HOMEWORK - 3 CS541 DEEP LEARNING

Q1. NEWTON'S METHOD.

$$J(w) = \frac{1}{2n} \sum_{i=1}^{n} \left(\hat{y}^{(i)} - y^{(i)}\right)^2$$

~~$\hat{y}^{(i)} = x^{T(i)} w$~~

$$\hat{y}^{(i)} = w^T x^{(i)} = x^{(i)T} w$$

By Newton's method,

$$w^{k+1} = w^k - H f(w^k)^{-1} \nabla_w f(w^{(k)}). \quad \to (1).$$

$$\nabla_w J(w) = \nabla_w \left(\frac{1}{2n} \sum_{i=1}^{n} \left(\hat{y}^{(i)} - y^{(i)}\right)^2 \right)$$

$$= \nabla_w \left(\frac{1}{2n} \sum_{i=1}^{n} \left(w^T x^{(i)} - y^{(i)}\right)^2 \right)$$

$$= 2 \times \frac{1}{2n} \sum_{i=1}^{n} \left(w^T x^{(i)} - y^{(i)}\right) \left(x^{(i)}\right)$$

$$= \frac{1}{n} \sum_{i=1}^{n} x^{(i)} \left(w^T x^{(i)} - y^{(i)}\right) \quad \to (2)$$

Now ~~$H f(w)$~~

$$H\, J(w) = \nabla_w \left(\nabla_w (J(w)) \right)$$

$$= \nabla_w \left(\frac{1}{n} \sum_{i=1}^{n} x^{(i)} \left(\hat{y}^{(i)} - y^{(i)} \right) \right)$$

$$= \nabla_w \left(\frac{1}{n} \sum_{i=1}^{n} x^{(i)} \left(x^{(i)T} w - y^{(i)} \right) \right)$$

$$= \nabla_w \left(\frac{1}{n} \sum_{i=1}^{n} x^{(i)} x^{(i)T} w - x^{(i)} y^{(i)} \right)$$

$$= \frac{1}{n} \sum_{i=1}^{n} x^{(i)} x^{(i)T}$$

$$H = \frac{1}{n} X X^T \rightarrow (3)$$

Eqn. (2) can be written as

$$= \frac{1}{n} \sum_{i=1}^{n} x^{(i)} \left(x^{(i)T} w^k - y^{(i)} \right)$$

$$= \frac{1}{n} X \left(X^T w^k - Y \right). \rightarrow (4)$$

Substituting (3) & (4) in (1) we get.

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

$$= w^k - \frac{\left(X X^T w^k - X Y \right)}{X X^T}$$

$$= \cancel{w^k} - \cancel{w^k} + \frac{X Y}{X X^T}$$

$$\therefore \; w^{k+1} = \frac{XY}{(XX^T)}$$

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

Hence Proved.

(2) Given

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

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

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

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

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

For case $k = \ell$.

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

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

∂ for $f = \frac{a}{b}$.

$$f' = \frac{a'b - b'a}{a^2}.$$

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

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

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

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

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

$$= x^{(i)} \hat{y}_l^{(i)} - \frac{\exp(x^T w^{(l)} + b_l)^{(i)}}{\sum_{k'=1}^{C} \exp(x^T w^{(k')} + b_{k'})^{(i)}} * \frac{\nabla_{w}^{(l)}\left(\sum_{k'=1}^{C} \exp(x^T w^{(k')} + b_{k'})^{(i)}\right)}{\sum_{k'=1}^{C} \exp(x^T w^{(k')} + b_{k'})^{(i)}}$$

$$= x^{(i)} \hat{y}_l^{(i)} - \hat{y}_l^{(i)} * \frac{x^{(i)} \exp(x^T w^{(l)} + b_l)^{(i)}}{\sum_{k'=1}^{C} \exp(x^T w^{(k')} + b_{k'})^{(i)}}$$

$\nabla_{w^{(l)}}\left(\sum_{k'=1}^{C} \exp(x^T w^{(k')} + b_{k'})^{(i)}\right) = 0$ for $\boxed{k' \neq l}$.

$\hat{y}_k^{(i)} = \hat{y}_l^{(i)}$ since $\boxed{l = k}$

$$\therefore \boxed{\nabla_{w^{(l)}} \hat{y}_k^{(i)} = x^{(i)} \hat{y}_l^{(i)} - \hat{y}_k^{(i)} x^{(i)} * \hat{y}_l^{(i)}}$$

$$\therefore \frac{\exp(x^T w^{(l)} + b_l)^{(i)}}{\sum_{k'=1}^{C} \exp(x^T w^{(k')} + b_{k'})^{(i)}} = \hat{y}_k^{(i)}.$$

for $k = l$

$$\therefore \boxed{\nabla_{w^{(l)}} \hat{y}_k^{(i)} = x^{(i)} \hat{y}_l^{(i)} (1 - \hat{y}_k^{(i)})} \rightarrow (1)$$

