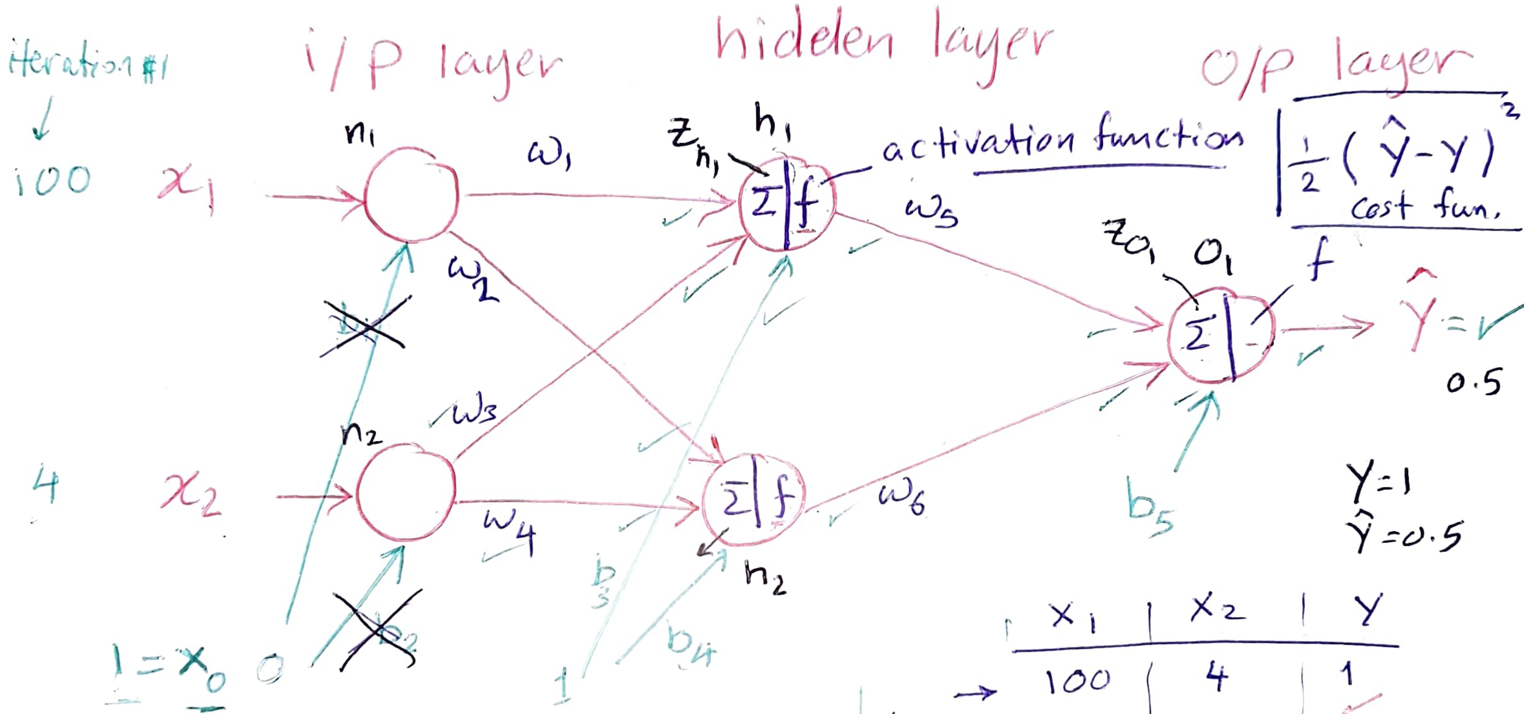
Numerical Optimization, AI45 Mans. session 5

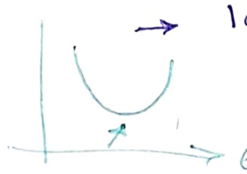
11/2/2025

$x_0\theta_0 + \theta_1 x_1 + \cancel{\theta_2 x_2}$

→ Neural Network



x_1	x_2	Y
100	4	1



at h_1 :

$$z_{h_1} = x_1 \omega_1 + x_2 \omega_3 + b_3$$

$$Out_{h_1} = f(z_{h_1}) = f(x_1 \omega_1 + x_2 \omega_3 + b_3)$$

$$Out_{h_1} = \frac{1}{1+e^{-z_{h_1}}} = \frac{1}{1+e^{-(x_1\omega_1 + x_2\omega_3 + b_3)}}$$

