

## Lecture 5. Coordinate Descent

Tuesday, April 24, 2018 6:23 AM

$$\underset{\beta \in \mathbb{R}^d}{\text{Min}} \quad \frac{1}{n} \|y - X^\top \beta\|_2^2 + \lambda \|\beta\|_1$$

Least-squares

$$\frac{1}{n} \sum_{i=1}^n (y_i - x_i^\top \beta)^2$$

$\ell_1$ -norm penalty

$$\|\beta\|_1 = \sum_{j=1}^d |\beta_j|$$

$\beta^*$  is sparse

$$\beta = [0, 0, 3, 0, -2]^\top$$

$$\beta = [0, 0, 3, 0, -2]'$$

$$\beta = \begin{pmatrix} 0 \\ 0 \\ 3 \\ 0 \\ -2 \end{pmatrix}$$

$$x^T = (x^1, x^2, x^3, x^4, x^5)$$

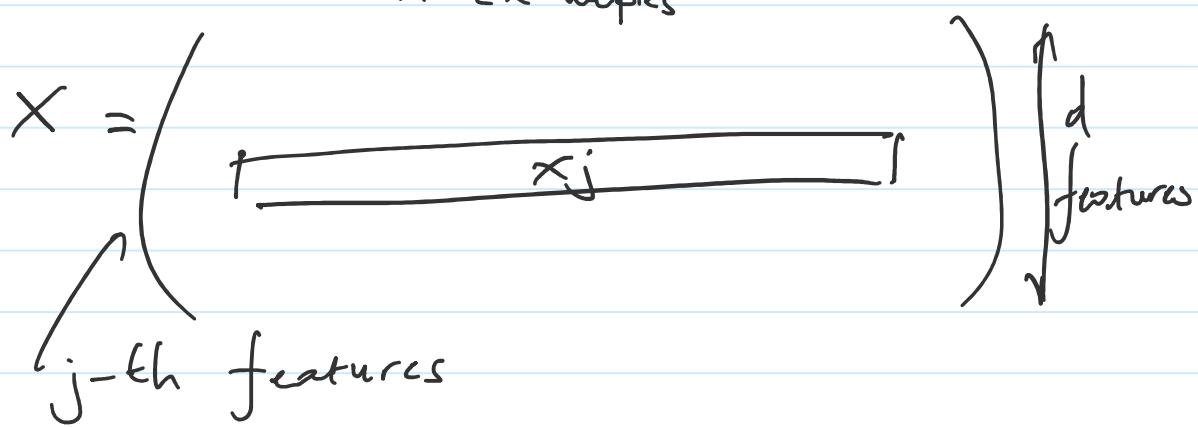
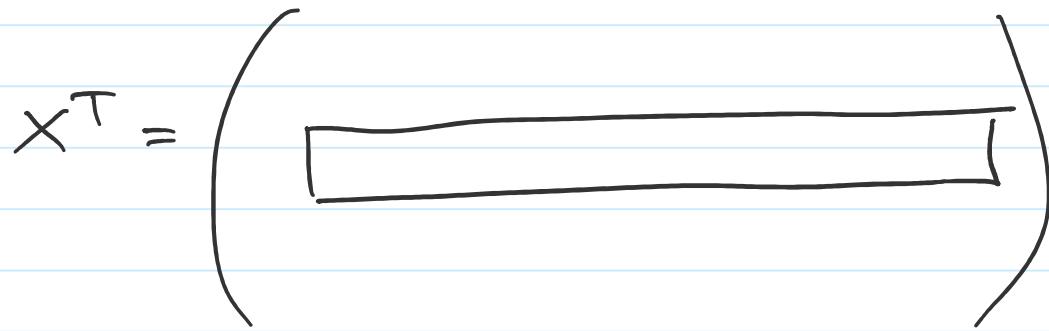
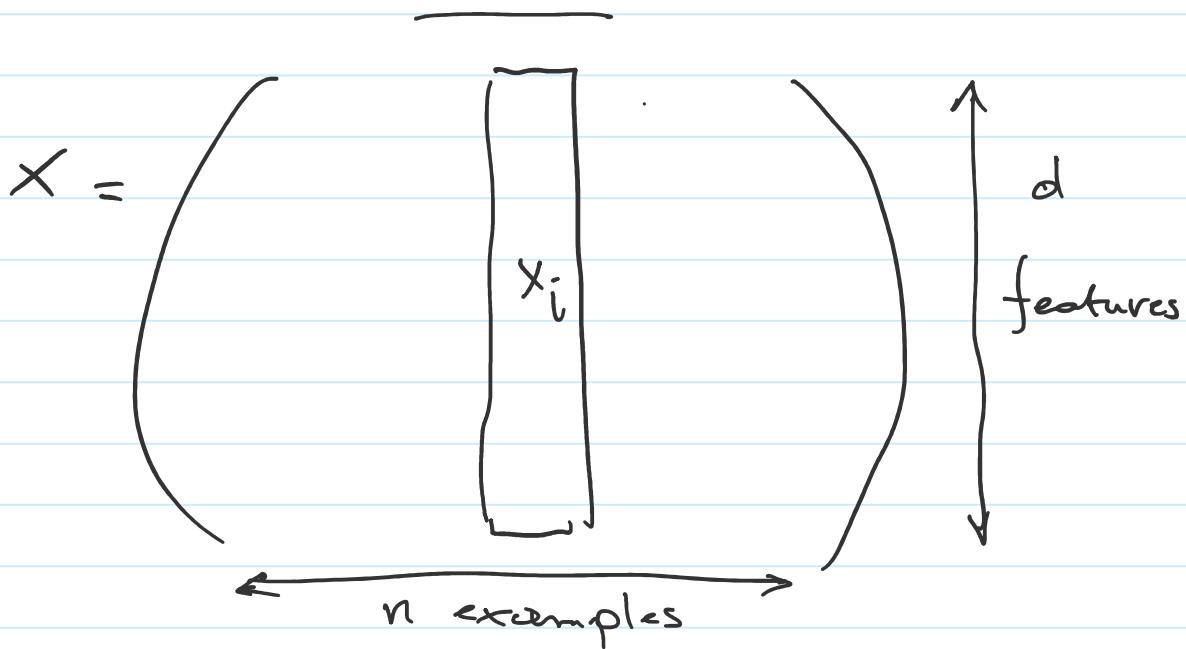
↑  
1st  
coordinate  
of  $x$

