COL 774
Machine Learning
Aug 31, 2021

Last class:-

1. Normalization]
2. Probabilistic interpretation of linear regression:-

$y^{(i)} | x^{(i)}; \theta \sim N(\theta^T x^{(i)}, \sigma^2)$

i.i.d $\epsilon \sim N(0, \sigma^2)$

↳ ML estimate

$$\arg\max_\theta \left[\prod_{i=1}^{m} P(y^{(i)} | x^{(i)}; \theta) \right] \equiv \arg\min_\theta J(\theta)$$

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

Newton's Method for Optimization:-

Till Now:-

$\arg\min_\theta f(\theta) \rightarrow \arg\min_\theta J(\theta)$

$\theta \in \mathbb{R}^{n+1}$

uses first order information [↳ gradient descent] converges to local minima. convex:- global minima.

$\nabla_\theta f(\theta)$

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

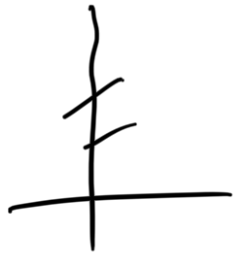
↳ uses second order information $[\nabla^2_\theta f(\theta))$ (curvature) $\equiv$ (Hessian matrix)]

Hopefully:- that you can make "faster" progress

Newton's Method for Optimization :-

↳ for finding zeroes of a function.

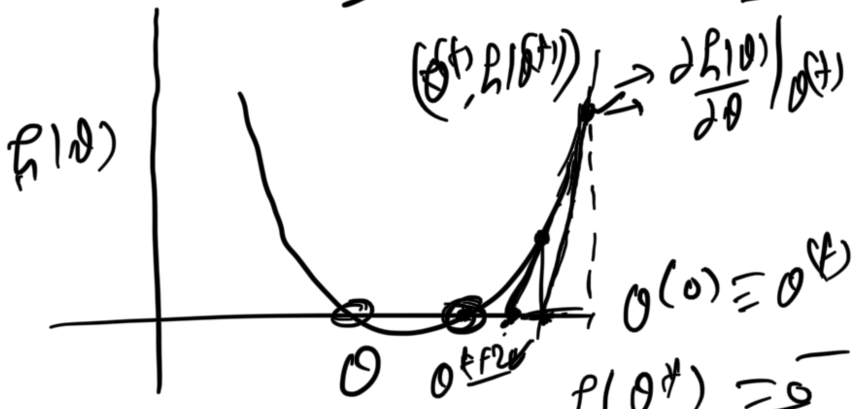
$f(\theta) = \cancel{(\theta-3)^2 + 1}$



Newton's method :-

$\theta \in \mathbb{R}$

$h(\theta)$ | find zeroes of this function.



$\theta^{(0)} = \theta^{(t)} \mid t=0$

$h(\theta^*) = 0$

$(\theta^{(t+1)}, 0)$

(i) iteration method :-

$$\frac{h(\theta^{(t+1)}) - h(\theta^{(t)})}{\theta^{(t+1)} - \theta^{(t)}} = h'(\theta)\Big|_{\theta^{(t)}}$$

$$\boxed{\theta^{(t+1)} - \theta^{(t)} = -\frac{h(\theta^{(t)})}{h'(\theta^{(t)})}}$$

↳

$$\theta^{(t+1)} = \theta^{(t)} - \frac{h(\theta^{(t)})}{h'(\theta^{(t)})}$$

Newton's update | for finding zeroes of a function

↳ (i) how does it converge

with our problem?

$\arg\min_\theta f(\theta)$] convex]

finding zeros of $f'(\theta)$]

$\leftarrow \nabla_\theta f(\theta)$ | $\left[\frac{\partial f(\theta)}{\partial \theta}\right]$

$\Rightarrow$ Even if function is not convex, we will still find (local optima)

Newton's method for optimization:-

$\arg\min_\theta f(\theta)$

$$\left[\theta^{(t+1)} \leftarrow \theta^{(t)} - \left[\left(\frac{f'(\theta^{(t)})}{f''(\theta^{(t)})}\right)\right]\right.$$

Newton's update

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

$\theta^{(t)}$

Newton's update:-

$$\theta^{(t+1)} \leftarrow \theta^{(t)} - \frac{\cancel{2}(\theta-3)}{\cancel{2}}\Big|_{\theta^{(t)}}$$

$$\boxed{\theta^{(t+1)} \leftarrow 3}$$

$\therefore$ Newton's method converges in single iteration.

