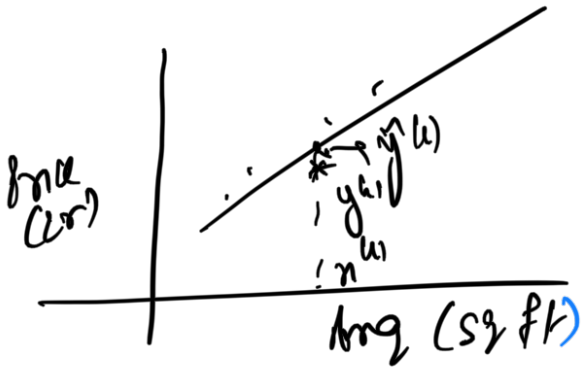
COL774
Machine Learning
Sat Aug 21, 2021

Make-up:-
Sunday Aug 22, 2021
8am-9am

Last class:-
Linear Regression:-



$\hat{y} = h_\theta(x) = \theta_1 x_1 + \theta_0$

↳ ① Hypothesis Space?
↳ linear
$h_\theta(x) = \theta^T x$
↳ Hypothesis $\sum_{j=0}^{d} \theta_j x_j$ $x_0 = 1$

② A good fit?

Loss fn:- $J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left[y^{(i)} - h_\theta(x^{(i)}) \right]^2$

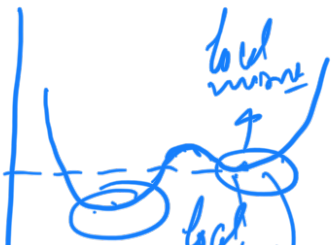
③ optimization: $\arg\min_\theta J(\theta)$

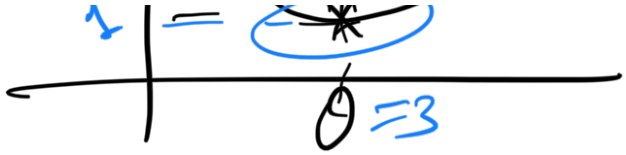
How to we compute $\arg\min_\theta J(\theta)$

↳ Detour:-
~~multivariate~~ Univariate Case

Local minima
global minima

$f(\theta) = (\theta - 3)^2 + 1$





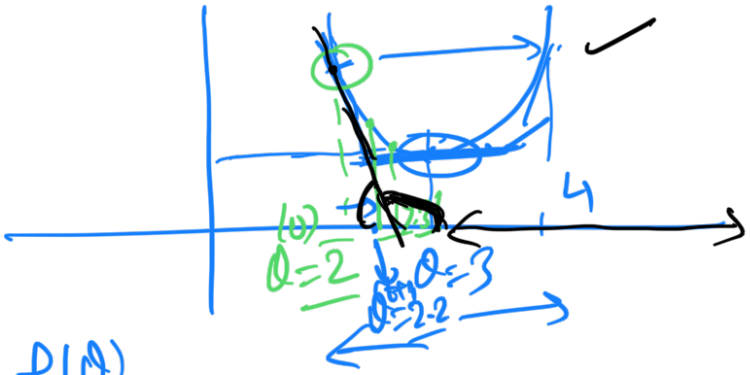
⇒ often be interested in finding global minima ⇒ local minima

Convex function ⇒ Convex function:- Local minima ≡ Global minima

"Gradient Descent":-

"Gradient Ascent" →

$f(\theta) = (\theta-3)^2 + 1$



$f(\theta)$

$f'(\theta) = 2(\theta-3)|_{\theta=2}$

$= -2$

Convex fn

At local minima:

$f'(\theta) = 0$ at any (local) minima for f

$\frac{\partial f(\theta)}{\partial \theta}$

$f'(\theta) = 2(\theta-3) + 0$

↳ Finding zeros of gradient

$\Rightarrow f'(\theta) = 0 \Rightarrow \boxed{\theta = 3}$

$\frac{\partial^2 f(\theta)}{\partial \theta^2} = 2 > 0$

$\Rightarrow \theta = 3$ is of local min.

⇒ Interested:- To find the points of local minima of a fn ⇒ $\frac{\partial f(\theta)}{\partial \theta} = 0$] may not always be possible

Gradient Descent:-

analytically

$$\theta^{(t+1)} \leftarrow \theta^{(t)} - \eta \left.\frac{\partial f(\theta)}{\partial \theta}\right|_{\theta^{(t)}}$$

(η: learning rate)

