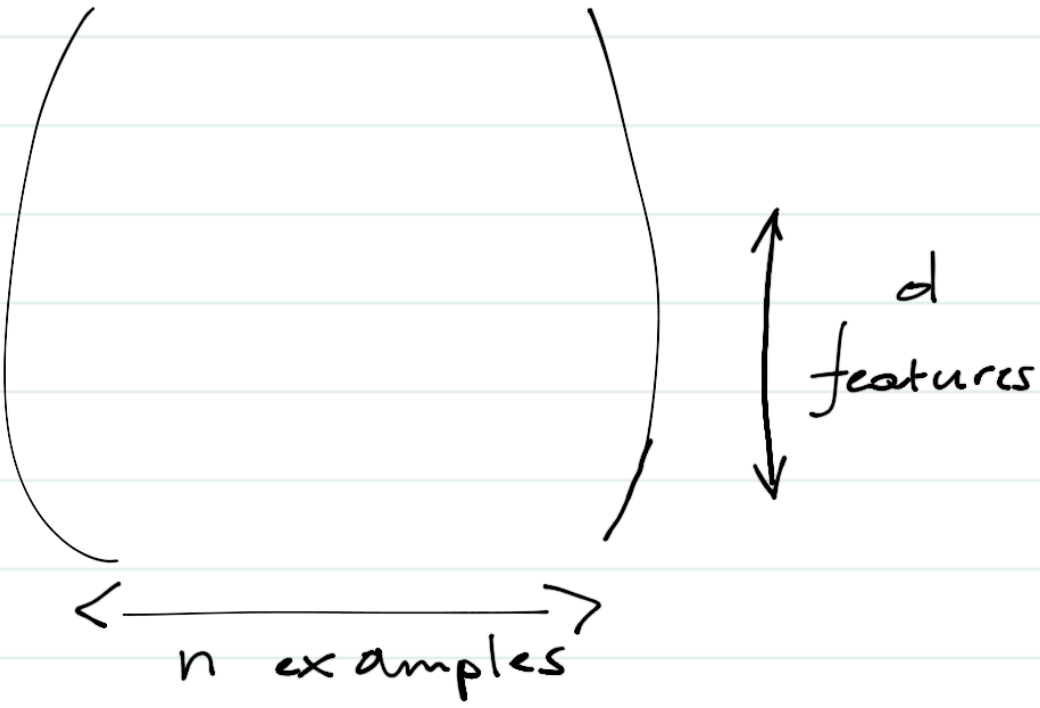
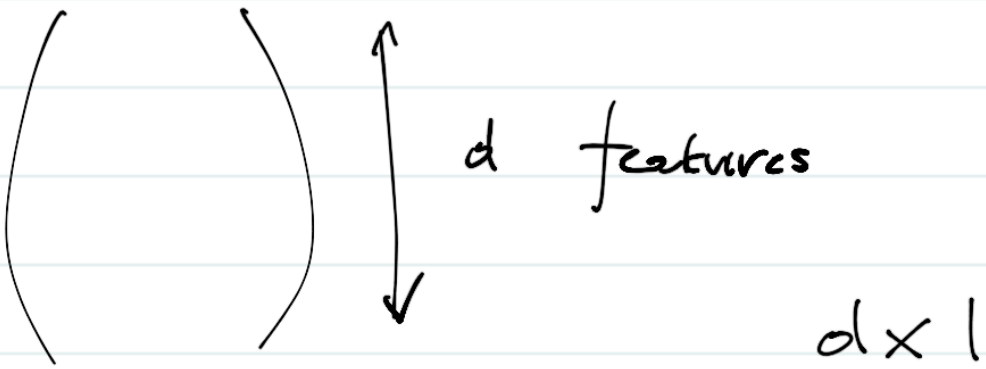
Linear regression Lecture 2