$$Out_{h_2} = f(x_1 \omega_2 + x_2 \omega_4 + b_4)$$

$$= \underline{\qquad\qquad}$$

$$z_{o_1} = Out_{h_1} w_5 + Out_{h_2} w_6 + b_5$$

$$\hat{y} = f(z_{o_1}) = \frac{1}{1+e^{-(Out_{h_1} w_5 + Out_{h_2} w_6 + b_5)}}$$

$$= \frac{1}{1+e^{-z_{o_1}}},$$

ex. $\hat{y} - y = 0.5 =$ error $\Rightarrow J(\;)$

$$\frac{\partial J(\;)}{\partial w_5}$$

- update $w_{5_{new}} = w_{5_{old}} - \alpha \frac{\partial J(\;)}{\partial w_5}$
- update $w_{1_{new}} = w_{1_{old}} - \alpha \frac{\partial J(\;)}{\partial w_1}$

for all paramerters (w's and b's)

---

e.g., $J(w_i) = \frac{1}{2}(\hat{y} - y)^2$

$$\hat{y} = h_{w,b}(\vec{x}) = \boxed{}$$

$$\frac{\partial J(\;)}{\partial w_5} = \frac{\partial J(\;)}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial z_{o_1}} \times \frac{\partial z_{o_1}}{\partial w_5} = 0.1$$

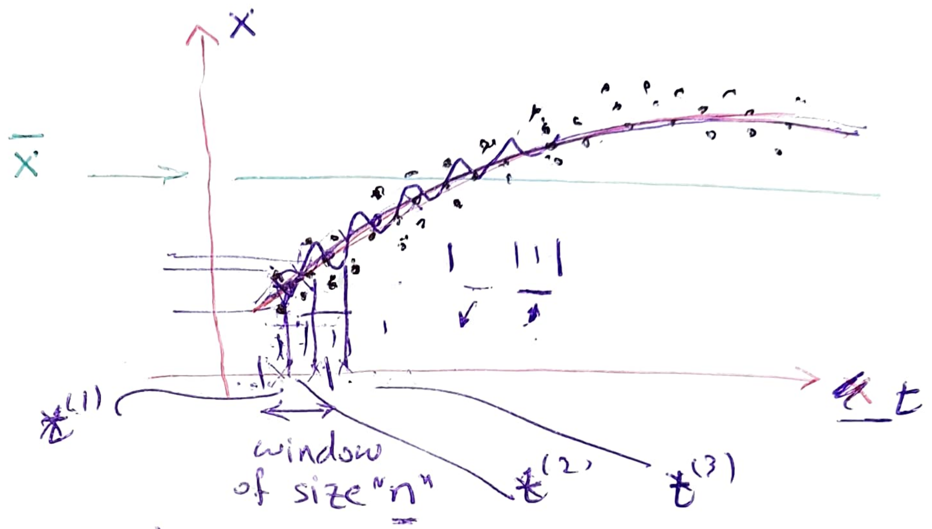
$\Rightarrow$ update $w_{5_{new}} = w_{5_{old}} - \alpha\, 0.1 = w_{5_{new}}$

$$\frac{\partial J(\;)}{\partial w_1} = \frac{\partial J}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial z_{o_1}} \times \overset{1.3}{\frac{\partial z_{o_1}}{\partial out_{h_1}}} \times \frac{\partial Out_{h_1}}{\partial z_{h_1}} \times \frac{\partial z_{h_1}}{\partial w_1}$$

$\rightarrow$ Chain Rule $\leftarrow$

→ Exponentially Weighted Moving Average (EWMA)

average



$$v^{(t)} = \beta\, v^{(t-1)} + (1-\beta)x$$

$$v^{(0)} = 0$$

$\hat{v}_3^{(1)}$

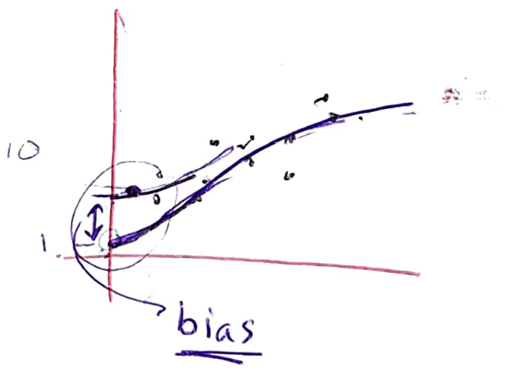
$$v^{(1)} = \beta\, v^{(0)} + (1-\beta)x^{(1)}$$
$$\searrow 0$$

