Lecture 5. Coordinate Descent

Tuesday, April 24, 2018 6:23 AM

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

- Least-squares

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

- L_1-norm penalty

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

β^* is <u>sparse</u>

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

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

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

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

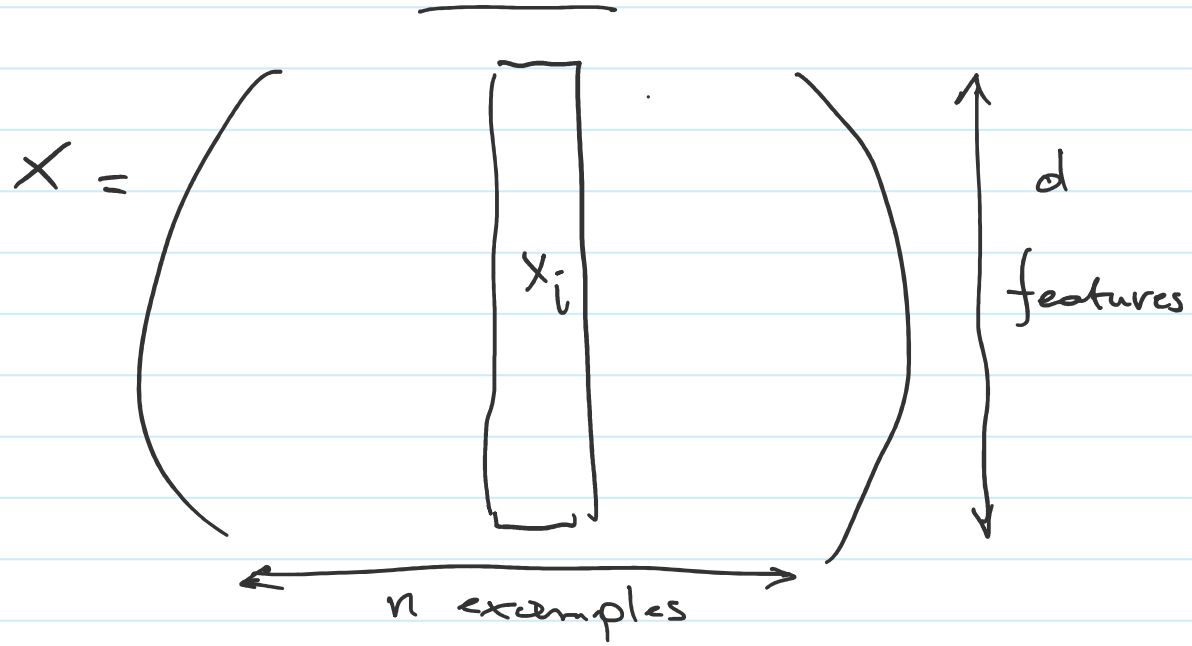
1st coordinate of x

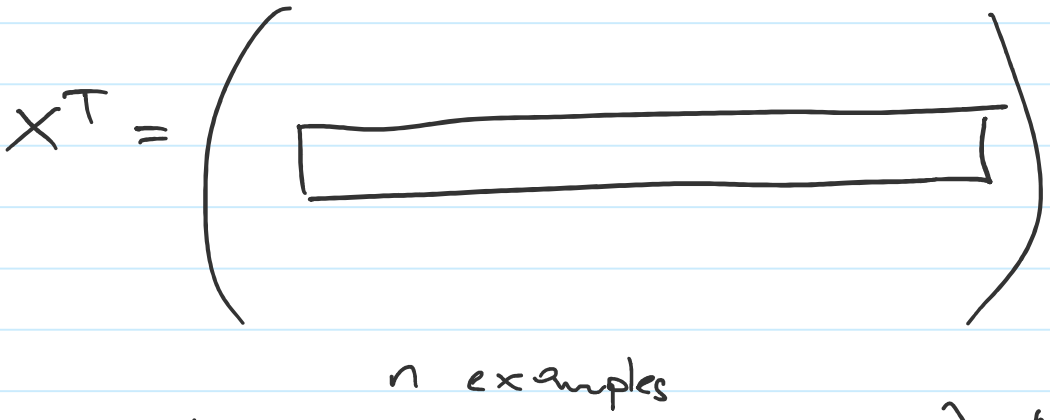
$$x^T\beta = (x^1, x^2, x^3, x^4, x^5) \begin{pmatrix} 0 \\ 0 \\ 3 \\ 0 \\ -2 \end{pmatrix}$$

