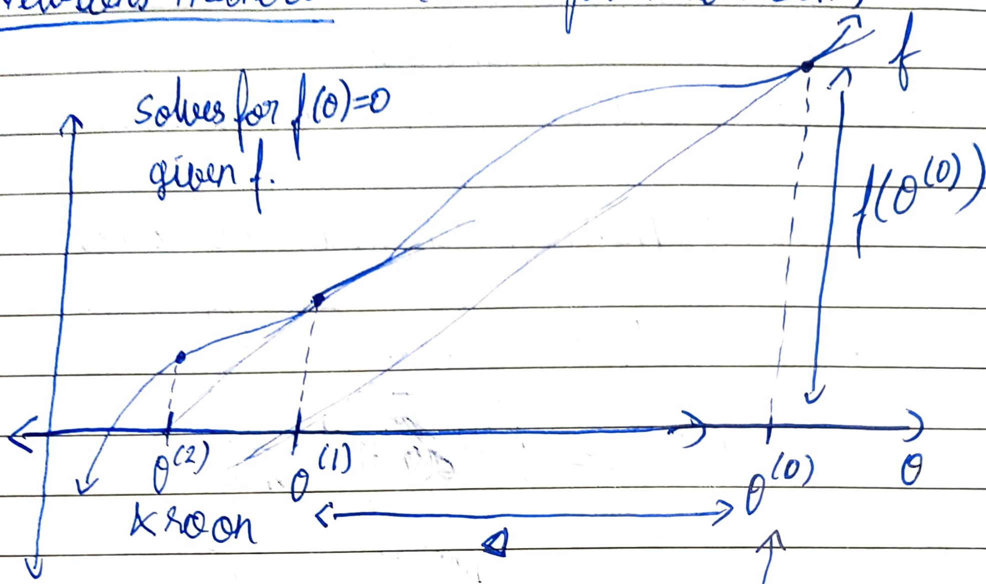
Newton's method – (much faster than BGD)



⇒ start with random $\theta^{(0)}$, find tangent at $f(0)$ & use that as a first order approximation; solve for $(f(\theta))^{linear} = 0$ & set that to next θ.
& repeat!

$$\theta^{(1)} := \theta^{(0)} - \Delta$$ → must solve for this.

clearly, $f'(\theta^{(0)}) = \dfrac{f(\theta^{(0)})}{\Delta}$

$$\Rightarrow \Delta = \frac{f(\theta^{(0)})}{f'(\theta^{(0)})}$$

$$\therefore \theta^{(t+1)} := \theta^{(t)} - \frac{f(\theta^{(t)})}{f'(\theta^{(t)})}$$

⇒ in our case $f(\theta) = l'(\theta)$, log likelihood ($\because$ at $l'(\theta)=0$ max.)

$$\therefore \theta^{(t+1)} := \theta^{(t)} - \frac{l'(\theta^{(t)})}{l''(\theta^{(t)})}$$

Newton's method enjoys "Quadratic Convergence")
i.e. no. of significant digits that we have converged to minimum doubles every iteration.

⇒ near the min. very fast convergence.

✩ when θ is a vector: called the Newton-Raphson method

$$\theta^{(t+1)} := \theta^{(t)} - \underbrace{H^{-1}}_{\mathbb{R}^{(n+1)\times(n+1)}} \underbrace{\nabla_\theta l}_{\text{vector } \mathbb{R}^{n+1}}$$

H is the Hessian Matrix.

$$H_{ij} = \frac{\partial^2 l}{\partial \theta_i \partial \theta_j}$$

∴ very expensive for high dimensional θs.

✩✩ when Newton's Method is used for maximizing $l(\theta)$ of logistic regression, the method is also called Fisher Scoring. ✩✩