$$x^T \beta = (x^1, x^2, x^3, x^4, x^5) \begin{pmatrix} 0 \\ 0 \\ 3 \\ 0 \\ -2 \end{pmatrix}$$

$$= 0x^1 + 0x^2 + 3x^3 + 0x^4 + (-2)x^5$$

$$= 3x^3 - 2x^5$$

$$= 3x^3 - 2x^5$$



example  
 observation  
 datapoint  
 input

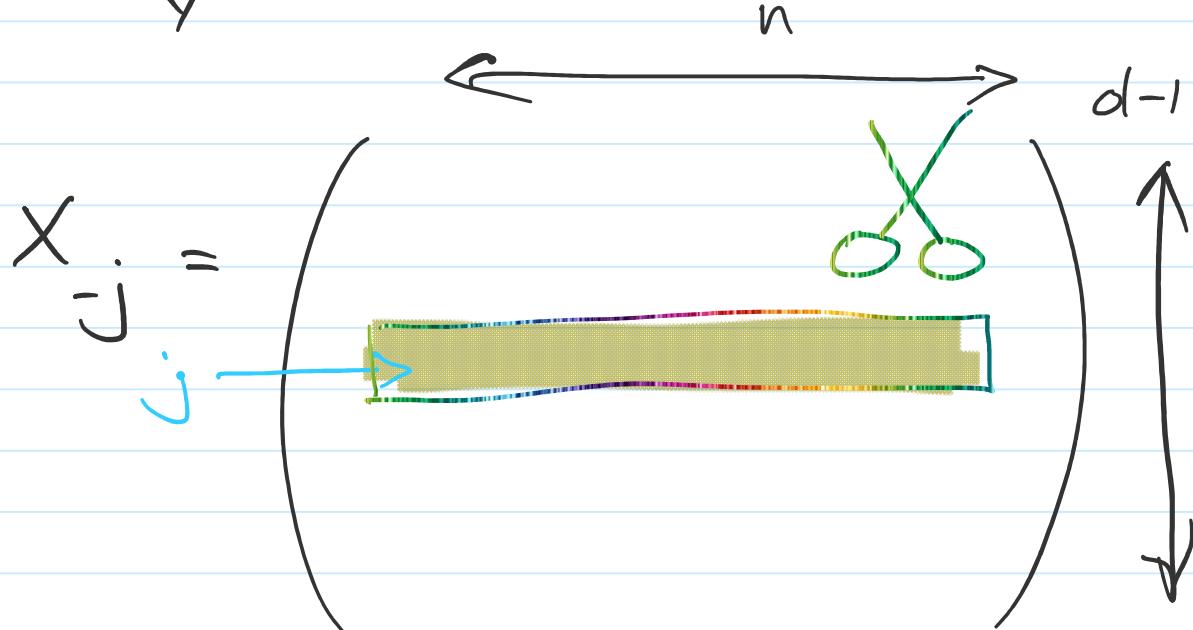
feature  
 Covariate

$x$

$x_j$

label  
 response  
 output

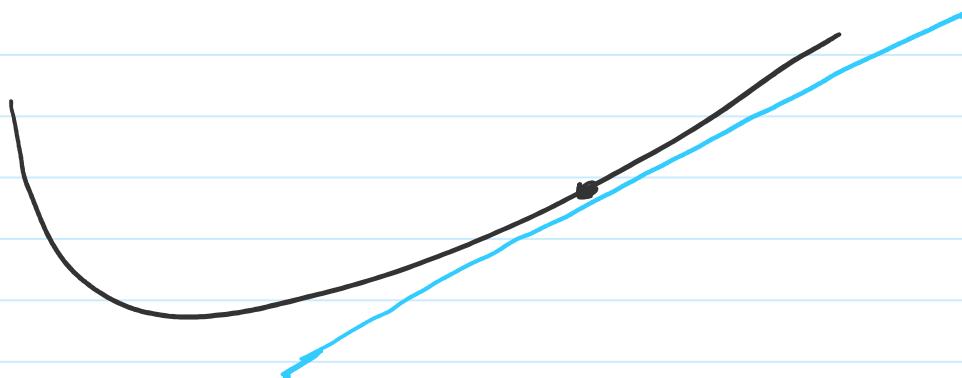
$y$



$$X_{-j}^T \quad n \times (d-1)$$

$$F(\beta) = \frac{1}{n} \| y - x^\top \beta \|^2_2 + \lambda \|\beta\|_1$$

$\beta$  convex differentiable



$$\partial F(\beta) = \underbrace{\{\nabla F(\beta)\}}_{\text{F is differentiable}}$$