→ $t=0$

$\theta^{(t)} = 2$ $\quad \eta = 0.1$

$$\theta^{(t+1)} = 2 - 0.1\left(\left.2(\theta-3)\right|_{\theta=2}\right)$$

$$= 2 - 0.1 * 2(-1)$$

$$= 2 + .2 = 2.2$$

$t \leftarrow t+1 \quad | t=1$

$$\theta^{(t+1)} = 2.2 - 0.1\left(\left.2(\theta-3)\right|_{2.2}\right)$$

$$= 2.2 - 0.1 * 2(2.2-3)$$

$$= 2.2 + .16 = 2.36$$

$\theta^{(2)} = 2.36$

→ $t \leftarrow t+1 \quad | t=2$

$$\theta^{(t+1)} = \theta^{(t)} - \eta \cdot f'(\theta)|_{\theta^{(t)}}$$

$$= 2.36 - 0.1 * \left[\left.2(\theta-3)\right|_{2.36}\right]$$

$$= 2.36 - 0.1 * 2[\ldots]$$

$= 2.36 + .128$

$= \underline{2.488} \rightarrow$

⇒ Using this process, we can get arbitrarily close to local minima if you choose sufficiently small η

Gradient Descent:-

$f(\theta)$ Univariate Cost

$t \leftarrow 0;$

$\theta^{(t)} \leftarrow \text{init}();$

do {

$\theta^{(t+1)} \leftarrow \theta^{(t)} - \eta \cdot \frac{\partial f(\theta)}{\partial \theta}$

$t \leftarrow t+1;$

} while (! converged);

Multivariate Cost: $J(\theta) = f_{\theta}(\theta_1, \theta_0)$

$f(\theta) =:$ Monotone

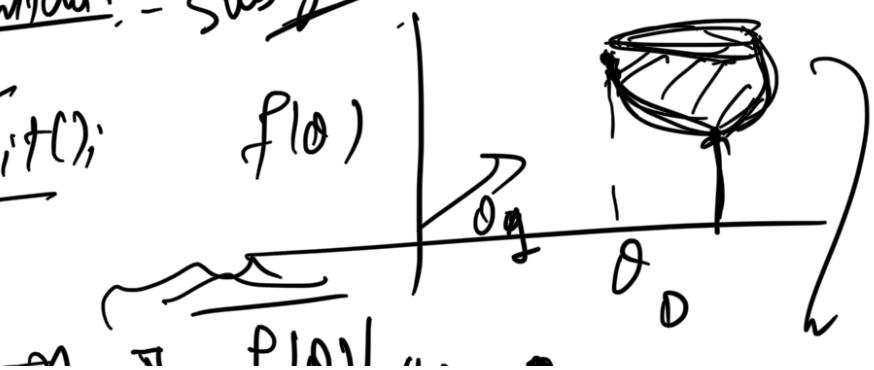
Gradient Descent Algorithm:- Sub-gradient $f_\theta(x) = \theta_1 x + \theta_0$

θ

$t=0$

$\theta^{(t)} \leftarrow \text{init}();$

do {

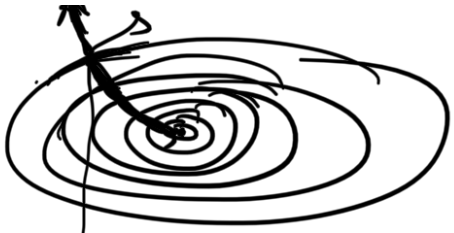


$\theta^{(t+1)} \leftarrow \theta^{(t)}$ — $\nabla_\theta f(\theta)|_{\theta^{(t)}}$

$t \leftarrow t+1$

3 while (! converged)

$\nabla_\theta f(\theta)$:- vector :-



$$\begin{bmatrix} \frac{\partial f(\theta)}{\partial \theta_n} \\ \frac{\partial f(\theta)}{\partial \theta_n} \\ \vdots \\ \frac{\partial f(\theta)}{\partial \theta_0} \end{bmatrix}_{n+1} \quad \theta \in \mathbb{R}^{n+1}$$

contour plot

"Newtons Method"

$$\left[\theta^{(t+1)} \leftarrow \theta^{(t)} - \eta \cdot \nabla_\theta f(\theta) \right] \Leftrightarrow \forall j :- \theta_j^{(t+1)} \leftarrow \theta_j^{(t)} - \eta \cdot \frac{\partial f(\theta)}{\partial \theta_j}\Big|_{\theta^{(t)}}$$

~~$\Leftarrow \theta_j^{(t)} - \eta \partial f(\theta)$~~

$\Leftrightarrow \forall j :- \quad j \in \{0 - n\}$

$\Rightarrow$:-
1. Impact of η
2. How do we decide convergence

---

① Adaptive η

η :- depends of learning iteration no.

$\eta \propto \frac{1}{t}$ or $\frac{1}{\sqrt{t}}$

② Different η for different coefficients

↳ "Large Diff." $\Rightarrow$

$$\eta = \frac{y_0}{t} \propto \frac{v_0^2}{\sqrt{t}}$$