$X =$



$x_i =$



$d \times 1$

$$y_i = x_i^\top \beta + \varepsilon_i$$

y_i: response

ε_i: noise

$$y_i = \sum_{j=1}^{d} x_{i,j}\,\beta_j + \varepsilon_i$$

$y_i \in \mathbb{R}$

ε_i: Gaussian noise

$$x_i = \begin{pmatrix} \\ \\ \\ \end{pmatrix} \quad d \qquad d \times 1$$

$$x_i^T = \begin{pmatrix} & & & \end{pmatrix} \qquad 1 \times d$$

$$\beta = \begin{pmatrix} \\ \\ \\ \end{pmatrix} \quad d \times 1$$

$x_i^T \beta$

$1 \times d \quad d \times 1 \qquad\qquad 1 \times 1$

$$\min_{\beta \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(y_i - x_i^T \beta \right)^2 + \lambda \| \beta \|_2^2 \right\}$$

n examples

d features

$\underline{\lambda = 0}$ $\min_{\beta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - x_i^T \beta \right)^2$

least-squares

$$n \gg d$$

$\beta \in \mathbb{R}^d \qquad \beta = \begin{pmatrix} \\ \\ \\ \end{pmatrix} \} d$

(n)

$$\min_{\beta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - x_i^T \beta \right)^2 + \lambda \sum_{j=1}^{d} \beta_j^2$$

Constrained Least - squares

$$\begin{cases} \min_{\beta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - x_i^T \beta\right)^2 \\ \text{subject to} \quad 0 \le \beta_1 \le \beta_{max} \\ \vdots \\ 0 \le \beta_d \le \beta_{max} \end{cases}$$

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

(Mathematical) Optimization

- $\beta \in \mathbb{R}^d$

$\beta \in \mathcal{B}$

p^d

(grid-search, brute-force)