Set

$$= -\frac{2}{n} x (y - x^\top \beta)$$

$\begin{bmatrix} d \times n & n \times 1 \end{bmatrix}$



$d \times 1$

$$\partial g(\beta) = \underbrace{\lambda \partial \{ ||\beta||_1 \}}_S$$

$v \in S$  if and only if

for all  $j = 1, \dots, d$

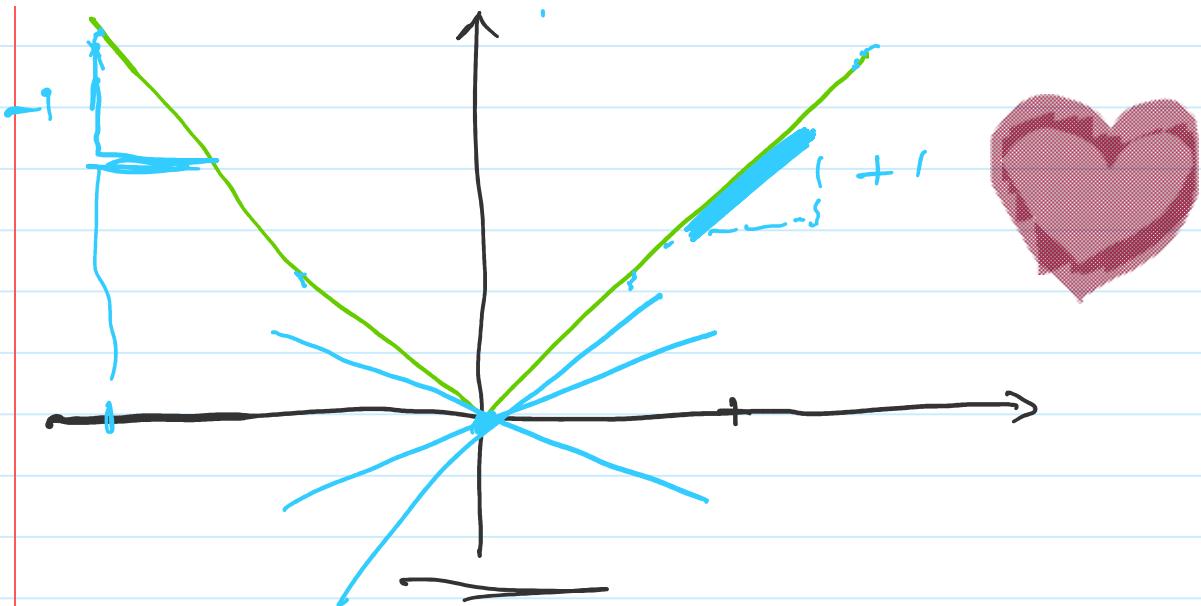
$$v_j = \begin{cases} +1 & \text{if } \beta_j > 0 \\ -1 & \text{if } \beta_j < 0 \\ [-1, +1] & \text{if } \beta_j = 0 \end{cases}$$

$$h(\gamma) = |\gamma| \quad S' = \partial h$$

$v \in S'$  if and only if

$$v = \begin{cases} +1 & \text{if } \gamma \geq 0 \\ -1 & \text{if } \gamma < 0 \\ [-1, +1] & \text{if } \gamma = 0 \end{cases}$$





in other words

$$u \in S \Leftrightarrow \text{For all } j = 1, \dots, d$$

$$\sigma_j = \begin{cases} \text{sign}(\beta_j) & \text{if } \beta_j \neq 0 \\ [-1, +1] & \text{if } \beta_j = 0 \end{cases}$$

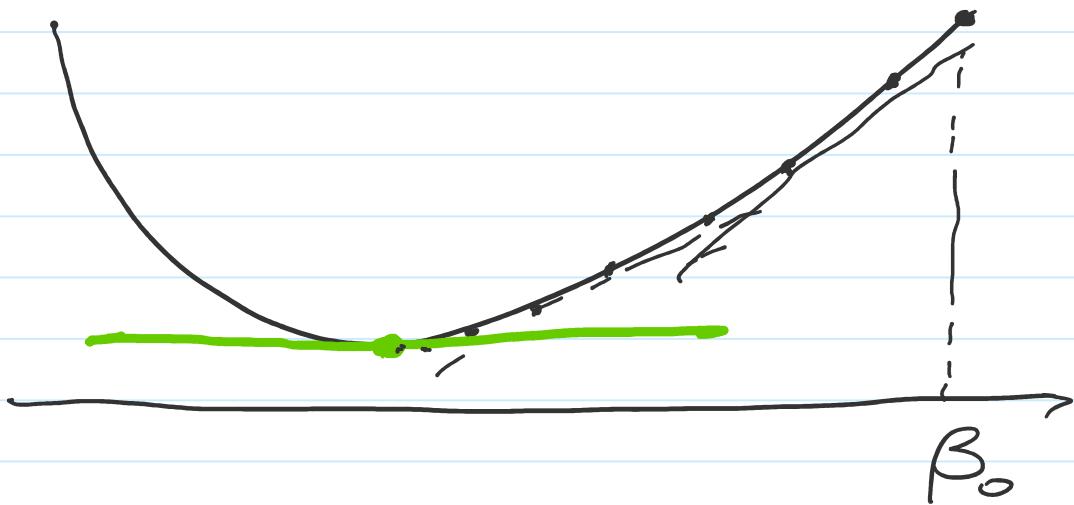
$\beta^*$  is such that  $F(\beta^*) = \min_{\beta} F(\beta)$

