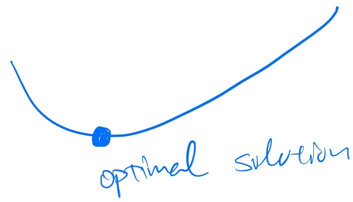
- Goal:

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



- assumptions here:
  - $\exists \ \nabla^2 f(w)$

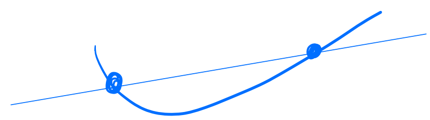
- convex function

$f: R^n \to R$

FACT f is called convex i.f.f.

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

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



- refer to ELEC5470 etc....

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. $\nabla^2 f(x) \in$ 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)
- stop if $\lim_{t \to \infty} \underbrace{\|\nabla f(w^t)\|}_{} = 0$ → we want $\nabla f(x) \to 0$

- 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:

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

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

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

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

- we can also do newton's method yet two slow
- we kw best α

- a function is L-Lipschitz continuous:

L is lipschitz constant

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

- a function is L-smooth if its gradient is Lipschitz continuous:

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

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

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

<u>FACT</u> let L be a Lipschitz constant

$\nabla^2 f(x) \preceq LI$ for all ,

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