$$v^{(2)} = \beta\, v^{(1)} + (1-\beta)\, x^{(2)}$$
$$= \beta(1-\beta)x^{(1)} + (1-\beta)\, x^{(2)}$$

$$v^{(3)} = \beta v^{(2)} + (1-\beta)\, x^{(3)}$$

$$v^{(3)} = \cancel{\beta}\, (\beta^2(1-\beta))x^{(1)} + \beta(1-\beta)\, x^{(2)} + (1-\beta)\, x^{(3)}$$

$$v^{(4)} = \beta^3(1-\beta)\, x^{(1)} + \beta^2(1-\beta)x^{(2)} + \beta(1-\beta)x^{(3)} + (1-\beta)\, x^{(4)}$$

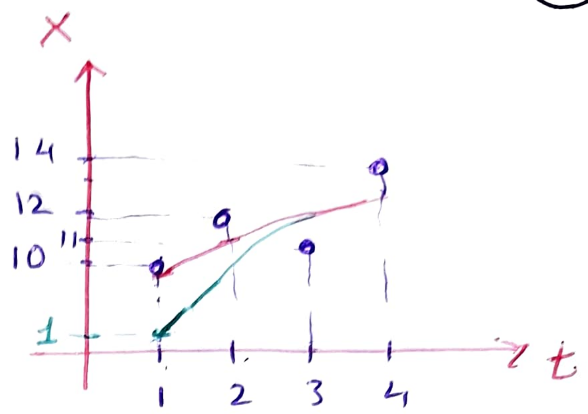


for bias Correction

$$\hat{v}(t) = \frac{v(t)}{1-\beta^t}$$

Bias correction

let $\beta = 0.9$



$v^{(0)} = 0$

$$v^{(1)} = \beta v^{(0)} + (1-\beta) x^{(1)}$$
$$= 0.9 \times 0 + 0.1 \times 10$$
$$= 1$$

$$v^{(2)} = \beta v^{(1)} + 0.1 \times 12$$
$$= 1 + \cancel{1.2}$$

Bias corrected

$$\hat{v}^{(1)} = \frac{v^{(1)}}{1-\beta^1}$$
$$= \frac{1}{1-0.9^1} = \frac{1}{0.1} = 10$$

$\vdots$

$$\hat{v}^{(5)} = \frac{v^{(5)}}{1-\beta^5}$$

1

=

for large time index.

$$1-\beta^t \approx 1$$

$$\hat{v}^{(t)} \approx v^{(t)}$$

Review Momentum

Standard form

$$\nu^{(t)} = \beta\, \nu^{(t-1)} + \alpha\, \nabla(J(\theta^{(t)}))$$

$$\theta^{(t+1)} = \theta^{(t)} - \nu^{(t)}$$

$$\theta^{(t+1)} = \theta^{(t)} - \beta\, \nu^{(t-1)} - \alpha \cdot \text{grad}.$$

EWMA form

$$\nu^{(t)} = \beta\, \nu^{(t+1)} + (1-\beta)\, \nabla(J(\theta^{(t)}))$$

velocity, $\nu^{(t)}$

momentum, $m^{(t)}$

$$\theta^{(t+1)} = \theta^{(t)} - \alpha\, \nu^{(t)}$$

$$\theta^{(t-1)} = \theta^{(t)} - \alpha \beta\, \nu^{(t+1)} - \alpha(1-\beta) \cdot \text{grad}.$$

scaled learning rate

$$\boxed{\tilde{\alpha} = \frac{\alpha}{1-\beta}}$$

Review RMSprop.

$$v^{(t)} = \beta\, v^{(t-1)} + (1-\beta)(\text{grad}\,\ell\,)^2$$

$$\theta^{(t+1)} = \theta^{(t)} - \frac{\alpha}{\sqrt{v^{(t)}} + \epsilon} \nabla(J(\theta^{(t)}))$$

Adam (Adaptive Momentum)

