# Lecture 16

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

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

↑ how many non-zero elements of $x$

---

LASSO : Least-Absolute Shrinkage and Selection Operator

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

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

force $\hat{\beta}_1 = 0$ then I effectively select out $X_1$

Consider the constrained optim problem

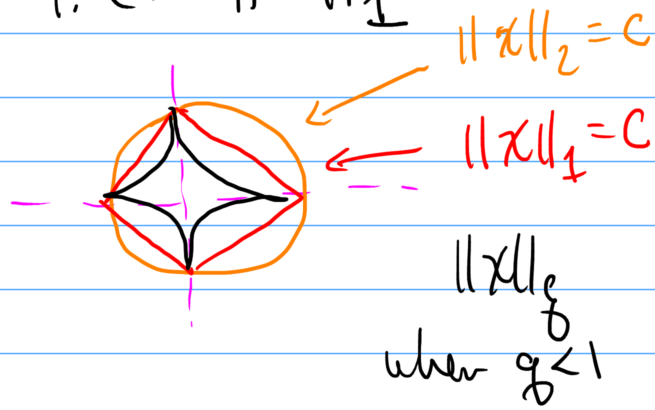
$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\, L(\beta) \quad \text{s.t. } \|\beta\|_0 \leq t$$

build best model w/ $\leq t$ variables

<u>Problem</u>: generally optimizing using a $\|\cdot\|_0$ constraint is difficult / intractable

not convex

So instead, we can work w/ the convex-relaxation of $\|\cdot\|_0$ i.e. $\|\cdot\|_1$

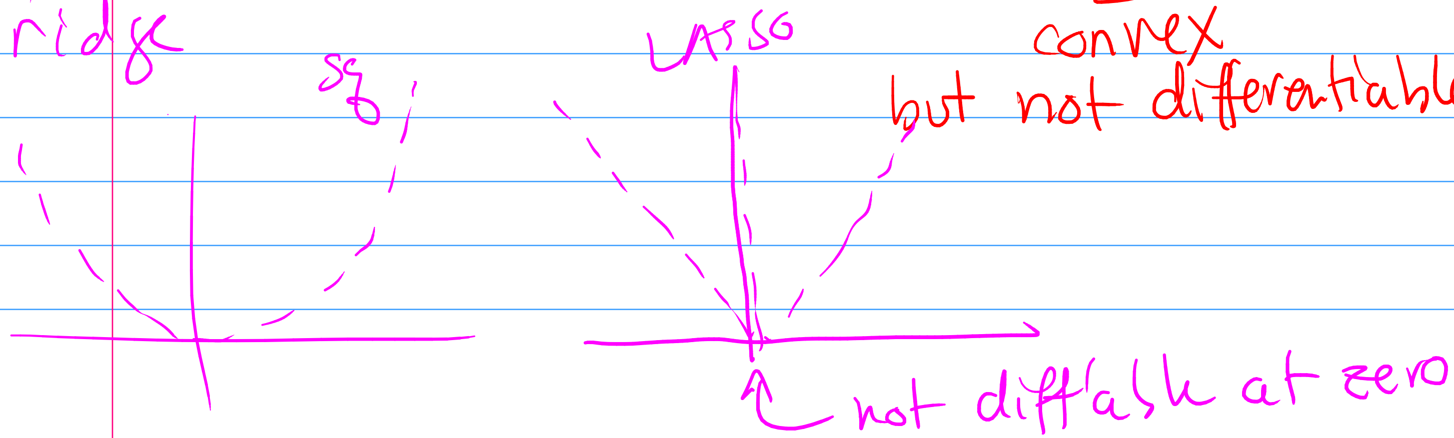



<u>LASSO</u>:

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

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

convex but not differentiable

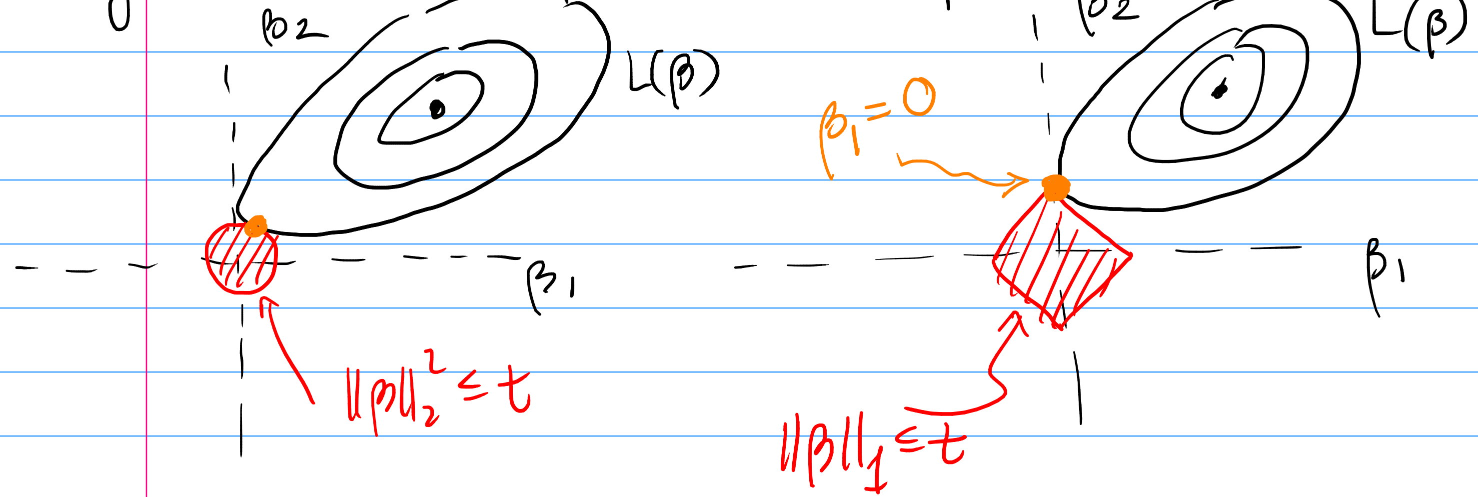


b/c $\|\beta\|_1$ isn't diff'able there is no closed-form soln for $\hat{\beta}^{LASSO}$

So need to use some numerical method.

---

Why do we want to use $\|\beta\|_1$?




Ridge  LASSO

LASSO tends to force constr. optimum at points of constraint i.e. zero out elements of $\beta$

Comparison  Assume that $X$ is orthogonal.

① Variable Selection (Hard-thresholding)

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

② Ridge:

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

③ LASSO:  soft thresholding

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

---

Elastic Net : L1 + L2

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

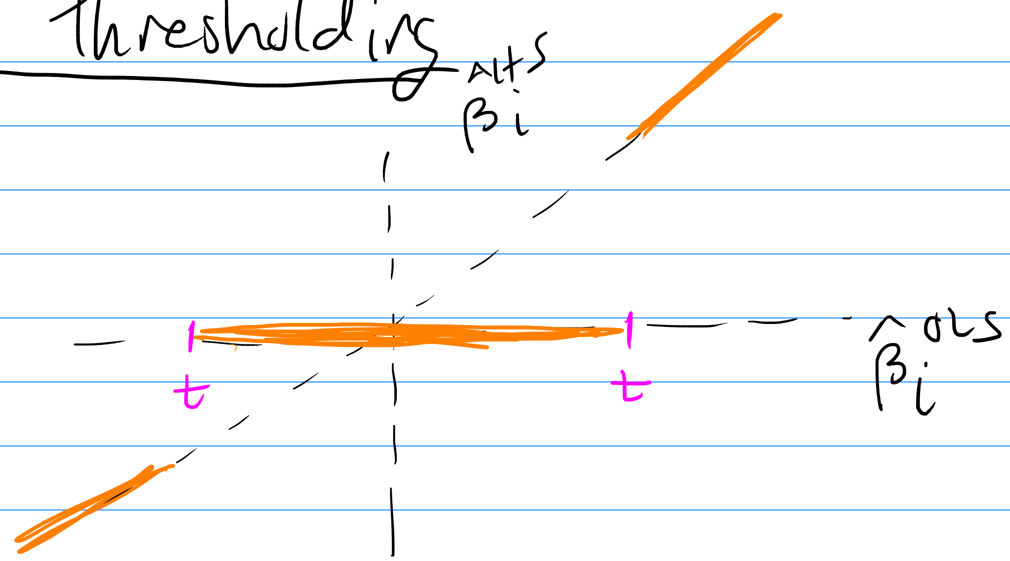
$0 \leq \alpha \leq 1$

tradeoff between L1 / L2 penalty

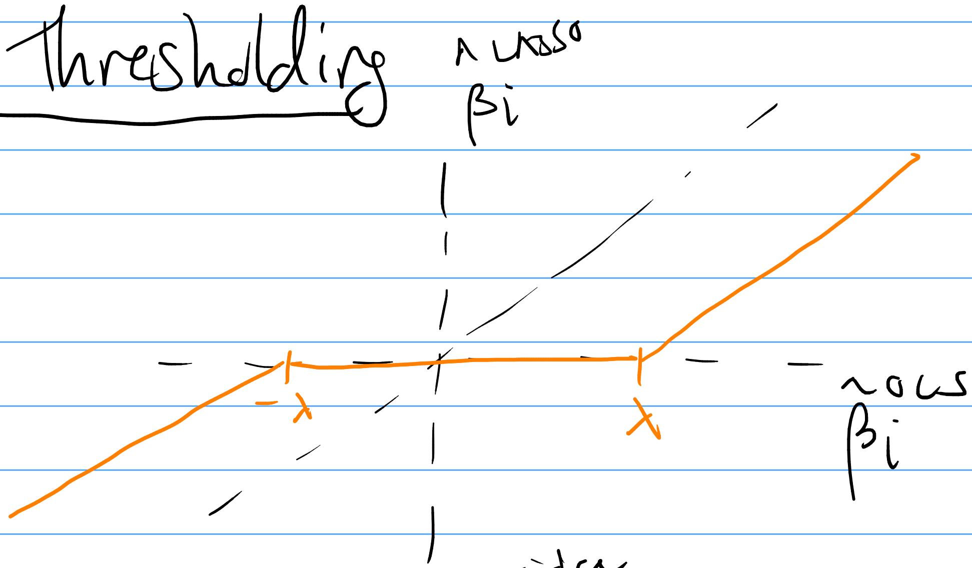
$\alpha = 0 \Rightarrow$ ridge

$\alpha = 1 \Rightarrow$ LASSO

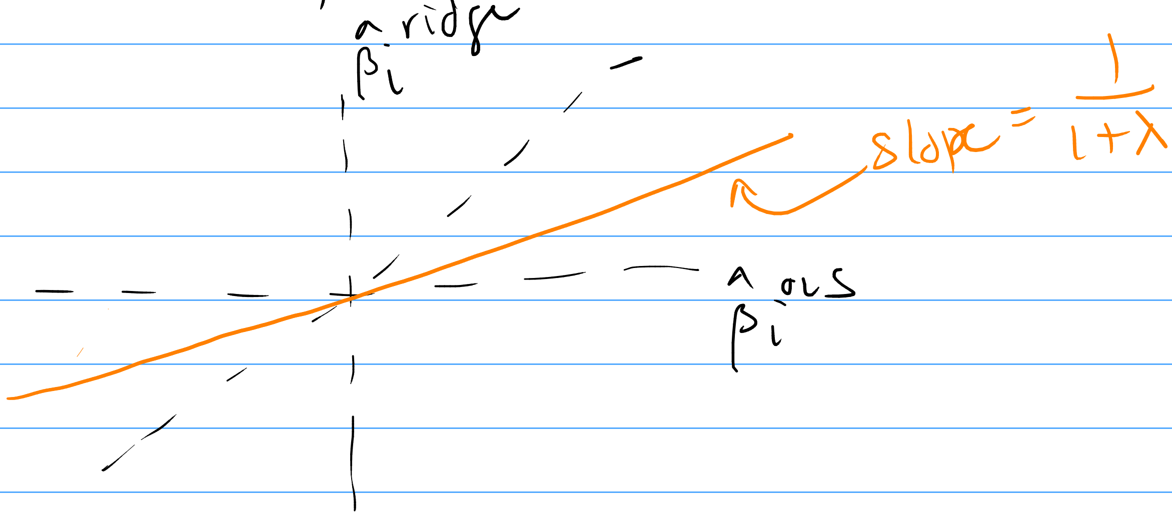
Hard thresholding




Soft thresholding

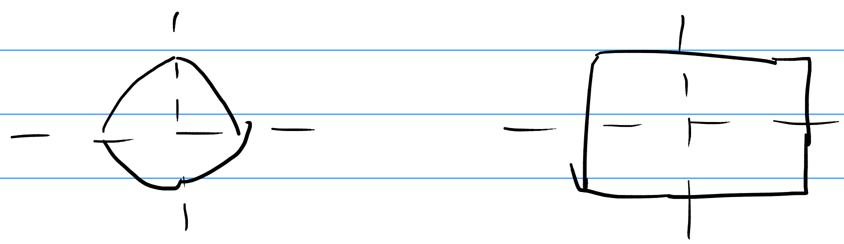



Ridge:




---

$q = 1.5$



Can do this w/ all sorts of other methods.

Ex. Logistic Regression

$$y_n \mid \underline{X}_n = \underline{x}_n \sim \text{Bern}(p(\underline{x}_n))$$

$$p(\underline{x}_n) = \text{logistic}(X\beta)$$ like regr.

$$\ell(\beta) = \text{(neg.) log. like. fn}$$

$$\hat{\beta} = \underset{\beta}{\text{argmin}}\ \ell(\beta) + \lambda\|\beta\|^2$$

or $+ \lambda\|\beta\|_1$

---

More generally,

$$\hat{f} = \underset{f \in \mathcal{F}}{\text{argmin}}\ L(f)$$

Could consider a penalized version

$$\hat{f} = \underset{f \in \mathcal{F}}{\text{argmin}}\ \underbrace{L(f)}_{\text{data term}} + \lambda \underbrace{J(f)}_{\text{penalty term}}$$

penalty fn for f that measures complexity