COL 774

Aug 2?, 2021

Yesterday:- Linear Regression; optimization / finding the local optima.

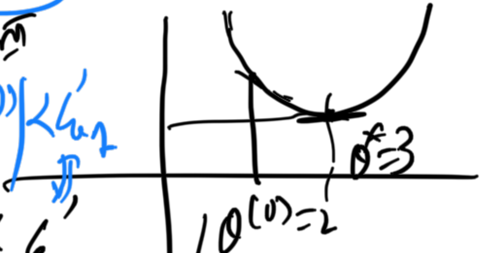
$f(\theta) = (\theta - 3)^2 + 1$

Gradient Descent :-

- Generally applicable
- → Does not make any assumption, $\theta \in \mathbb{R}^{n+1}$
- → Any function (differentiable)

```
t = 0
θ^(t) ← init();
do {
    θ^(t+1) ← θ^(t) − η · ∇_θ f(θ)      [η: learning rate; ∇_θ f(θ): gradient]
    t ← t+1;
} (while ! converged);
```

"overfitting" / early stopping



$\theta^{(0)} = 2$

$\theta^{(1)} = 2.2$

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

When do we stop? How close is really close?

↳ Converged:-

① $|\theta^{(t+1)} - \theta^{(t)}| < \epsilon_0$ ⟺ ② $\|\nabla_\theta f(\theta)\|_1 < \epsilon_0'$

$\forall j: |\theta_j^{(t+1)} - \theta_j^{(t)}| < \epsilon_{0,1}$

one-norm: $\forall j: \left|\frac{\partial f(\theta)}{\partial \theta_j}\right| < \epsilon_{0,1}'$

$\sum_j \left|\frac{\partial f(\theta)}{\partial \theta_j}\right|$

③ $|f(\theta^{(t+1)}) - f(\theta^{(t)})| < \epsilon$ ← threshold

$\theta^{(t+1)} = \theta^{(t)} - \nabla_\theta f(\theta) \cdot \eta$

$\epsilon_0' \Leftarrow \epsilon_0 = \epsilon_0' * \eta$

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

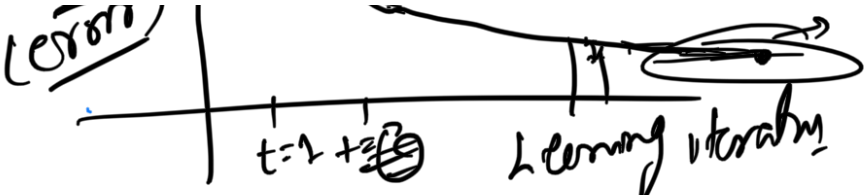
$\theta^{(t)}_{t=0} = 2$

$= J(\theta)$

Gradient Descent

SGD — Stochastic Gradient Descent

"converged"

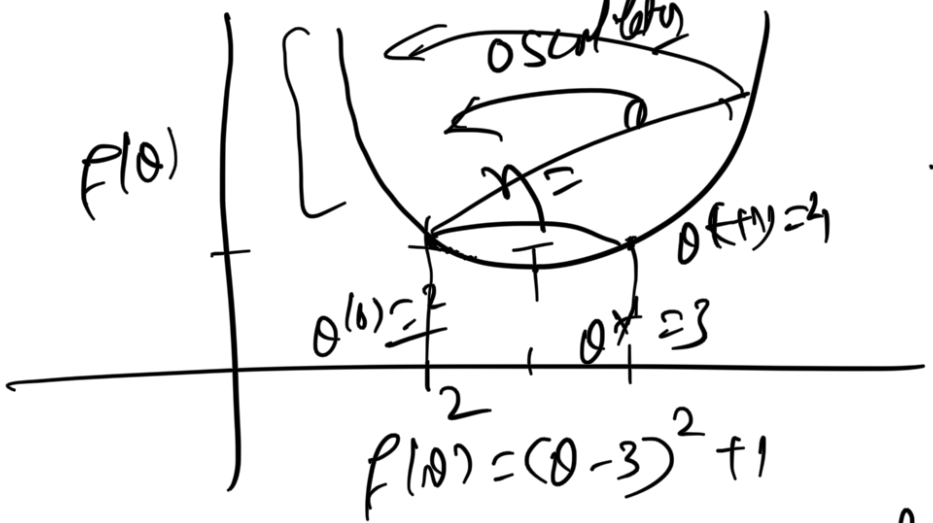


$\Rightarrow$ often!:- & Run for a $\begin{cases} \text{"certain" no. of} \\ \text{(large enough)} \\ \text{iterations} \end{cases}$ "epoch"

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

$\Rightarrow$ Can be expensive if m is large.