Optimality condition

$$o \in \partial F(\beta) \\ (\partial F(\beta) \ni o)$$

Minimize  $F(\beta)$   
 $\beta \in \mathbb{R}^d$

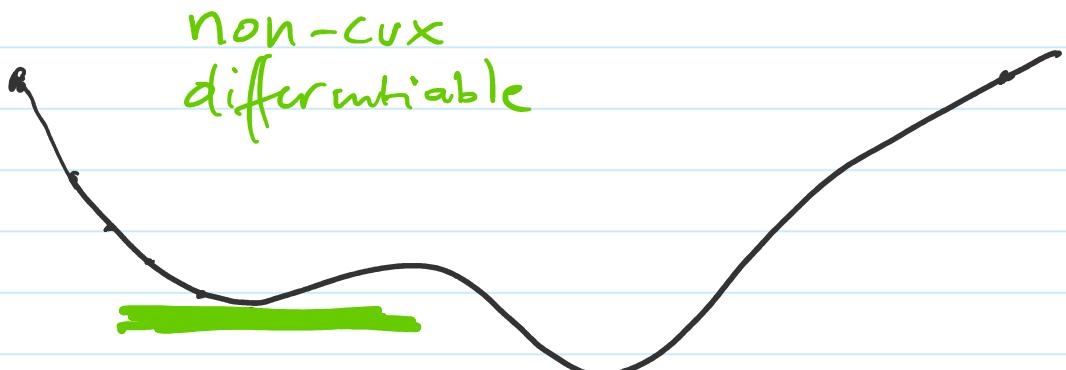
$F$  is convex  
differentiable



$$\nabla F(\beta) = 0$$

$$\|\nabla(\beta)\|_2 \leq \varepsilon.$$

non-convex  
differentiable



$\beta_0$

convex  
non-differentiable

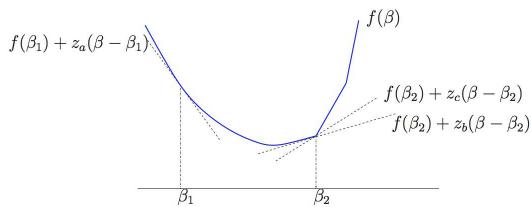


Figure 5.3 A convex function  $f : \mathbb{R} \rightarrow \mathbb{R}$ , along with some examples of subgradients at  $\beta_1$  and  $\beta_2$ .

LASSO

$$\underset{\beta \in \mathbb{R}^d}{\text{Min}} \quad F(\beta) = \frac{1}{n} \| Y - X^\top \beta \|^2 + \lambda \|\beta\|_1$$

$$0 \in \partial F(\beta)$$

$$0 \in \left\{ \nabla F(\beta) + \lambda \partial \{ \|\beta\|_1 \} \right\}$$

||

$$0 \in \{v, v = \nabla f(\beta) + \lambda s\}$$

with  $s \in \partial \{||\beta||_1, 3\}$

there exists some  $s \in \partial \{||\beta||_1, 3\}$

such that

$$0 = -\frac{2}{n} X(Y - X^T \beta) + \lambda s$$

$$s \in \partial \{||\beta||_1, 3\}$$



$\Rightarrow$  for all  $j = 1, \dots, d$

$$s_j = \begin{cases} \text{sign}(\beta_j) & \text{if } \beta_j \neq 0 \\ [-1, +1] & \text{if } \beta_j = 0 \end{cases}$$