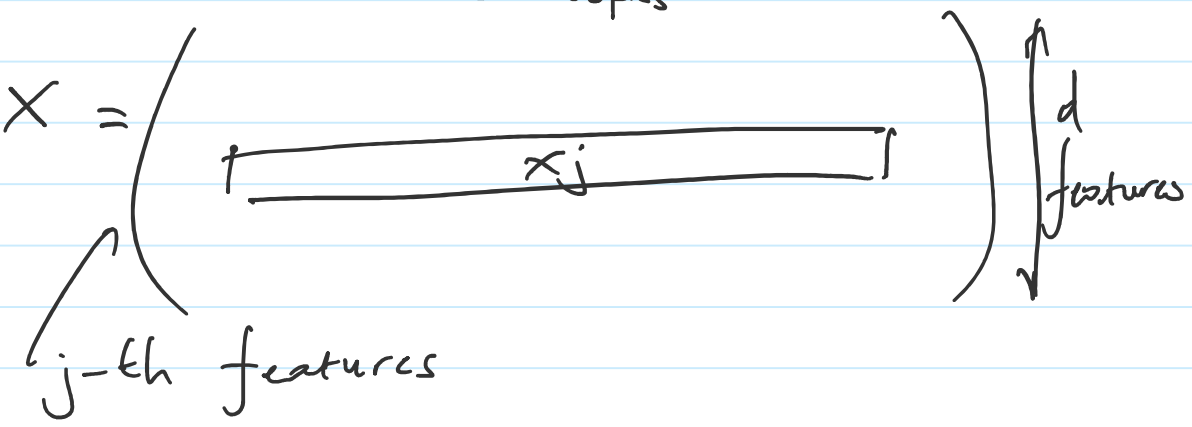
$$= 0x^1 + 0x^2 + 3x^3 + 0x^4 + (-2)x^5$$