Choosing "right" η :-



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

$$\theta^{(t+1)} \leftarrow \theta^{(t)} - \eta \cdot \nabla_\theta J(\theta)$$

$$\nabla_\theta J(\theta) = 2(\theta - 3)$$

$\eta = 0.1$

$\theta^{(0)} = 2$

$\theta^{(1)} = 2.2$

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

$\vdots$

$\Rightarrow$ what happens if η :-

(a) too large

too fast

$t=0$ $\eta = 1$ $\theta^{(0)} = 2$

$\theta^{(t+1)} \leftarrow 2 - 1 \cdot 2(2-3)$

$= 2 - (-2)$

$= 4$

$t = 1$

$\theta^{(t+1)} \leftarrow 4 - 1.2(4-3)$

$[2(\theta-3)]$ gradient $= 4 - 2 = 2$ $\theta=4$

& $\gamma \geq 1$

⇒ (b) γ :- too small.

Too slow

$t=0$ $\gamma = .01$ $\theta^{(t)} = 2$

$\theta^{(t+1)} \leftarrow 2 - .01 * 2(2-3)$

$= 2 - (-).02 = 2.02$

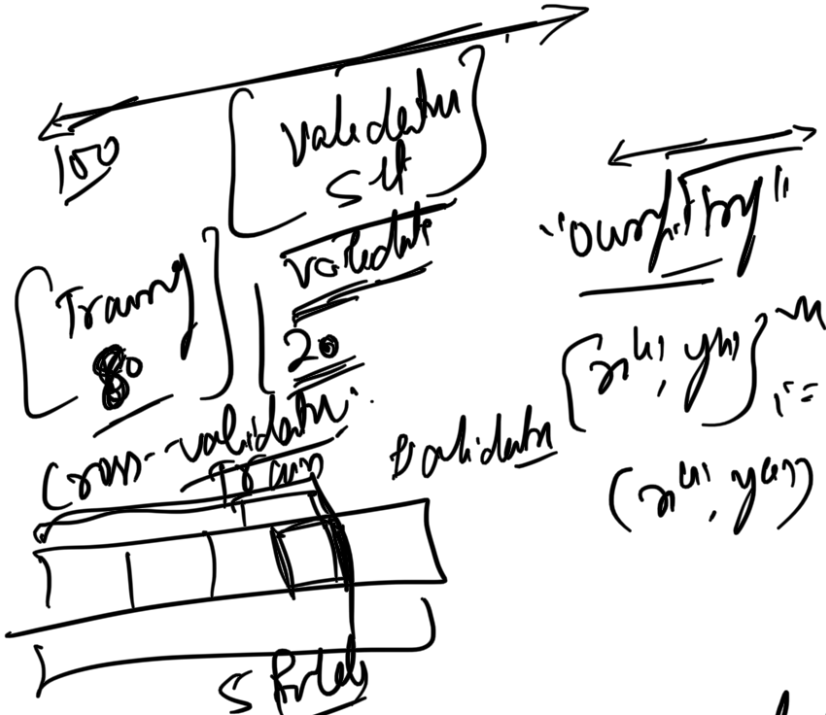
$t = 1$

$\theta^{(t+1)} \leftarrow = 2.02 - .01 * 2(2.0 - 3$

$= 2.02 - \underset{*.01}{(-1)}(1.96)$

$= 2.02 + .0196$

$= 2.0396$



"our Trng" $\{x^{(i)}, y^{(i)}\}_{i=1}^{m}$

$(x^{(i)}, y^{(i)}) \sim$ Dist

Example:- computing partial Derivatives

$\theta_0^2 + \theta_1^2 + 2\theta_0\theta_1 + 2\theta_0 + 3\theta_1$

$f(\theta) = \theta_1 + \theta_2 + \dots$

$\nabla_\theta f(\theta) = \begin{bmatrix} 2\theta_2 + \\ 3\theta_1 + 2 \\ 2\theta_1 + 3\theta_2 \end{bmatrix}$ $\leftarrow$ $\begin{matrix} \frac{\partial f(\theta)}{\partial \theta_2} \\ \frac{\partial f(\theta)}{\partial \theta_1} \end{matrix}$ $\Big|_{\theta_2 = 1, \theta_1 = 1}$

$= \begin{bmatrix} 7 \\ 5 \end{bmatrix}$

Linear Regression:

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

$$\nabla_\theta J(\theta) = \downarrow = \frac{1}{2m} \sum_{i=1}^{m} \underbrace{\left[y^{(i)} - \theta^T x^{(i)} \right]^2}_{\text{Quadratic fn}}$$

$\Rightarrow$ Unique Local minima of θ (Convex fn.)