For $k \neq l$

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

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

Since $k \neq l$ $\nabla_{w^{(l)}} \exp(x^T w^{(k)} + b_k)^{(i)} = 0.$

$\Rightarrow$

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

3

$$= -\frac{\exp(x^T w^{(k)} + b_k)^{(i)}}{\sum_{k'=1}^{C} \exp(\cancel{w^{(k')}} x^T w^{(k')} + b_{k'})^{(i)}} * \frac{\nabla_{w^{(l)}} \left[\sum_{k'=1}^{C} \exp(x^T w^{(k')} + b_{k'})^{(i)} \right]}{\sum_{k'=1}^{C} \exp(x^T w^{(k')} + b_{k'})^{(i)}}$$

$$= -\hat{y}_k^{(i)} * x^{(i)} * \frac{\exp(x^T w^{(l)} + b_l)^{(i)}}{\sum_{k'=1}^{C} \exp(x^T w^{(k')} + b_{k'})^{(i)}}$$

$$= -\hat{y}_k^{(i)} \times x^{(i)} \times \hat{y}_l^{(i)}$$

$$\boxed{\nabla_{w^l} \hat{y}_k^{(i)} = -x^{(i)} \hat{y}_k^{(i)} \hat{y}_l^{(i)}.} \rightarrow 2$$

$\therefore \nabla_{w^{(l)}} \sum_{k'=1}^{C} \exp(x^T w^{k'} + b_{k'})^{(i)} = 0$ for all values $\boxed{k' = l}$

$$\Rightarrow \nabla_{w^{(l)}} \sum_{k'=1}^{C} \exp(x^T w^{(k')} + b_{k'})^{(i)} = x^{(i)} \exp(x^T w^{(l)} + b_l)^{(i)}$$

Total Gradient w.r.t to $w^{(k)}$ = sum over all examples and all classes.

$= \sum_k$

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

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

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

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

from ① & ②

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

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

$$\cancel{= \frac{-1}{n}\sum_{i=1}^{n}\left[y_l^{(i)} x^{(i)} \sum_{k=l}\right.}$$

$$\cancel{\sum_{k=l}}$$

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

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

$$= \cancel{\frac{-1}{n} \sum_{i=1}^{n} \left[y_l^{(i)} x^{(i)} - x \sum_{k=1}^{c} y_k^{(i)} x^{(i)} \right.}$$

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

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

$\sum_{k=1}^{c} y_k^{(i)} = 1$ since ~~tof~~ sum of ~~of~~ probabilities of a class for each example is '1'.

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

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

$$\therefore \boxed{\nabla_{w^{(l)}} f_{CE}(w, b) = \frac{-1}{n} \sum_{i=1}^{n} x^{(i)} \left(y_l^{(i)} - \hat{y}_l^{(i)} \right)}$$

Gradient wrt 'b'

To prove.

$$\nabla_b f_{CE}(w,b) = -\frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)} - \hat{y}^{(i)}\right)$$

For $k = l$.

$$\cancel{\nabla_{b_l} f_{CE}} =$$

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

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

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

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

$$= \hat{y}_l^{(i)} - \hat{y}_l^{(i)} \hat{y}_l^{(i)}$$

$$\boxed{\nabla_{b_l} \hat{y}_k^{(i)} = \hat{y}_l^{(i)} \left(1 - \hat{y}_l^{(i)}\right)} \rightarrow 3$$

$$\nabla_{b_l} \sum_{k'=1}^{c} \exp(x^T w^{(k')} + b_{k'}) = 0 \text{ for all values } k' \neq l$$

$$\hat{y}_k^{(i)} = \hat{y}_l^{(i)}$$

For $k \neq l$.

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

$$= \frac{0 \times \sum_{k'=1}^{c} \exp(x^T w^{(k')} + b_{k'})^{(i)} - \exp(x^T w^{(k)} + b_k)^{(i)} * \nabla_{b_l} \left(\sum_{k'=1}^{c} \exp(x^T w^{(k')} + b_{k'})^{(i)}\right)}{\left[\sum_{k=1}^{c} \exp(x^T w^{(k')} + b_{k'})\right]^2}$$

3

Since $k \neq l$ $\quad \nabla_{b_l}$ wrt $k = 0$

$$= -\frac{\exp(x^T w^{(k)} + b_k)^{(i)}}{\sum_{k'=1}^{c} \exp(x^T w^{(k')} + b_{k'})^{(i)}} \times \frac{\exp(x^T w^{(l)} + b_l)^{(i)}}{\sum_{k'=1}^{c} \exp(x^T w^{(k')} + b_{k'})^{(i)}}$$

$$= -\hat{y}_k^{(i)} \, \hat{y}_l^{(i)}$$

$$\therefore \boxed{\nabla_{b_l} \hat{y}_k^{(i)} = -\hat{y}_k^{(i)} \, \hat{y}_l^{(i)}} \rightarrow 4$$

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

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

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

2

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

From (3) & (4)

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

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

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

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

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

$\sum_{k=1}^{c} y_k^{(i)} = 1$ → sum of probabilities of all class for a training example is '1'

$$\Rightarrow \boxed{\nabla_b f_{CE}(w,b) = \frac{-1}{n}\sum_{i=1}^{n}\left[y_l^{(i)} - \hat{y}_l^{(i)}\right]}$$

$$\cancel{= \frac{-1}{n}\sum_{i=1}^{n}\left[y^{(i)} - \hat{y}\right]}$$

$$\boxed{\nabla_b f_{CE}(w,b) = \frac{-1}{n}\sum_{i=1}^{n}\left[y^{(i)} - \hat{y}^{(i)}\right]}$$

where $\hat{y}_l^{(i)}$ & $y_l^{(i)}$ are represented as vector ~~[illegible]~~ $\hat{y}^{(i)}$ & $y^{(i)}$

(3)

BEST NO. OF EPOCHS = 400

BEST REGULARIZATION CONSTANT = 0.0005

BEST BATCH SIZE = 128

BEST LEARNING RATE = 0.2

TESTING LOSS = 0.54853963

ACCURACY = 81.24 %