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

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


X =
x_i
d
features
n examples


X^T =
n examples


X =
x_j
d
features
j-th features

example
observation
datapoint
input

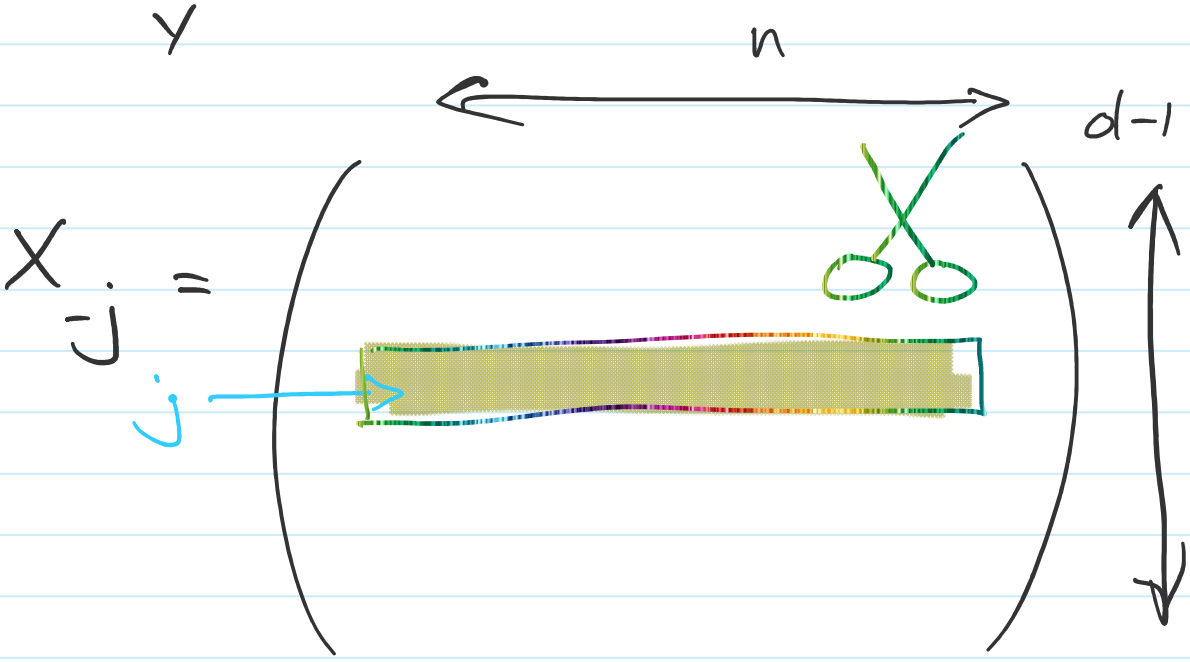
x

feature
covariate

x_j

label
response
output

y

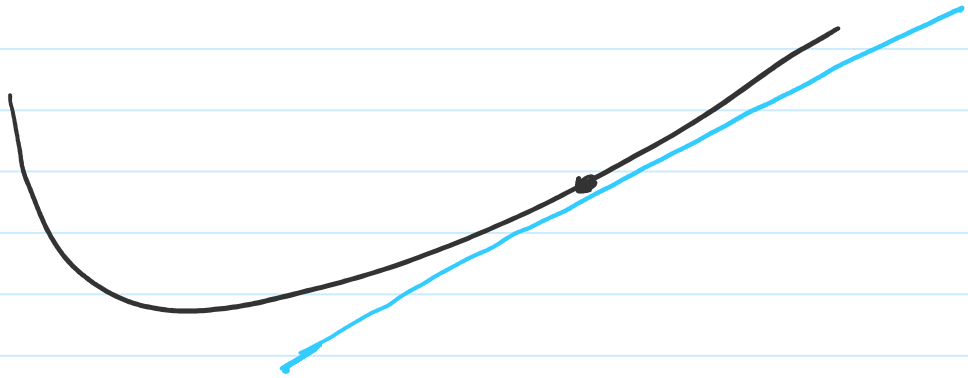


$X_{\cdot j}^T$ $\quad n \times (d-1)$

$$F(\beta) = \underbrace{\frac{1}{n}\|Y - X^T\beta\|_2^2}_{f(\beta)} + \underbrace{\lambda\|\beta\|_1}_{g(\beta)}$$

---

f cvx differentiable



$$\partial f(\beta) = \{\nabla f(\beta)\} \quad [f \text{ is differentiable}]$$

set

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

$d \times n \quad n \times 1$

$d \times 1$

- $\partial g(\beta) = \underbrace{\lambda \partial \{ \|\beta\|_1 \}}_{S}$

$v \in S$ if and only if

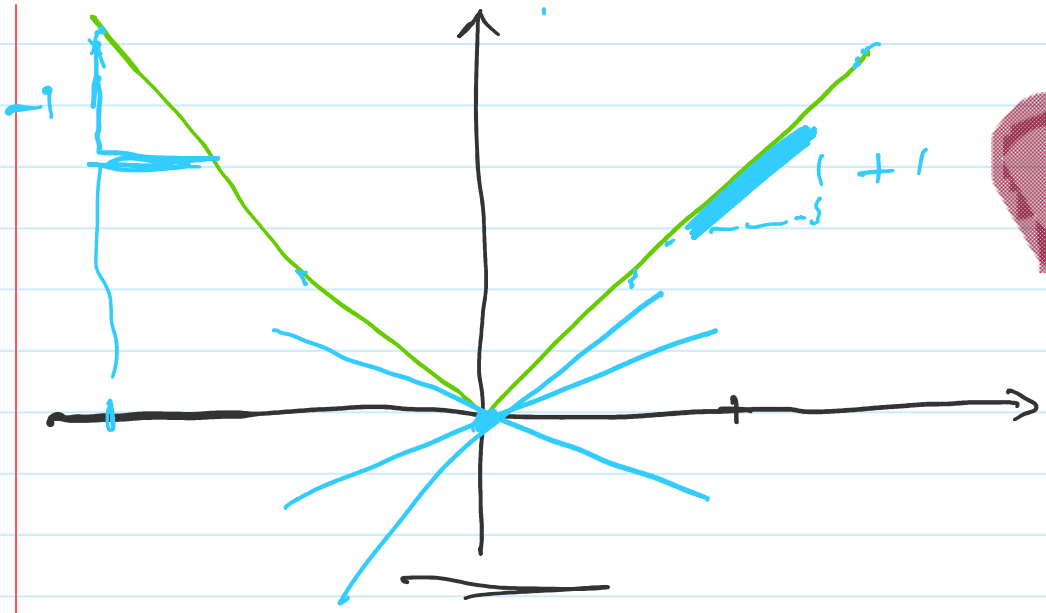
for all $j = 1, \ldots, d$

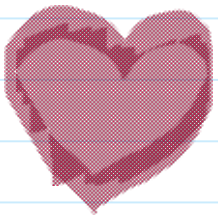
$$v_j = \begin{cases} +1 & \text{if } \beta_j > 0 \\ -1 & \text{if } \beta_j < 0 \\ [-1, +1] & \text{if } \beta_j = 0 \end{cases}$$