"faster" numerical maximizer

$f''(\theta) > 0$ minima

< 0 maxima

Higher Dimensions:-

univariate: scalar update

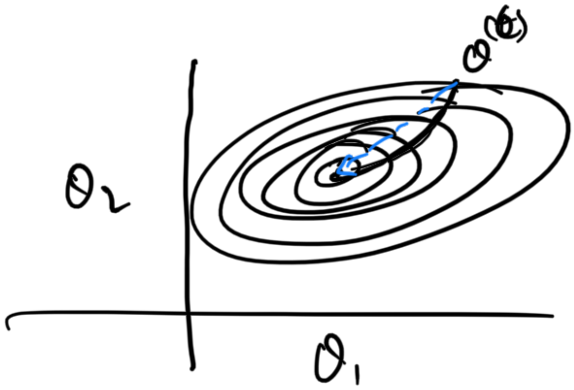
$$\theta^{(t+1)} \leftarrow \theta^{(t)} - [f''(\theta^{(t)})]^{-1} [f'(\theta^{(t)})]$$

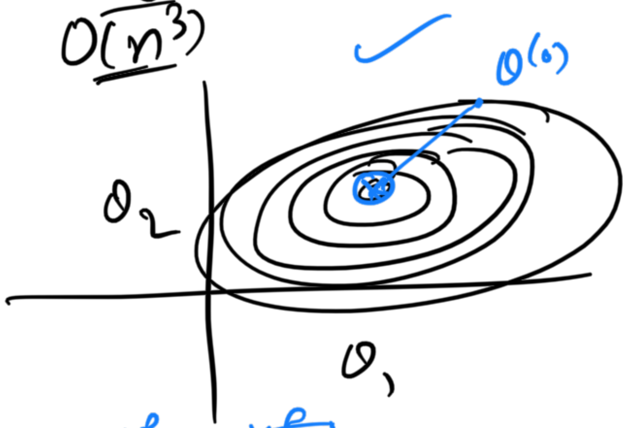
multivariate:

$$[\theta^{(t+1)} \leftarrow \theta^{(t)} - H^{-1}\big|_{\theta^{(t)}} \nabla_\theta f(\theta) \big|_{\theta^{(t)}}]$$

vector update

$O(n^3)$



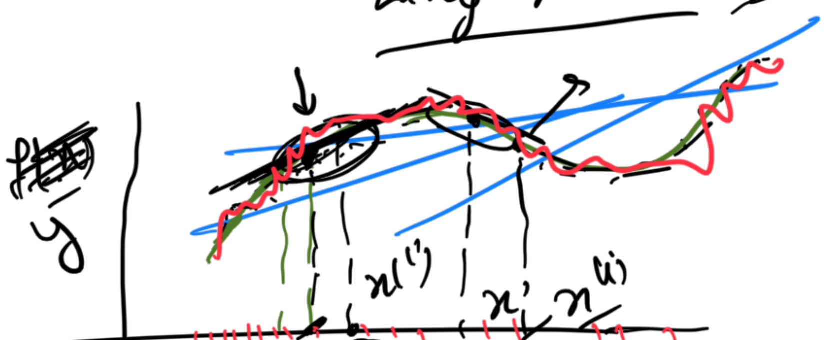


[why would one even prefer gradient descent over Newton's method?]

Locally weighted Linear Regression:-

Non-parametric methods:-

"Lazy Learners"



$h_\theta(x) \approx y$

Multiple Linear Approximations to the fn

linearity

$\{x^{(i)}, y^{(i)}\}_{i=1}$

inversely proportional to distance of $x^{(i)}$ from x

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

↳ 1

$$w^{(i)} = e^{-\frac{(x - x^{(i)})^2}{2\tau^2}} \rightarrow \tau : \text{bandwidth}$$

locally weighted linear regression.

underfit

$\tau \rightarrow \infty \qquad w^{(i)} \rightarrow e^{-0} = 1$

↳ Linear Regression

overfit

$\tau \rightarrow 0 \qquad$ You pay more & more attention to close by examples

"overfit"

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

n×n matrix

$$w^{(i)} \leftarrow e^{-\frac{(x - x^{(i)})^T \Sigma^{-1} (x - x^{(i)})}{2}}$$