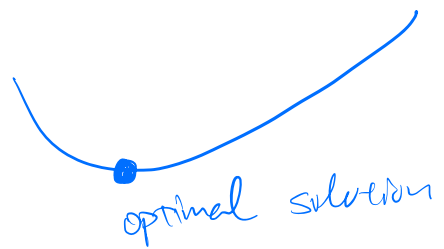


- Goal:

$$\underset{w}{\text{minimize}} \quad f(w)$$



- assumptions here:

- $\exists \nabla^2 f(w)$

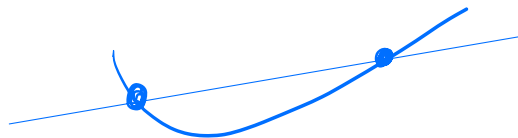
- convex function

$$f: \mathbb{R}^n \rightarrow \mathbb{R}$$

FACT f is called convex i.f.f.

$$\forall x_1, x_2, \forall t \in [0, 1]$$

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$



- rather to EULER to Rah....

FACT f is convex i.f.f. $f(x) \geq f(x_0) + \nabla f(x)^T (x - x_0)$,
 $\forall x, x_0$

FACT f is convex i.f.f. $\nabla f(x^*) = 0$ & $f(x^*)$ is min.
 i.f.f. $\nabla^2 f(x) \in \text{P.S.D.}$

e.g. linear regression, logistic regression

• Gradient descent

$$= W^{t+1} \leftarrow W - \alpha \nabla f(W^t)$$

- $\alpha > 0$ is the step size / learning rate (hyper-parameter)

$$= \text{stop if } \lim_{t \rightarrow \infty} \underbrace{\|\nabla f(W^t)\|}_{\text{we want}} = 0$$

$$\nabla f(x) \rightarrow 0$$

- first-order Taylor expansion

$$f(W+d) \approx g(d) := f(W^t) + \nabla f(W^t)d + \frac{1}{2\alpha} \|d\|^2$$

recall:

$$f(W+d) = f(W^t) + \nabla f(W^t)d + \frac{1}{2!} d^T \nabla^2 f(W^t) d + \dots$$

$$g(d) = f(W^t) + \nabla f(W^t)d + \frac{1}{2\alpha} \|d\|^2$$

$$d^* = \arg \min_d g(d)$$

$$\nabla g(d^*) = 0 \Rightarrow \nabla f(W^t) + \frac{1}{\alpha} d^* = 0 \Rightarrow d^* = -\alpha \nabla f(W^t)$$

- we can also do Newton's method yet two slow

- we hv best α

① a function is L -Lipschitz continuous: L is Lipschitz constant

$$\|f(x_1) - f(x_2)\|_2 \leq L \|x_1 - x_2\|_2$$

② a function is

L -smooth if its gradient is Lipschitz continuous:

$$\|\nabla f(x_1) - \nabla f(x_2)\|_2 \leq L \|x_1 - x_2\|_2$$

$$\nabla^2 f(x) \preceq LI$$

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{1}{2} L \|y - x\|^2$$

FACT let L be a Lipschitz constant

$$\nabla^2 f(x) \preceq LI \text{ for all } x$$

gradient descent converges if $\alpha < \frac{1}{L}$