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

$v \in S'$ if and only if

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





in other words

$$u \in S \iff \text{for all } j = 1, \dots, d$$

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

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

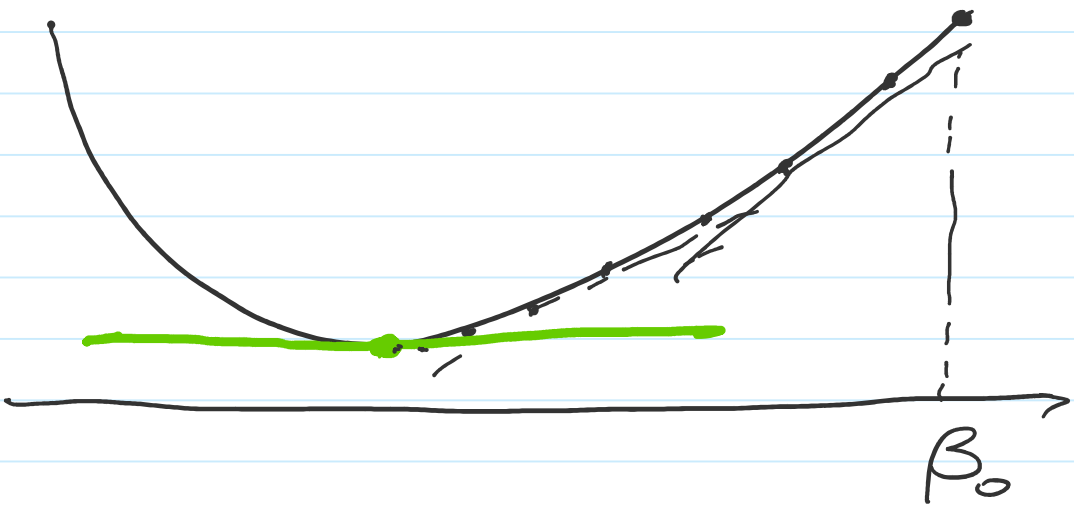
Optimality condition

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

$$\left(\partial f(\beta) \ni 0\right)$$

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

f is convex differentiable



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

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

non-cvx
differentiable

cux
non-differentiable

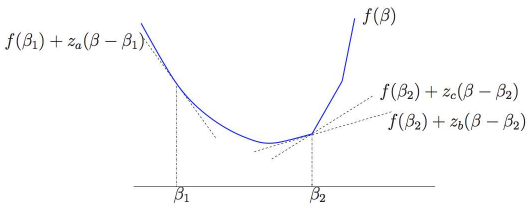


Figure 5.3 *A convex function $f : \mathbb{R} \to \mathbb{R}$, along with some examples of subgradients at β_1 and β_2.*

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

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

$0 \in \{ \nabla g(\beta) + \lambda \partial \{ \|\beta\|_1 \} \}$

$\|$

$$0 \in \left\{ v, \quad v = \nabla f(\beta) + \lambda s \right.$$
$$\left. \text{with} \quad s \in \partial \{ \|\beta\|_1 \} \right\}$$

there exists some $s \in \partial \{\|\beta\|_1\}$ such that

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

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

$\Leftrightarrow$ for all $j = 1, \ldots, d$

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

$$\begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}_d = -\frac{2}{n} X (Y - X^T \beta) + \lambda \begin{bmatrix} s_1 \\ \vdots \\ s_d \end{bmatrix}$$

$\beta_j = 0$ $\qquad s_j \in [-1, +1]$

zoom on the j-th row

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

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

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

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

$\beta_j \neq 0$

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

Comparison with Ridge Regression

$$\min_{\beta \in \mathbb{R}^d} f(\beta) = \frac{1}{n} \| Y - X^T \beta \|_2^2 + \lambda \|\beta\|_2^2$$

Optimality condition (cvx differentiable)

