1) As per the Newton's Method for 2-layer Neural Network:

$$x^* = x^{(0)} - H(f)(x^{(0)})^{-1} \nabla f(x^{(0)})$$

in our case, weight ~~vector~~ variable = w

$\therefore$ replacing x with w:

$$w^* = w^{(0)} - H(f)(w^{(0)})^{-1} \nabla f(w^{(0)}) \quad \text{——— (I)}$$

Given the cost function:

$$f = \frac{1}{2n} \sum_{i=1}^{n} (\hat{y}^{(i)} - y^{(i)})^2$$

for linear network, $\hat{y} = x^T w$

$$\therefore f(w) = \frac{1}{2n} \sum_{i=1}^{n} (x^T w - y^{(i)}) \quad \text{——— (II)}$$

taking gradient of eqn (II) wrt w

$$\nabla_w f(w) = \frac{1}{n}(X)(X^T w - Y) \quad \text{——— (III)}$$

again taking gradient of eqn (III) wrt w

$$H = \frac{1}{n}(X X^T) \quad \text{——— (IV)}$$

substituting values of H and ∇f in eqn (I):

$$w^* = w^{(0)} - \left(\frac{1}{n}(XX^T)\right)^{-1} \left(\frac{1}{n}(X)(X^T w^{(0)} - Y)\right)$$

Scanned with CamScanner

$$w^* = w^{(0)} - \left[\cancel{n}(XX^T)^{-1} \cdot \frac{1}{\cancel{n}}(XX^T w^{(0)} - XY)\right]$$

$$w^* = w^{(0)} - \left[(XX^T)^{-1} \cdot (XX^T w^{(0)} - XY)\right]$$

$$w^* = w^{(0)} - w^{(0)} + (XX^T)^{-1} XY$$

$$\boxed{w^* = (XX^T)^{-1} XY}$$

Therefore, it is clear from the above value that whatever be the starting point, the solution for a 2-layer Neural Network using Newton's method always converges to the optimal solution.

2

$$\hat{y}_k = \frac{\exp z_k}{\sum_{k'=1}^{c} \exp z_{k'}}$$

$$z_k = x^T w^{(k)} + b_k$$

$$f_{CE}(W,b) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{c} y_k^{(i)} \log \hat{y}_k^{(i)} \quad \text{———} \quad (a)$$

$$\nabla_{w^{(l)}} f_{CE}(W,b) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{c} y_k^{(i)} \nabla_{w^{(l)}} \log \hat{y}_k^{(i)}$$

$$= -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{c} y_k^{(i)} \left(\frac{\nabla_{w^{(l)}} \hat{y}_k^{(i)}}{\hat{y}_k^{(i)}}\right) \quad \text{———} \quad (1)$$

if $l = k$:

$$\nabla_{w^{(l)}} \hat{y}_k^{(i)} = \nabla_{w^{(l)}} \left(\frac{\exp z_k}{\sum_{k'=1}^{c} \exp z_{k'}}\right)^{(i)}$$

$$= \nabla_{w^{(l)}} \left(\frac{\exp(x^T w^{(l)} + b_l)}{\sum_{k'=1}^{c} \exp(x^T w^{(k')} + b_{k'})}\right)^{(i)}$$

$$= \frac{\left[x^{(i)} \exp(x^T w^{(l)} + b_l) \sum_{k'=1}^{c} \exp(x^T w^{(k')} + b^{k'})\right] - \left[\exp(x^T w^{(l)} + b_l) \cdot \left(x^{(i)} \exp(x^T w^{l} + b_l)\right)\right]}{\left(\sum_{k'=1}^{c} \exp(x^T w^{k'} + b_{k'})\right)^2}$$

$$= \frac{x^{(i)} \exp(x^T w^{(l)} + b^{(l)}) \left[\sum_{k'=1}^{c} \exp(x^T w_{k'} + b_{k'}) - \exp(x^T w_l + b_l)\right]}{\sum_{k'=1}^{c} \exp(x^T w_{k'} + b_{k'}) \sum_{k'=1}^{c} \exp(x^T w_{k'} + b_{k'})}$$

$$= x^{(i)} \hat{y}_l^{(i)} \left[1 - \frac{\exp(x^T w_l + b_l)}{\sum_{k'=1}^{c} \exp[x^T w_{k'} + b_{k'}]} \right]$$

$$\boxed{\nabla_{w^{(l)}} \hat{y}_k^{(i)} = x^{(i)} \cdot \hat{y}_l^{(i)} (1 - \hat{y}_l^{(i)})} \text{ ——— (ii)}$$