EWMA momentum

$$m^{(t)} \equiv v_1^{(t)} = \beta_1 v_1^{(t-1)} + (1-\beta_1)\,\underbrace{\nabla(J(\theta^{(t)}))}_{\text{grad.}}$$

$$v_2^{(t)} = \beta_2\, v_2^{(t-1)} + (1-\beta_2)\left(\nabla J(\theta^{(t)})\right)^2$$

RMS prop.

$$\theta^{(t+1)} = \theta^{(t)} - \frac{\alpha}{\sqrt{\hat{v}_2^{(t)}} + \epsilon}\,\hat{v}_1^{(t)}$$

bias-corrected $v_2^{(t)}$

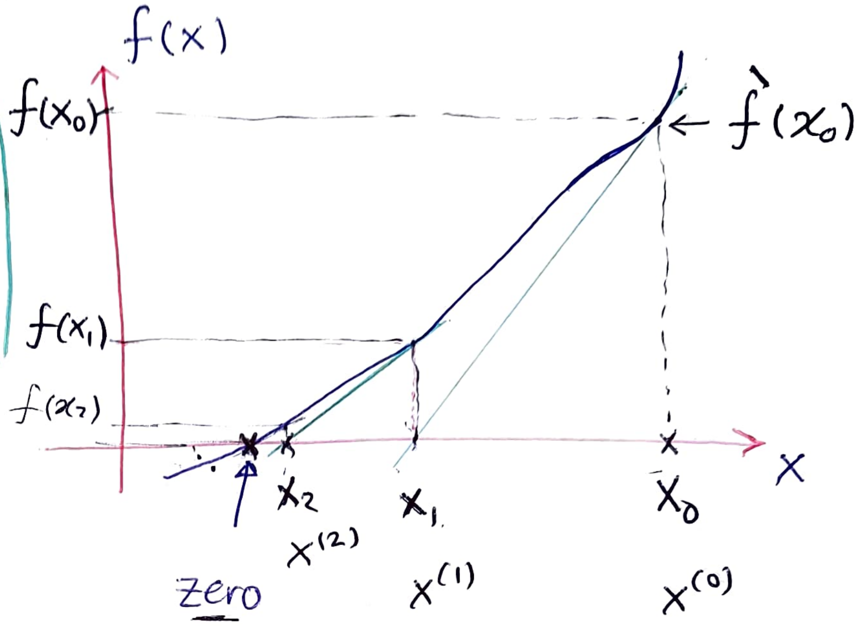
bias-corrected $v_1^{(t)}$

Commonly $\beta_1 = 0.9$ $\beta_2 = 0.999$ $\epsilon = 1\times10^{-8}$

Newton's Method

Zero of a function;

value of x s.t. $f(x)=0$



x_0 ✓

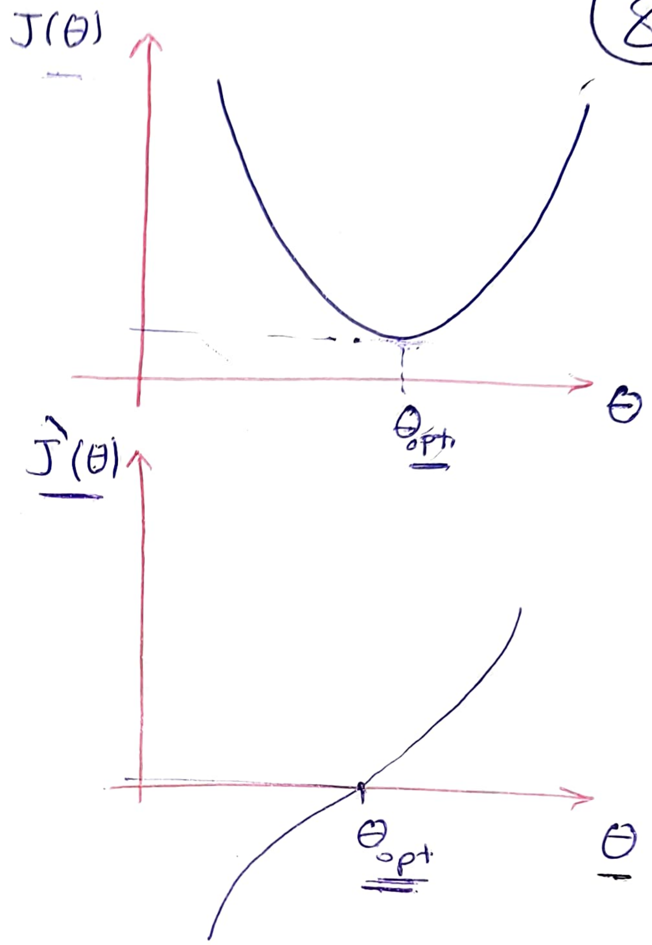
$$\Rightarrow f'(x_0) = \frac{\Delta Y}{\Delta X} = \frac{f(x_0) - 0}{x_0 - x_1}$$