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

f differentiable

$\rightarrow$ Gradient Descent

Algo : $\beta = 0$

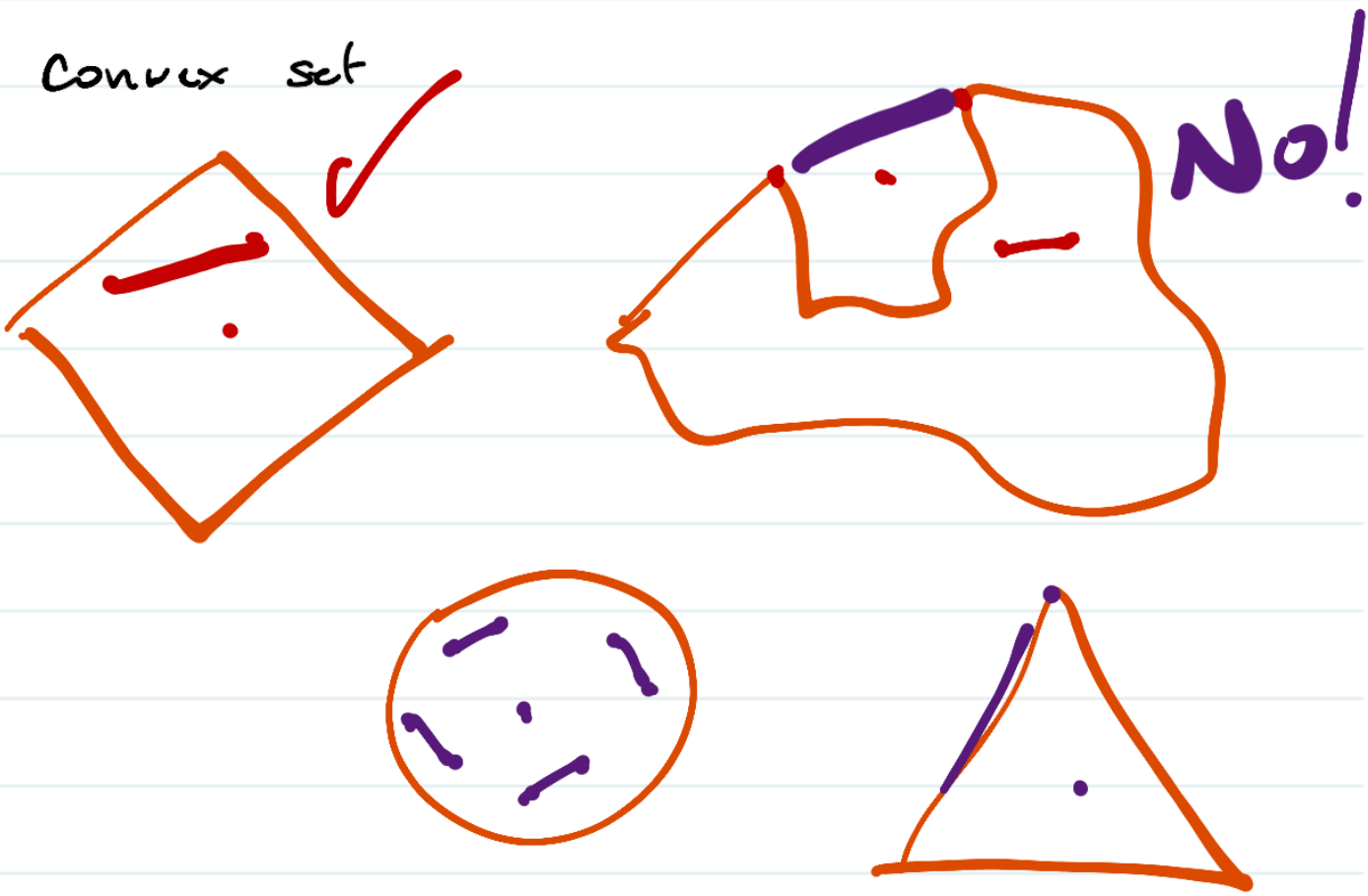