Hence Proved

if $l \neq k$

$$\nabla_{w^{(l)}} \hat{y}_k^{(i)} = \nabla_{w^{(l)}} \frac{\exp(z_k^{(i)})}{\sum_{k'=1}^{c} \exp z_{k'}^{(i)}}$$

$$= \nabla_{w^{(l)}} \frac{\exp(x^T w_k + b_k)}{\sum_{k'=1}^{c} \exp(x^T w_{k'} + b_{k'})}$$

$$= \frac{0 \cdot \sum_{k'=1}^{c} \exp(x^T w_{k'} + b_{k'}) - x^{(i)} \exp(x^T w_l + b_l) \exp(x^T w_k + b_k)}{\left(\sum_{k'=1}^{c} \exp(x^T w_{k'} + b_{k'}) \right)^2}$$

$$= \frac{0 - x^{(i)} \exp(x^T w_l + b_l) \exp(x^T w_k + b_k)}{\sum_{k'=1}^{c} \exp(x^T w_{k'} + b_{k'}) \sum_{k'=1}^{c} (x^T w_{k'} + b_{k'})}$$

$$\boxed{\nabla_{w^{(l)}} \hat{y}_k^{(i)} = -x^{(i)} \hat{y}_k^{(i)} \hat{y}_l^{(i)}} \text{ ——— (iii)}$$

Hence Proved.

Now substituting eq^n (II) & (III) in eq^n (I)

$$\nabla f_{CE}(W,b) = -\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{\substack{k=l \\ (k=l)}} \frac{y_k^{(i)}}{\hat{y}_k^{(i)}} \nabla_{w(l)}\left(\hat{y}_k^{(i)}\right) + \sum_{k\neq l} \frac{y_k^{(i)}}{\hat{y}_k^{(i)}} \nabla_{w(k)} \hat{y}_k^{(i)}\right]$$

$$= -\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{k=l} \frac{y_l^{(i)}}{\cancel{\hat{y}_l^{(i)}}} x^{(i)} \cancel{\hat{y}_l^{(i)}} \left(1-\hat{y}_l^{(i)}\right) + \sum_{k\neq l} \frac{y_k^{(i)}}{\cancel{\hat{y}_k^{(i)}}}\left(-x^{(i)} \cancel{\hat{y}_k^{(i)}} \hat{y}_l^{(i)}\right)\right]$$

$$= -\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{k=l} y_l^{(i)} x^{(i)} \left(1-\hat{y}_l^{(i)}\right) + \sum_{k\neq l} y_k^{(i)}\left(-x^{(i)}\right) \hat{y}_l^{(i)}\right]$$

$$= -\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{k=l}\left(y_l^{(i)} x^{(i)} - y_l^{(i)} x^{(i)} \hat{y}_l^{(i)}\right) + \sum_{k\neq l}\left(-x^{(i)}\right) y_k^{(i)} \hat{y}_l^{(i)}\right]$$

$$\cancel{= -\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{k=l} x^{(i)} y_l^{(i)} - \sum_{k=1}^{c} x^{(i)} \left(y_l^{(i)} \hat{y}_l^{(i)} + \sum_{k\neq l}\right.\right.}$$

$$= -\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{k=l}\left(x^{(i)} y_l^{(i)}\right) - \sum_{k=l} x^{(i)} y_l^{(i)} \hat{y}_l^{(i)} - \sum_{k\neq l} x^{(i)} \hat{y}_l^{(i)} y_k^{(i)}\right]$$

$$= -\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{k=l} x^{(i)} y_l^{(i)} - \sum_{k=1}^{c}\left(x^{(i)} \hat{y}_l^{(i)} y_k^{(i)}\right)\right]$$

since $\displaystyle\sum_{k=l} x^{(i)} y_l^{(i)} \hat{y}_l^{(i)} + \sum_{k\neq l} x^{(i)} \hat{y}_l^{(i)} y_k^{(i)} = \sum_{k=1}^{c} x^{(i)} \hat{y}_l^{(i)} y_k^{(i)}$

$$= -\frac{1}{n}\sum_{i=1}^{n}\left[x^{(i)} y_l^{(i)} - x^{(i)} \hat{y}_l^{(i)} \sum_{k=1}^{c} y_k^{(i)}\right]$$

and $\sum_{k=1}^{c} y_k^{(i)} = 1$

$$\therefore \nabla_{w} f_{CE}(w,b) = -\frac{1}{n}\sum_{i=1}^{n}\left[x^{(i)} y_k^{(i)} - x^{(i)} \hat{y}_k^{(i)}\right]$$

$$\boxed{\nabla_{w^{(i)}} f_{CE}(w,b) = -\frac{1}{n}\sum_{i=1}^{n} x^{(i)}\left(y_k^{(i)} - \hat{y}_k^{(i)}\right)}$$ Hence Proved

