Useful things

1. Linearity of differentiation

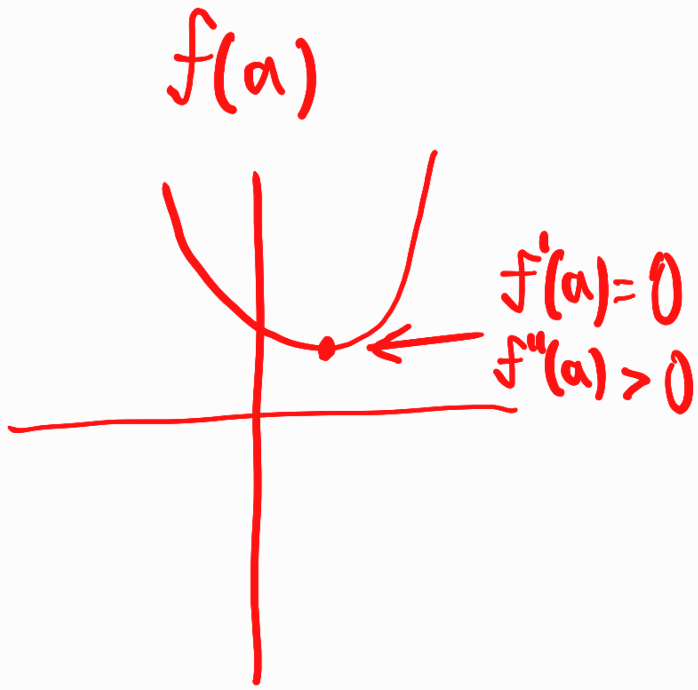
$$\frac{d}{d\alpha} \sum_i g_i(\alpha) = \sum_i g_i'(\alpha)$$

2. Conditions for minimizing a function $f(\alpha)$

minimum found when:

$$f'(\alpha) = 0$$

$$f''(\alpha) > 0$$



Consider the objective function

$$f(\alpha) = \sum_{i=1}^{n} (x^i - \alpha)^2$$

(i) Compute α^* that minimizes $f(\alpha)$

minimum found when:

$$f'(\alpha^*) = 0$$

$$f''(\alpha^*) > 0$$

$$f'(\alpha) = \frac{d}{d\alpha} \sum_{i=1}^{n} (x^i - \alpha)^2$$

linearity

$$= \sum_{i=1}^{n} \frac{d}{d\alpha} (x^i - \alpha)^2$$

chain

$\downarrow$ rule

$$= \sum_{i=1}^{n} 2(x^i - a)\frac{d}{da}(x^i - a)$$

$$= \sum_{i=1}^{n} 2(x^i - a)(-1)$$

$$f'(a) = -2\sum_{i=1}^{n}(x^i - a)$$

Need to find a^* s.t. $f'(a) = 0$

$$f'(a^*) = \cancel{-2}\sum_{i=1}^{n}(x^i - a^*) = 0$$

$$\Rightarrow \sum_{i=1}^{n} x^i - na^* = 0$$

$$\Rightarrow \sum_{i=1}^{n} x^i = na^*$$

$$\Rightarrow a^* = \frac{1}{n}\sum_{i=1}^{n} x^i$$

What is this?

Need to confirm its a minimum: $f''(a^*) > 0$

$$f'(a) = -2\sum_{i=1}^{n}(x^i - a)$$

$$f''(a) = \frac{d}{da} f'(a) = -2\sum_{i=1}^{n}\frac{d}{da}(x^i - a)$$

$$= -2\sum_{i=1}^{n}(-1)$$

$$= -2(-n)$$

$$f''(a) = 2n$$

$f''(a^*) = 2n > 0$ (assuming your sample size $n > 0$)