- Iterate for $t = 1, \dots, T$

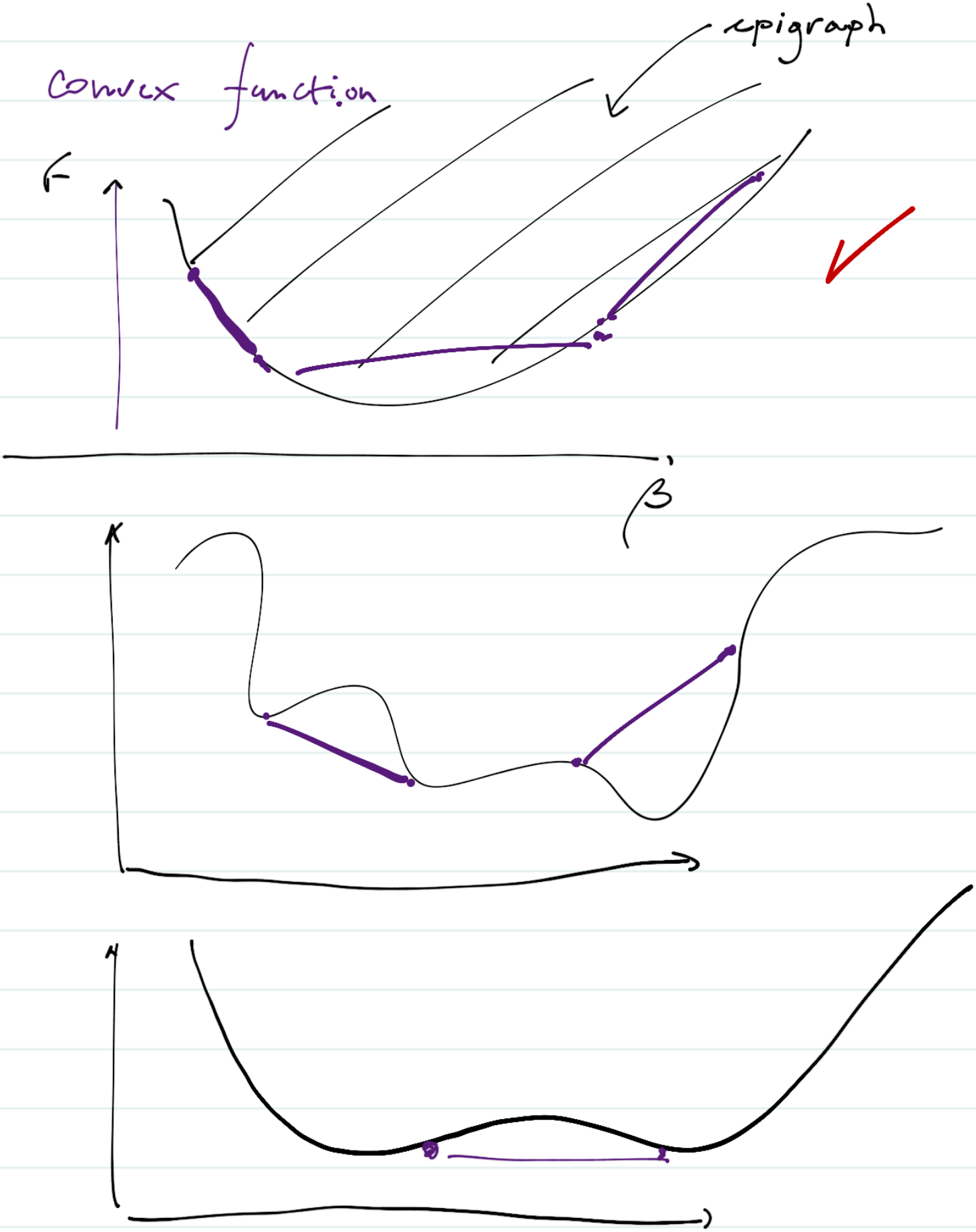
$$\beta^{(t+1)} = \beta^{(t)} - \eta_t \nabla f(\beta^{(t)})$$