Taking gradient of eq^n (a) wrt b:

$$\nabla_b f_{CE}(w,b) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{c} y_k^{(i)} \nabla_b \log \hat{y}_k^{(i)}$$

$$\nabla_b f_{CE}(w,b) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{c} y_k^{(i)}\left[\frac{\nabla_b \hat{y}_k^{(i)}}{\hat{y}_k^{(i)}}\right] \quad \text{--- (IV)}$$

$$\nabla_b \hat{y}_k^{(i)} = \nabla_b\left[\frac{\exp(x^T w_k + b_k)}{\sum_{k'=1}^{c}\exp(x^T w_{k'} + b_{k'})}\right]$$

if $k = l$:

$$\nabla_b \hat{y}_{k\,(k=l)}^{(i)} = \frac{\exp(x^T w_k + b_k)\sum_{k'=1}^{c}\exp(x^T w_{k'} + b_{k'}) - \exp(x^T w_k + b_k)\exp(x^T w_l + b_l)}{\left[\sum_{k'=1}^{c}\exp(x^T w_{k'} + b_{k'})\right]^2}$$

$$= \frac{\exp(x^T w_k + b_k)}{\sum_{k'=1}^{c}\exp(x^T w_{k'} + b_{k'})} \cdot \left[1 - \frac{\exp(x^T w_l + b_l)}{\sum_{k'=1}^{c}\exp(x^T w_{k'} + b_{k'})}\right]$$

$$\boxed{\nabla_b \hat{y}_{k \,(k=l)}^{(i)} = \hat{y}_l^{(i)}\left(1-\hat{y}_l^{(i)}\right)} \quad \text{—— (V)}$$

if $k \neq l$:

$$\nabla_b \hat{y}_{k\,(k\neq l)}^{(i)} = \nabla_b \left[\frac{\exp\left(x^T w^{(k)} + b^{(k)}\right)}{\sum_{k'=1}^{c} \exp\left(x^T w^{(k')} + b^{(k')}\right)} \right]$$

$$= 0 - \frac{\exp(x^T w_k + b_k)\, \exp\left(x^T w_l + b_l\right)}{\left[\sum_{k'=1}^{c} \exp\left(x^T w_{k'} + b_{k'}\right)\right]^2}$$

$$\boxed{\nabla_b \hat{y}_{k\,(k\neq l)}^{(i)} = -\hat{y}_k^{(i)} \hat{y}_l^{(i)}} \quad \text{—— (VI)}$$

substituting (V) & (VI) in eq^n (IV) :

$$\nabla_b f_{CE}(w,b) = \frac{-1}{n} \sum_{i=1}^{n} \left[\sum_{k=l} \frac{y_k^{(i)}}{\hat{y}_k^{(i)}} \hat{y}_l^{(i)}\left(1-\hat{y}_l^{(i)}\right) + \sum_{k\neq l} \frac{y_k^{(i)}}{\cancel{\hat{y}_k^{(i)}}} \left(-\cancel{\hat{y}_k^{(i)}}\, \hat{y}_l^{(i)}\right) \right]$$

$$= \frac{-1}{n} \sum_{i=1}^{n} \left[\frac{y_l^{(i)}}{\cancel{\hat{y}_l^{(i)}}} \cancel{\hat{y}_l^{(i)}}\left(1-\hat{y}_l^{(i)}\right) + \sum_{k\neq l}\left(-y_k^{(i)} \hat{y}_l^{(i)}\right) \right]$$

$$= \frac{-1}{n} \sum_{i=1}^{n} \left[y_l^{(i)} - y_l^{(i)} \hat{y}_l^{(i)} - \sum_{k\neq l} y_k^{(i)} \hat{y}_l^{(i)} \right]$$

$$= \frac{-1}{n} \sum_{i=1}^{n} \left[y_l^{(i)} - \hat{y}_l^{(i)} \sum_{k=1}^{c} y_k^{(i)} \right]$$

since $\sum_{k=1}^{c} y_k^{(i)} = 1$

$\therefore \quad \nabla_b f_{CE}(W, b) = -\frac{1}{n} \sum_{i=1}^{n} \left[y_k^{(i)} - \hat{y}_k^{(i)} \right]$

$\Rightarrow \boxed{\nabla_b f_{CE}(W, b) = -\frac{1}{n} \sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)} \right)}$ Hence proved.