$$\begin{bmatrix} 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \end{bmatrix}^d$$

$$= -\frac{2}{n} X(Y - X^T \beta) + \lambda \begin{bmatrix} s_1 \\ \vdots \\ s_d \end{bmatrix}$$

$$\beta_j = 0$$

$$s_j \in [-1, +1]$$

zoom on the  $j$ -th row

$$\underbrace{0}_{\lambda} = -\frac{2}{n} x_j (y - x^T \beta) + \lambda s_j$$

$$\frac{2}{n} x_j (y - x^T \beta) = \lambda s_j$$

$$\begin{bmatrix} -1, +1 \end{bmatrix}$$

$$\left| \frac{2}{n} x_j (y - x^T \beta_j) \right| \leq \lambda$$

$$\underline{\beta_j \neq 0}$$

$$\left| \frac{2}{n} x_j (y - x^T \beta_j) \right| = \lambda$$

# Comparison with Ridge Regression

$$\underset{\beta \in \mathbb{R}^d}{\text{Min}} \quad f(\beta) = \frac{1}{n} \|y - x^\top \beta\|_2^2 + \lambda \|\beta\|_2^2$$

Optimality condition  $\left( \begin{array}{l} \text{C} \times \\ \text{differentiable} \end{array} \right)$

$$\nabla F(\beta) = 0$$

$$-\frac{2}{n} x \left( y - x^\top \beta \right) + 2\lambda \beta = 0$$

$$\beta_j = 0$$

$$-\frac{2}{n} x_j \left( y - (x^\top \beta) \right) + 2\lambda \beta_j = 0$$

$$\sum_j x_j \left( y - (x^\top \beta) \right) = 0$$

$j$ -th  
feature

residual of the  
 $j$ -th feature

$n$  examples

$r \beta$

$\underbrace{1 \quad 1 \quad \dots \quad 1}_{n \text{ examples}}$

$$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_d \end{pmatrix}$$

$$x = \begin{pmatrix} & & \\ \uparrow & & \\ \text{feature} & & \\ & & \downarrow \end{pmatrix}$$

$$\underbrace{x^T \beta}_{n \times d \quad d \times 1}$$

$n \times 1$

$n \times 1$

$$\cdot -\frac{2}{n} \times (y - x^T \beta) + 2\lambda \beta = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$$

$$\cdot -\frac{2}{n} \sum_j (y - x^T \beta) + 2\lambda \beta_j = 0$$

RR

$$\beta_j = 0 \Rightarrow -\frac{2}{n} \sum_{\text{residuals}} (y - x^T \beta) = 0$$

$j$ - features

LASSO

$$\beta_j = 0 \Rightarrow \left| \frac{2}{n} x_j (y - x^T \beta) \right| \leq \lambda$$

$$\min_{\beta \in \mathbb{R}^d} \frac{1}{n} \|y - x^T \beta\|_2^2 + \lambda \|\beta\|_1$$

$$\frac{1}{n} \sum_{i=1}^n (y_i - x_i \beta_j)^2 + \lambda |\beta_j|$$

—

## Coordinate Descent

Init for all  $j = 1, \dots, d$

Repeat for  $t = 0, 1, 2, \dots$

- Pick a coordinate with index  $j$   
 $j \in \{1, \dots, d\}$

- find  $\beta_j^{\text{NEW}}$  by

$$\min_{\beta_j} F(\beta_1, \dots, \beta_{j-1}, \beta_j^{\text{NEW}}, \beta_{j+1}, \dots, \beta_d)$$