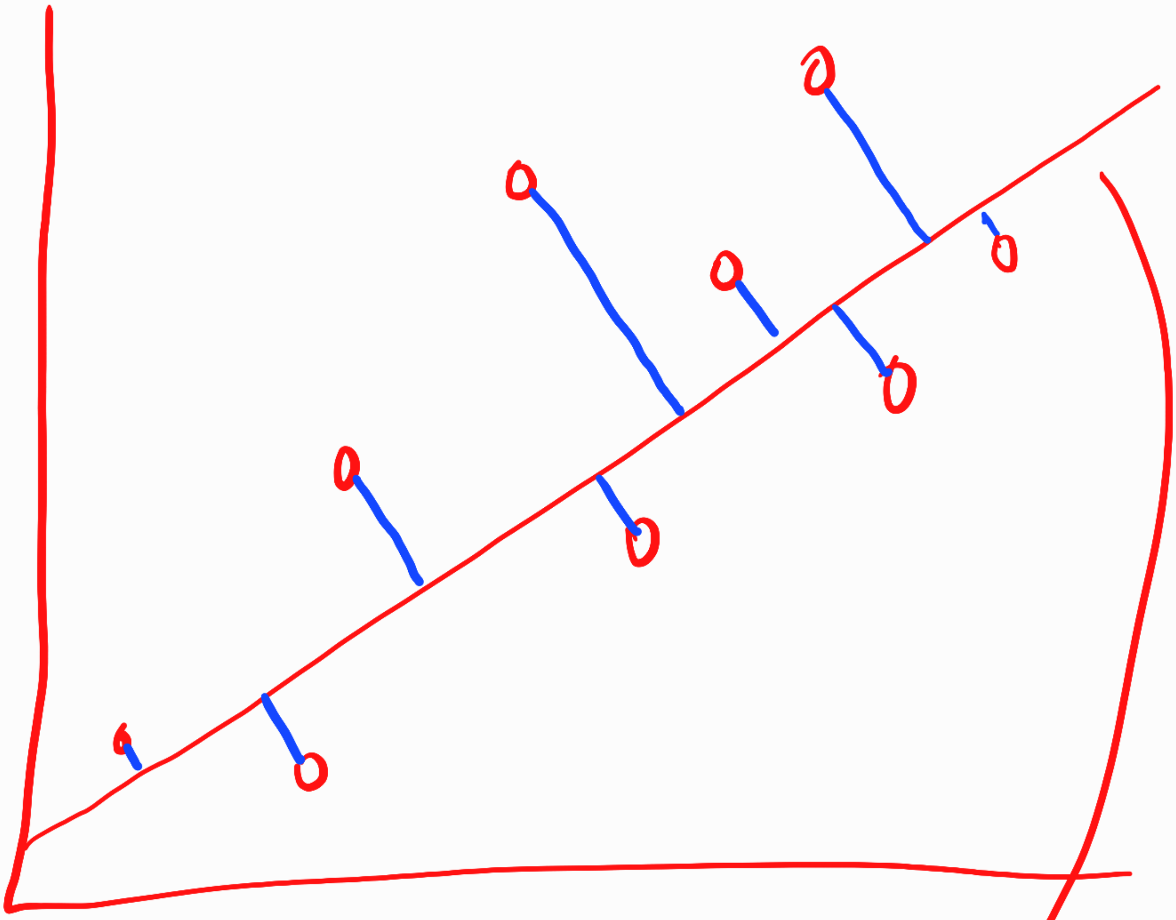
$\Rightarrow$ $\boxed{\text{For } a^*, \quad f'(a) = 0 \text{ and } f''(a) > 0 \quad \therefore f(a^*) \text{ is a minimum}}$

(ii) Intepret the solution in terms of the given data $x^1, x^2, \dots, x^n$

$$a^* = \frac{1}{n}\sum_{i=1}^{n} x^i = \bar{x}$$

The sample mean!

(iii) In probabilistic models, we are trying to find the parameter(s) that maximizes/minimizes some value (e.g. minimizes the loss)



$$y = b_1 x + b_0$$

$\text{loss}(b_1, b_0) =$ "sum of distances of sample points from line"

Idea: $\min_{b_1, b_0} \text{loss}(b_1, b_0)$ minimizes the distance – makes the best line.

(Optional Exercise)

(i) The distribution of N I.I.D. sample means approximates gaussian once N gets large enough.

(ii) Because of this, data in in the real would (which is always based on samples) is often model-able with a gaussian model.