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

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

$\beta_j = 0$

$$-\frac{2}{n} x_j \left(Y - (X^T \beta) \right) + \cancel{2\lambda \beta_j = 0}$$

$$x_j^T \underbrace{\left(Y - (X^T \beta) \right)}_{\text{residual of the } j\text{-th feature}} = 0$$

j-th feature

n examples

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

d features

$\underbrace{X^T\beta}$ $\qquad$ Y

$n \times d \quad d \times 1$

$n \times 1 \qquad\qquad n \times 1$

- $-\frac{2}{n} X (Y - X^T\beta) + 2\lambda\beta = \begin{pmatrix} 0 \\ | \\ 0 \end{pmatrix}$

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

RR

$\beta_j = 0 \Rightarrow -\frac{2}{\_\_} \underbrace{(Y - X^T\beta)}_{\text{residuals}} = 0$

j- features

LASSO

$$\hat{\beta}_j = 0 \quad \Rightarrow \quad \left| \frac{2}{n} x_j \left(Y - X^T \beta \right) \right| \leq \lambda$$

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

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

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 β_j^{NEW} by

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

- $\beta_1, \dots, \beta_{j-1}, \beta_{j+1}, \dots, \beta_d$ FIXED
- **Perform update**

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

$$\beta_\ell^{(t+1)} = \beta_\ell^{(t)} \quad \text{for } \ell \neq j$$

$f(\beta) + g(\beta)$

$$\min_\beta \frac{1}{2} \|Y - X^T\beta\|^2$$

(S) $\min_{\beta_j}$ others fixed

$$\frac{1}{n}\|Y - X^T\beta\|_2^2 + \lambda\|\beta\|_1$$

$$= \frac{1}{n}\|Y - \sum_{k \neq j} X_k\beta_k - X_j\beta_j\|_2^2 + \lambda\|\beta\|_1$$

$$= \frac{1}{n}\|R - X_j\beta_j\|_2^2 + \lambda\|\beta\|_1$$

R_j

$$\nabla f(\beta) = -\frac{2}{n} X (Y - X^T\beta)$$

R residuals

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

$f(\beta_1, \beta_2)$ $\quad \nabla f_{\beta_1} \quad \nabla f_{\beta_2}$

$\partial f_{\beta_1} \quad \partial f_{\beta_2}$

Optimality conditions for (P_j)

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

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

- $\partial_{\beta_j} g(\beta) = \partial_{\beta_j} \{$ [illegible] $+ |\beta_j| +$ [illegible] $\}$

$= \partial_{\beta_j} \{ |\beta_j| \}$

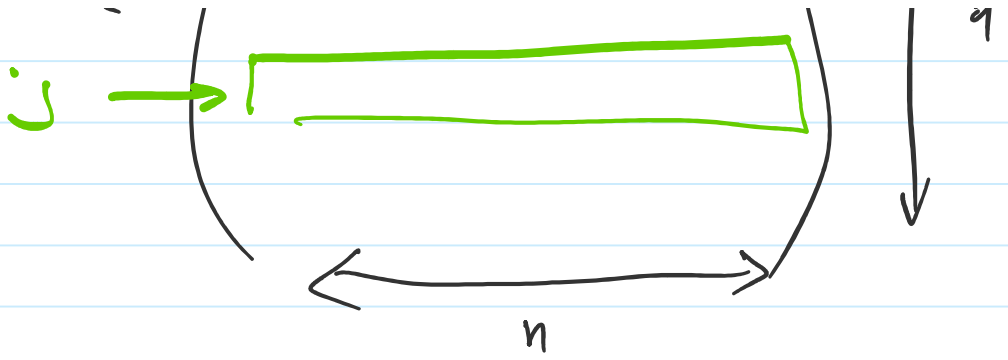
- $\nabla_{\beta_j} R(\beta)$ — $\frac{\partial R}{\partial \beta_j}$