$\beta_1, \dots, \beta_{j-1}, \beta_{j+1}, \dots, \beta_d$  **fixed**

- Perform update

$$\beta_j^{(t+1)} = \beta_j^{\text{NEW}}$$

$$\beta_\ell^{(t+1)} = \beta_\ell^{(t)}$$

for  $\ell \neq j$

$$f(\beta) \quad g(\beta)$$
$$\frac{1}{2} \|y - X^\top \beta\|^2$$

(S)  $\min$

$$\left(\begin{array}{l} \text{Min}_{\beta_j} \\ \beta_j \end{array}\right) \quad \frac{1}{n} \|y - x^T \beta\|_2^2 + \lambda \|\beta\|_1$$

others fixed

$\hat{R}_j$

$$\nabla f(\beta) = -\frac{2}{n} X \left( y - x^T \beta \right)$$

$R$   
residuals

$$\partial_{\beta_j} f(\beta_j)$$

$$f(\beta_1, \beta_2) \quad \nabla_{\beta_1}^F \quad \nabla_{\beta_2}^F$$

$$\frac{\partial F}{\beta_1} \quad \frac{\partial F}{\beta_2}$$

Optimality conditions for  $(\beta_j)$

$$\partial_{\beta_j} f(\beta_j) = 0$$

$$\partial_{\beta_j} + (\beta_j) = 0$$

$$\cdot \partial_{\beta_j} g(\beta) = \partial_{\beta_j} \left\{ \|\beta\|_1 + \|\beta_j\|_1 \right\}$$

$$= \partial_{\beta_j} \left\{ \|\beta_j\|_1 \right\}$$

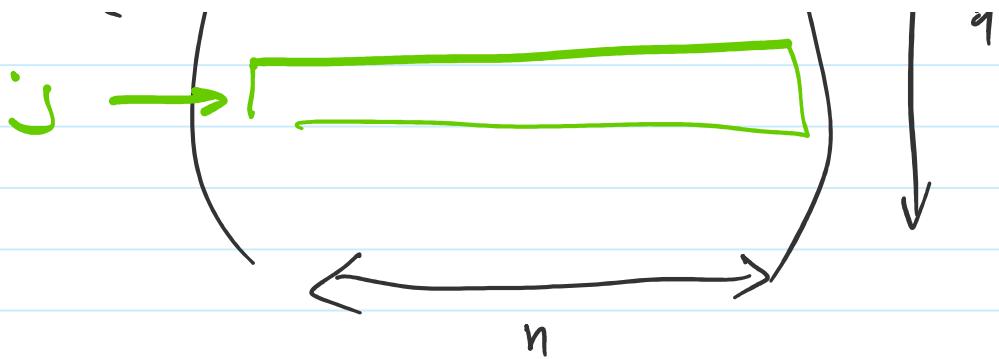
$$\cdot \nabla_{\beta_j} \beta(\beta)$$

$$= -\frac{2}{n} x_j (y - x^T \beta)$$

$$= -\frac{2}{n} x_j \left( y - \sum_j \beta_j x_j + x^T \beta \right)$$

$$x_j = \underbrace{\quad}_{n} \quad \quad 1 \times n$$

$$x = \underbrace{\quad}_{d}$$



$$x_{-j} = \begin{pmatrix} & \\ & \\ & \\ & \end{pmatrix} \quad d-1$$

↑      ↓  
n      n

$$\beta = \begin{pmatrix} & \\ & \\ & \\ & \end{pmatrix} \quad d$$

$$\beta_{-j} = \begin{pmatrix} & \\ & \\ & \\ & \end{pmatrix} \quad d-1$$

$$\begin{aligned}
 & -\frac{2}{n} \left| x_j \left( y - x_{-j}^T \beta_{-j} \right) \right| \rightarrow \in \mathbb{R} \\
 & -\frac{2}{n} \times (-!) \times \left( x_j x_j^T \beta_j \right) \frac{\|x_j\|_2^2}{\|x_j\|_2^2}
 \end{aligned}$$

