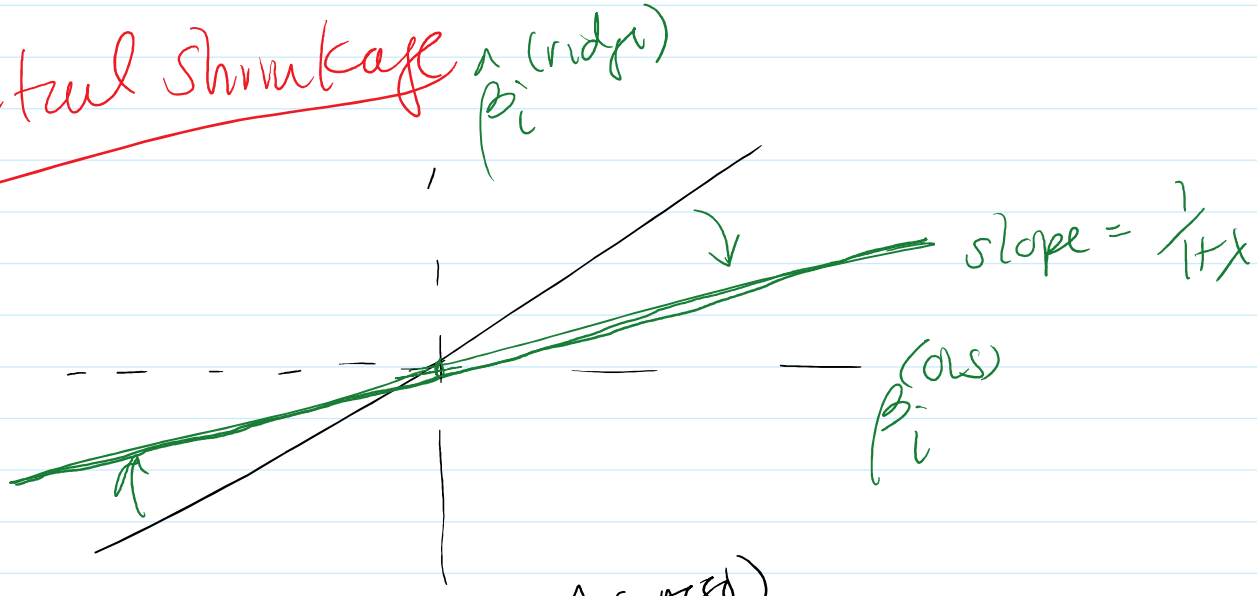
$$= \begin{cases} \hat{\beta}^{(OLS)} - \lambda & \text{if this is} \geq 0 \\ 0 & \text{else} \end{cases}$$

for $\hat{\beta}^{(OLS)} > 0$

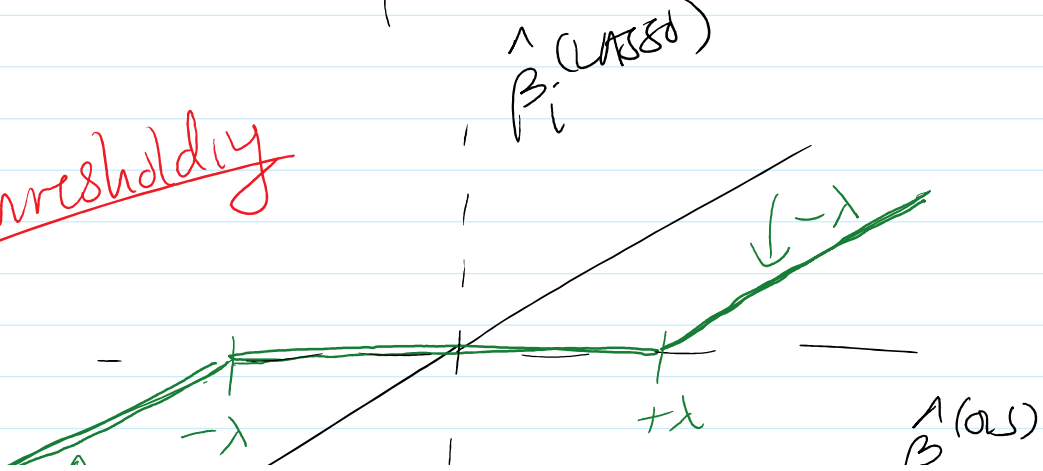
Hard thresholding

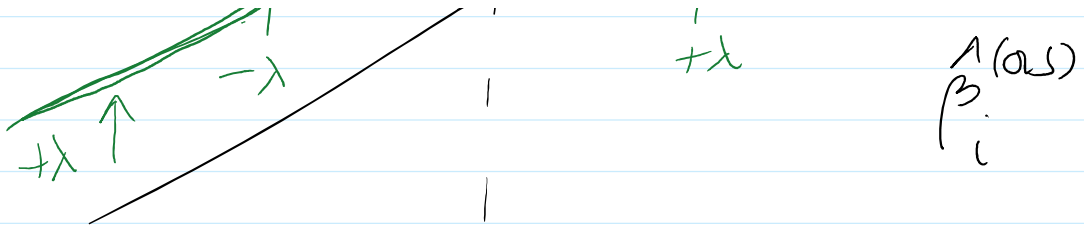


Proportional shrinkage



Soft thresholding





---

(1) Can combine ridge/LASSO

Elastic Net $\quad 0 \le \alpha \le 1$

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

$\alpha = 0 \Rightarrow$ ridge

$\alpha = 1 \Rightarrow$ LASSO

(2) Can do penalization w/ any method.

→ penalize any LOSS

→ Ex. penalized Logistic regression

→ ...