Return $\beta^{(T)}$

- T ? maximum number of iterations
- $\beta^{(T)}$? What can you say about it ?
- η_t ? how to set the step-size

F convex

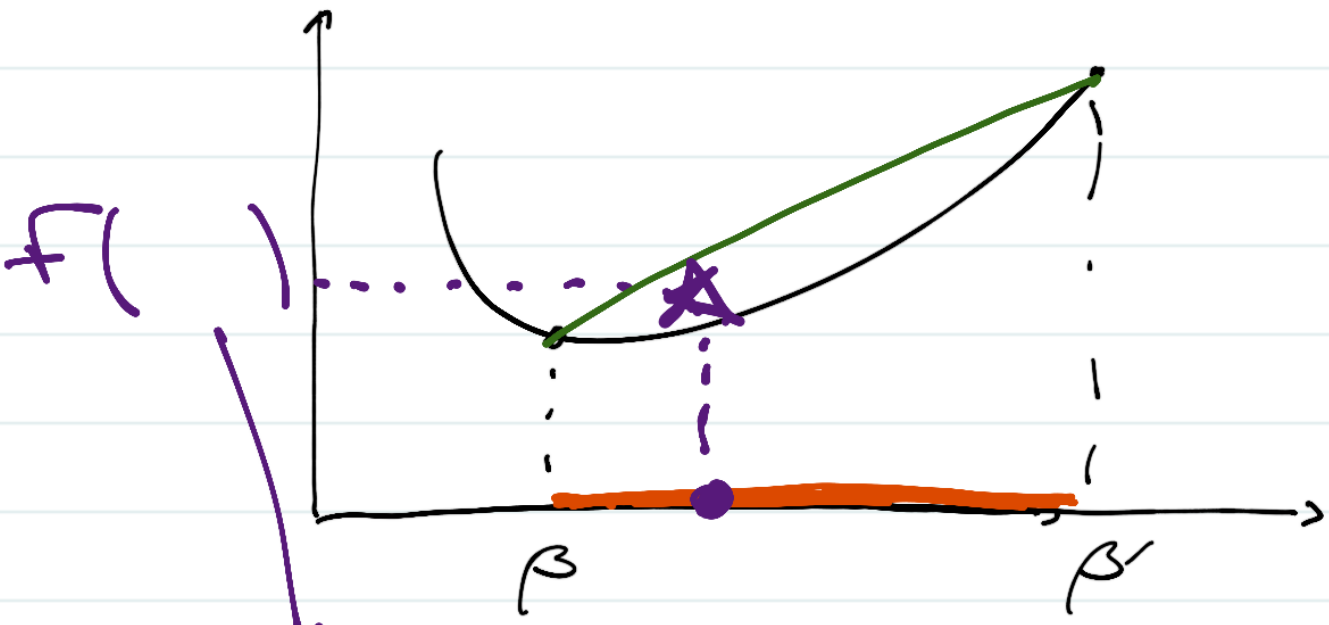



convex function
epigraph
F
β
K
H

$$f : \mathbb{R}^d \longrightarrow \mathbb{R}$$
$$\beta \longmapsto f(\beta)$$

For any $\beta, \beta' \in \mathbb{R}^d$

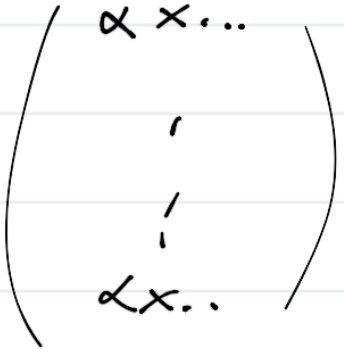
For any $\alpha \in [0, 1]$



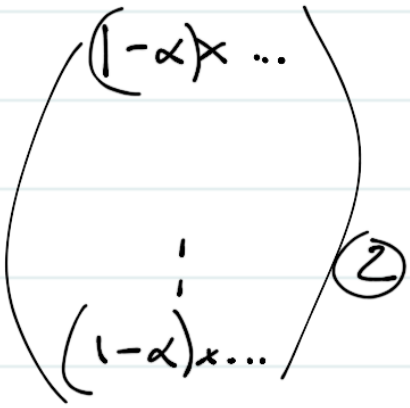
- $f\left(\alpha\beta + (1-\alpha)\beta'\right)$

$\leq \alpha f(\beta) + (1-\alpha) f(\beta')$

$\beta =$  α

$\alpha\beta =$  ①

$\beta' =$  α

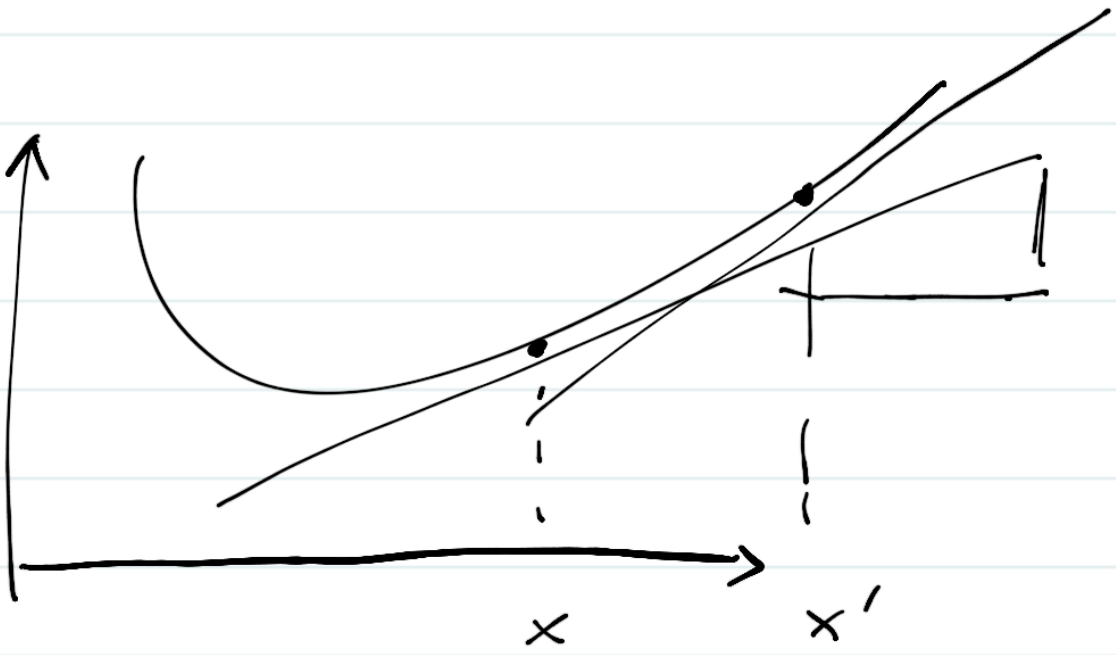
$(1-\alpha)\beta' =$  ②

$\alpha\beta + (1-\alpha)\beta' =$ ① + ②

Smooth functions

A function is L-smooth with constant L if F is continuously differentiable and for all x, x'

$$\| \nabla F(x) - \nabla F(x') \| \leq L \| x - x' \|$$



Convergence of gradient descent

Assume that F is :