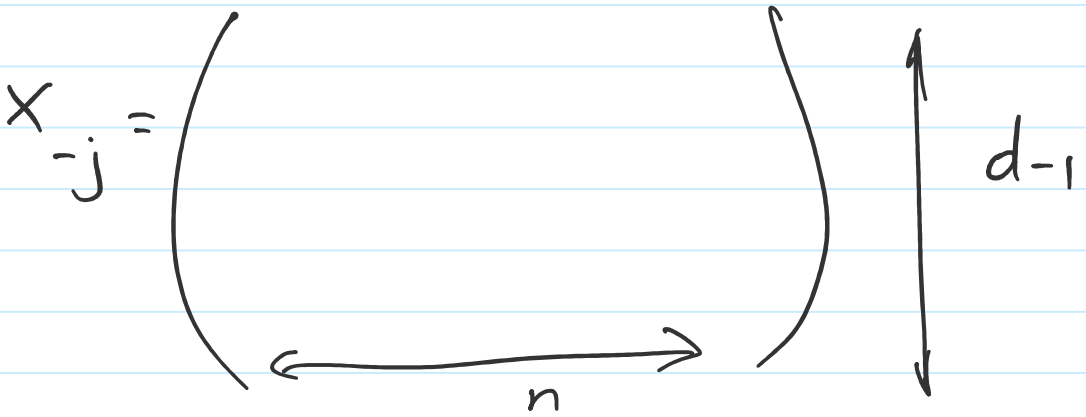
$= -\frac{2}{n} x_j (Y - X^T \beta)$

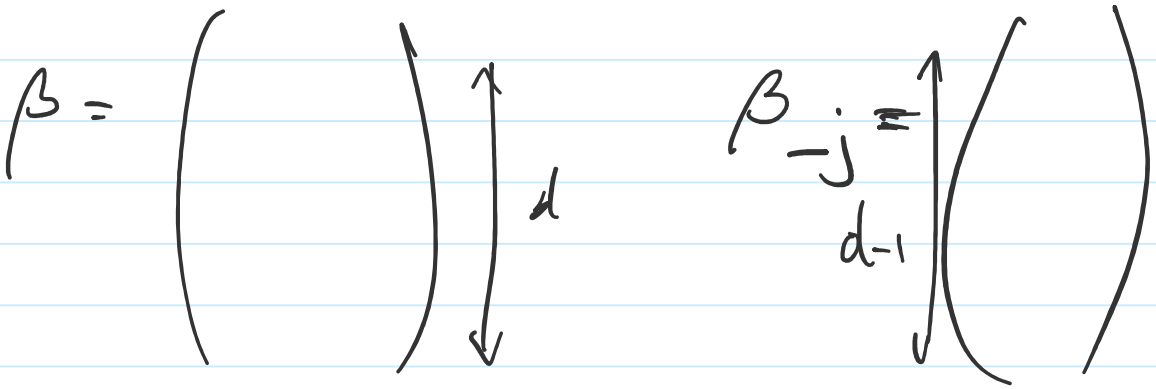
$= = \frac{2}{n} x_j (Y - X_{-j}^T \beta_{-j} + x_j^T \beta_j)$

$x_j =$ ($\longleftrightarrow$) $1 \times n$ (n)

$X =$ $\uparrow d$







$$-\frac{2}{n} x_j \left(Y - X_{-j}^T \beta_{-j} \right) \in \mathbb{R}$$

$$-\frac{2}{n} \times (-1) \times \left(x_j x_j^T \beta_j \right)$$

$$\|x_j\|_2^2$$

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

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

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

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

$$0 \in -\frac{2}{n}\left(X_j R^{-j} - \beta_j z_j\right) + \lambda \partial |\beta_j|$$

there exists s_j such that

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

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

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

$$= \left(X_j R^{-j} \quad \beta_j z_j\right) \quad \lambda = 0$$

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

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

$\beta_j = 0$

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

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

$$\left| \frac{2}{n}\left(x_j R^{-j} - \underbrace{\beta_j z_j}_{0}\right) \right| \leq \lambda$$

$$\left| \frac{2}{n} x_j R^{-j} \right| \leq \lambda$$

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 $\left| \frac{2}{n} x_j R^{-j} \right| \leq \lambda$

Then $\beta_j = 0$ ☺

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

A) $\ell(y, x^T\beta) = (y - x^T\beta)^2$

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

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

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

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

A1	ridge regression		FGM
A2	LASSO		CD
B1	L_2^2 - regularized	LR	FGM
B2	L_1 - regularized	LR	Acc Proximal Gradient Method

- $y = \sin(x)$ NO
- $y = \log(-x)$
- $y = \dfrac{1}{1+|x|}$ NO

- $y = \frac{}{1+|x|}$

- $\|\beta\|_p = \left(\sum_{j=1}^{d} |\beta_j|^p \right)^{1/p}$

- $\ell(y, x^T\beta) = (y - x^T\beta)^4$

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

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

$\beta \in \mathbb{R}^1 \quad \Omega(\beta) = |\beta|^2 = \beta^2$

- $\ell(y, x^T\beta) = \exp(-y x^T\beta)$