$$= -\frac{2}{n} \left( x_j R^j - \beta_j z_j \right)$$

$$= -\frac{2}{n} \left( x_j R^j - \beta_j z_j \right)$$

$$z_j = \sum_{i=1}^n x_{ij}^2$$

$$R^j = Y - X_{-j}^\top \beta_{-j}$$

$$\text{OE} = -\frac{2}{n} \left( x_j R^j - \beta_j z_j \right) + \lambda \boxed{\partial f / \beta_j}$$

there exists  $s_j$  such that

$$-\frac{2}{n} \left( x_j R^j - \beta_j z_j \right) + \lambda s_j = 0$$

$$s_j = \begin{cases} \text{sign}(\beta_j) & \text{if } \beta_j \neq 0 \\ \in [-1, +1] & \text{if } \beta_j = 0 \end{cases}$$

$$\beta_j \leq 0 \Rightarrow \text{sign}(\beta_j) = -1$$

$$= 1 \dots n-j \dots -1 \quad 1 \dots$$

$$-\frac{z}{n} (x_j R^{-j} - \beta_j z_j) - \lambda = 0$$

$$\beta_j = \frac{\lambda + \frac{z}{n} x R^{-j}}{\frac{z}{n} z_j}$$

$$\underline{\beta_j = 0}$$

$$-\frac{z}{n} (x_j R^{-j} - \beta_j z_j) + \lambda s_j = 0$$

$$\in [-1, +1]$$

$$\left| \frac{z}{n} (x_j R^{-j} - \underbrace{\beta_j z_j}_0) \right| \leq \lambda$$

$$\left| \frac{z}{n} x_j R^{-j} \right| \leq \lambda$$

Summary

if  $\frac{2}{n} X R^{-j} \leq -\lambda$

Then  $\beta_j = \frac{\lambda + \frac{2}{n} X_j R^{-j}}{\frac{2}{n} z_j}$

if  $\frac{2}{n} X R^{-j} \geq +\lambda$

Then  $\beta_j = \frac{-\lambda + \frac{2}{n} X_j R^{-j}}{\frac{2}{n} z_j}$

if  $|\frac{2}{n} X_j R^{-j}| \leq \lambda$

Then  $\beta_j = 0$

$$\underset{\substack{\text{Min}_{\beta} \\ \beta \in \mathbb{R}^d}}{\frac{1}{n} \sum_{i=1}^n \ell(y_i, x_i^T \beta)} + \lambda \Omega(\beta)$$

A)  $\ell(y, x^T \beta) = (y - x^T \beta)^2$   
 B)  $\ell(y, x^T \beta) = \log(1 + e^{-y x^T \beta})$

$$B) \ell(y, x^\top \beta) = \log(1 + e^{-y})$$

$$1) \Omega(\beta) = \|\beta\|_2^2$$

$$2) \Omega(\beta) = \|\beta\|_1$$

A1

ridge regression

A2

LASSO

B1

$\ell_2^2$ -regularized LR

B2

$\ell_1$ -regularized LR

FGM
CD
FGM
Acc Proximal Gradient Method

- $y = \sin(x)$  NO

- $y = \log(-x)$

- $y = \frac{1}{1+|x|}$  NO

$$\cdot \quad y = \frac{1}{1+|x|}$$

$$\cdot \quad \|\beta\|_p = \left( \sum_{j=1}^d |\beta_j|^p \right)^{\frac{1}{p}}$$

$$\cdot \quad \ell(y, x^\top \beta) = (y - x^\top \beta)^4,$$

$$\cdot \quad \Omega(\beta) = (1-\alpha) \|\beta\|_2^2 + \alpha \|\beta\|_1$$

$$\cdot \quad \Omega(\beta) = \|\beta\|_1^2$$

$$\begin{aligned} \beta \in \mathbb{R}^n & \quad \Omega(\beta) = |\beta|^2 \\ & = \beta^2 \end{aligned}$$

$$\cdot \quad \ell(y, x^\top \beta) = \exp(-y x^\top \beta)$$