- convex
- L - smooth

Then if $\eta_t = \frac{1}{L}$ (eta), we have

$$F(\beta^{(t)}) - F^* \leq \frac{2L \|\beta^{(0)} - \beta^*\|^2}{t}$$

- $F^* = \min_{\beta} F(\beta)$
- $\beta^* = \operatorname{Arg\,Min}_{\beta} F(\beta)$
- $F(\beta^{(t)}) - F^* = O\left(\frac{1}{t}\right)$

$$F(\beta^{(T)}) - F^* = \varepsilon$$

$$\frac{2L\|\beta^{(0)} - \beta^*\|}{t} \leq \varepsilon$$

$$t \geq \frac{2L\|\beta^{(0)} - \beta^*\|}{\varepsilon}$$

$$t \geq \frac{2LD}{\varepsilon}$$

$$T = \left\lceil \frac{2LD}{\varepsilon} \right\rceil + 1$$

- $\min_{\beta \in \mathbb{R}^d} F(\beta) := \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \|\beta\|_2^2$

$$\|\beta\|_2^2 = \sum_{j=1}^{d} \beta_j^2$$

- **Algo.** Gradient Descent

Input: L

Init: $\beta^{(0)}$

Iterate: for $t = 1, \dots, T$

$$\beta^{(t+1)} = \beta^{(t)} - \frac{1}{L} \nabla_\beta F(\beta^{(t)})$$

Return $\beta^{(T+1)}$

Simplest case

$n = 1 \qquad d = 1 \qquad \beta \in \mathbb{R}^1$

$$F(\beta) = \frac{1}{1} \sum_{i=1}^{1} (y_i - x_i \beta)^2 + \lambda \sum_{j=1}^{1} \beta_j^2$$

$$F(\beta) = (y - x\beta)^2 + \lambda \beta^2$$

- F convex ✓ sum of two quadratics $f'' \geq 0$
- F smooth ✓
- $F'(\beta) =$.

Chain rule $\qquad f(u) = g(h(u))$

$$f'(u) = h'(u) \cdot g'(h(u))$$

$$f(u) = au^2 \qquad f'(u) = 2au$$

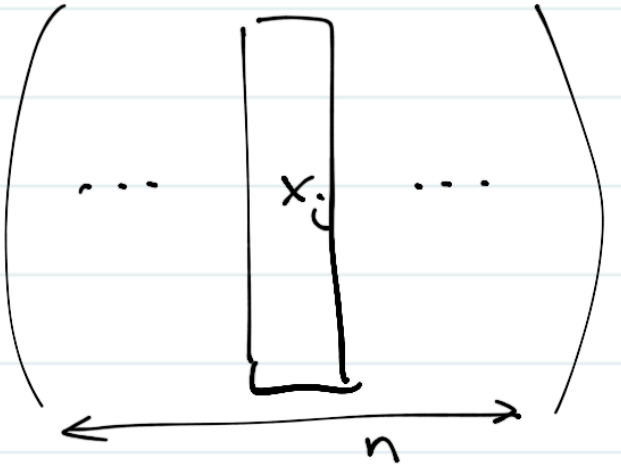
$$F'(\beta) = 2 \cdot (-x) \cdot (y - x\beta) + 2\lambda\beta$$

$n \geq 1$ $\qquad$ $d \geq 1$

$$F(\beta) = \frac{1}{n} \sum \cdots \qquad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_d \end{pmatrix}$$

$\bar{X}$ $\qquad$ β $\qquad$ $\bar{Y}$

$$\bar{X} = \begin{pmatrix} \cdots & x_j & \cdots \end{pmatrix} \quad \begin{matrix} \\ \\ \\ \end{matrix} \quad \bar{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$



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

$$X^T \beta = \underbrace{\begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix}}_{n \times d} \underbrace{\begin{pmatrix} \beta_1 \\ \vdots \\ \beta_d \end{pmatrix}}_{d \times 1}$$

(n rows, d columns; β has d entries)

$$n \times 1$$

$$X^T \beta \qquad Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \; n$$

$$\bar{Y} - \bar{X}\beta$$

$$= \begin{pmatrix} y_1 \\ | \\ | \\ y_n \end{pmatrix} - \underbrace{\bar{X}^T\beta}_{n \times 1}$$

.

$n \times 1$ $\qquad$ $n \times 1$ $\qquad$ (ok)

.