$$f'(x_0) = \frac{f(x_0)}{x_0 - x_1}$$

$$\Rightarrow x_0 - x_1 = \frac{f(x_0)}{f'(x_0)} \Rightarrow x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}$$

$$x^{(t)} = x^{(t-1)} - \frac{f(x^{(t-1)})}{f'(x^{(t-1)})}$$



Zero of $J'(\theta)$

$\Rightarrow$ minimum of $J(\theta)$

(or maximum, or saddle point, or local minimum!)

→ use Newton's method to find zero of $J'(\theta)$

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

---

for multivariate cases

$$\vec{\theta}^{(t+1)} = \vec{\theta}^{(t)} - \cancel{\frac{\nabla J(\theta^{(t)})}{?}}$$

$$\vec{\theta}^{(t+1)} = \vec{\theta}^{(t)} - (H^{-1}) \nabla J(\theta^{t})$$

GD $O(n)$ complexity ? $O(n^3)$

Remember

gradient $\nabla f = \begin{bmatrix} \partial f / \partial \theta_1 \\ \vdots \\ \partial f / \theta_n \end{bmatrix}$

e.g., $f(x,y) \Rightarrow \nabla f = \begin{bmatrix} \partial f / \partial x \\ \partial f / \partial y \end{bmatrix}$

Hessian $H \equiv \begin{bmatrix} \frac{\partial f}{\partial x^2} & \frac{\partial f}{\partial x \partial y} \\ \frac{\partial f}{\partial x \partial y} & \frac{\partial f}{\partial y^2} \end{bmatrix}$

$H = \begin{bmatrix} - H_{i,j} \end{bmatrix}$

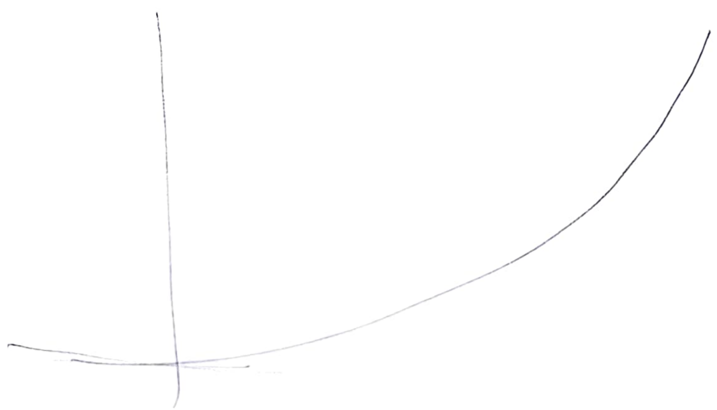
$$H_{i,j} = \frac{\partial f}{\partial \theta_i \partial \theta_j}$$

$$\nabla J(\theta_1, .., \theta_n) = \begin{bmatrix} \partial \\ \\ \\ \end{bmatrix}$$

- Quasi-Newton methods

- Secant method

- BFGS (Broyden, Fletcher, Goldfarb, Shanno,)



note

note that the same method can be expressed in different forms.

e.g., momentum method (standard form)

$$v^{(t)} = \beta v^{(t-1)} + \alpha \nabla J(\theta^{(t)})$$

$$\theta^{(t+1)} = \theta^{(t)} - v^{(t)}$$

OR

$$v^{(t+1)} = \beta v^{(t)} + \alpha \nabla J(\theta^{(t)})$$

$$\theta^{(t+1)} = \theta^{(t)} - v^{(t+1)}$$

OR

$$v^{(t+1)} = \beta v^{(t)} - \alpha \nabla J(\theta^{(t)}$$

$$\theta^{(t+1)} = \theta^{(t)} + v^{(t+1)}$$

and so on, ...

Regardless of how it is written, the parampeters are updated in the same manner.