$$\| \bar{Y} - \bar{X}\beta \|_2^2$$

$$\|v\|_p = \left(\sum_{i=1}^{n} |v_i|^p \right)^{\frac{1}{p}}$$

$p \geq 1$

L_1-norm

$$\|v\|_1 = \left(\sum_{i=1}^{n} |v_i|^1 \right)^{1/1}$$

$$= \sum_{i=1}^{n} |v_i|$$

L_2-norm

$$\|v\|_2 = \left(\sum_{i=1}^{n} |v_i|^2 \right)^{1/2}$$

$$= \left(\sum_{i=1}^{n} v_i^2 \right)^{1/2}$$

$$\|v\|_2^2 = \sum_{i=1}^{n} v_i^2$$

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

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

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

$$\frac{\partial h}{\partial \beta} = 2\lambda\beta \quad \cdots$$

useful general identity

$$\frac{\partial (x^T A x)}{\partial x} = \left(A + A^T \right) x$$

$A \leftarrow Id \quad ; \quad x \leftarrow \beta$

- $g(\beta) = \frac{1}{n} \| Y - X^T\beta \|_2^2$

$$g(\beta) = \frac{1}{n} \{ (Y - X^T\beta)^T (Y - X^T\beta) \}$$

$$= \frac{1}{n} \{ Y^T Y - Y^T X^T \beta - \beta^T X Y + \beta^T X X^T \beta \}$$

$$(X^T\beta)^T = \beta^T (X^T)^T = \beta^T X$$

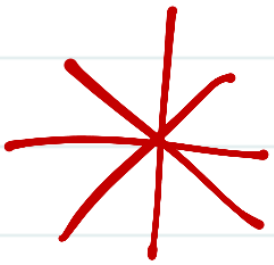
$$\frac{\partial g(\beta)}{\partial \beta}$$

$$= \frac{1}{n} \Big\{ 0 - \frac{\partial}{\partial \beta}\left((xy)^T \beta\right) * - \frac{\partial}{\partial \beta}\left(\beta^T (xy)\right) * + \frac{\partial}{\partial \beta}\left(\beta^T (xx^T)\beta\right) :$$

Useful formula ::

$$\frac{\partial}{\partial x}\left(a^T x\right) = \frac{\partial}{\partial x}\left(x^T a\right) = a$$

$$\frac{\partial}{\partial \beta}\left((xy)^T \beta\right) = xy$$



$$\frac{\partial}{\partial \beta}\left(\beta^T (xy)\right) = xy.$$

$$\frac{\partial}{\partial \beta}\left(\beta^T [\text{BLAH}]\, \beta\right)$$

$$(AB)^T = B^T A^T$$

$$= \left(\text{BLAH} + \text{BLAH}^T\right)\beta$$

$$\left(xx^T\right)^T = \left(x^T\right)^T x^T = xx^T$$

$$\frac{\partial}{\partial \beta}\left(\beta^T (xx^T)\beta\right).$$

$$= \left(xx^T + \left(xx^T\right)^T\right)\beta$$

$$= \left(xx^T + xx^T\right)\beta = 2xx^T\beta$$

Collecting everything

$$\frac{\partial g}{\partial \beta} = \frac{1}{n} \left\{ \begin{array}{l} -2xy \\ +2xx^T\beta \\ +0 \end{array} \right\}$$

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

$$\frac{\partial f}{\partial \beta} = \frac{\partial g}{\partial \beta} + \frac{\partial n}{\partial \beta}$$

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

Algo

Init $\beta^{(0)} = 0_{\mathbb{R}^d}$

Iterate for $t = 1, \ldots, T$

$$\beta^{(t+1)} = \beta^{(t)} - \frac{1}{L} \nabla_\beta F(\beta^{(t)})$$

Return: $\beta^{(T+1)}$

. L constant

$n = 1 \qquad d = 1$

$$F(\beta) = (y - x\beta)^2 + \lambda\beta^2$$

$$|\nabla F(\beta) - \nabla F(\beta')| \leq L\,|\beta - \beta'|$$

$$\nabla F(\beta) = -2x(y - x\beta) + 2\lambda\beta$$

$$\nabla F(\beta) - \nabla F(\beta')$$

$$= -2x(y - x\beta) + 2\lambda\beta$$

$$+ 2x(y - x\beta') - 2\lambda\beta'$$

$$= -2xy + 2x^2\beta + 2\lambda\beta$$

$$+ 2xy - 2x^2\beta' - 2\lambda\beta'$$

$$\nabla f(\beta) - \nabla f(\beta')$$

$$= 2x^2(\beta - \beta') + 2\lambda(\beta - \beta')$$

$$= 2(x^2 + \lambda)(\beta - \beta')$$

$$|\nabla f(\beta) - \nabla f(\beta')| \leq L|\beta - \beta'|$$

$$L = 2(x^2 + \lambda)$$