1.

i) 2-layer linear neural network : $\hat{y} = f_w(x) = w^T x$

cost function: $J(w) = \frac{1}{2n} \sum_{i=1}^{n} (\hat{y}^{(i)} - y^{(i)})^2$

Newton's Method: $x^* = x^{(0)} - H(f)(x^{(0)})^{-1} \nabla_x f(x^{(0)})$

Replace x with w: $w^* = w^{(0)} - H(f)(w^{(0)})^{-1} \nabla_w f(w^{(0)})$

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

$\nabla_w J(w) = \frac{1}{n} (x)(x^T w - y)$

$\nabla_w^2 J(w) = \frac{1}{n} (x x^T)$ → Hessian

$w^* = w^{(0)} - \left(\frac{1}{n}(x x^T)\right)^{-1} \left(\frac{1}{n}(x)(x^T w^{(0)} - y)\right)$

$= w^{(0)} - (x x^T)^{-1} (x x^T w^{(0)} - x y)$

$= w^{(0)} - w^{(0)} + (x x^T)^{-1} x y$

$= (x x^T)^{-1} x y$

∴ The Newton's Method converges to the optimal solution $w^* = (X X^T)^{-1} X y$ in 1 iteration no matter what the starting point $w^{(0)}$ of the search is.

2.

2) $\hat{y}_k = \frac{\exp z_k}{\sum_{k'=1}^{c} \exp z_{k'}}$, $z_k = x^T w^{(k)} + b_k$, $f_{CE}(W,b) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{c} y_k^{(i)} \log \hat{y}_k^{(i)}$

For $l = k$:

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

$$= \nabla w^{(l)} \left(\frac{\exp(x^T w^{(l)} + b_l)}{\sum_{k'=1}^{c} \exp(x^T w^{(k')} + b_{k'})}\right)^{(i)} \quad * d\left(\frac{f(x)}{g(x)}\right) = \frac{g(x)\,df(x) - f(x)\,dg(x)}{(g(x))^2}$$

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

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

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

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

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

$$= x^{(i)} \hat{y}_l^{(i)} (1 - \hat{y}_l^{(i)})$$

For $l \neq k$

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

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

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

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

log

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

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

(2) Subst. $\nabla_{w^{(L)}} \log \hat{y}_k^{(i)} = \dfrac{\nabla_{w^{(L)}} \hat{y}_k^{(i)}}{\hat{y}_k^{(i)}}$:

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

(3) Subst. $\nabla_{w^{(L)}} \hat{y}_L^{(i)} = x^{(i)} \cdot \hat{y}_L^{(i)}\left(1-\hat{y}_L^{(i)}\right)$ for $k=L$

and

Subst $\nabla_{w^{(L)}} \hat{y}_k^{(i)} = -x^{(i)} \hat{y}_k^{(i)} \hat{y}_L^{(i)}$ for $k\neq L$

Then cancel terms

$$: = -\frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_{(L)}^{(i)}\left(x^{(i)} \hat{y}_L^{(i)}\left(1-\hat{y}_L^{(i)}\right)\right)}{\hat{y}_L^{(i)}} + \sum_{k\neq L}^{c}\left[\frac{y_k^{(i)}\left(-x^{(i)} \hat{y}_k^{(i)} \hat{y}_L^{(i)}\right)}{\hat{y}_k^{(i)}} \right]\right]$$

$$= -\frac{1}{n}\sum_{i=1}^{n}\left[y_{(L)}^{(i)} x^{(i)} - y_{(L)}^{(i)} x^{(i)} \hat{y}_L^{(i)} + \sum_{k\neq L}^{c}\left[y_k^{(i)}\left(-x^{(i)}\right)\hat{y}_L^{(i)} \right]\right]$$

(4) use $\sum_k a_k = a_1 + \sum_{k\neq L} a_k$ to combine $y_L^{(i)} x^{(i)} \hat{y}_L^{(i)}$ with $\sum_{k\neq L}^{c}\left[y_k^{(i)} x^{(i)} \hat{y}_L^{(i)} \right]$

$$: = -\frac{1}{n}\sum_{i=1}^{n}\left[y_{(L)}^{(i)} x^{(i)} - \sum_{k=1}^{c}\left[x^{(i)} \hat{y}_L^{(i)} y_L^{(i)} \right]\right]$$

(5) use $\sum_{k=1}^{c} y_k^{(i)} = 1$:

$$= -\frac{1}{n}\sum_{i=1}^{n}\left[y_{(L)}^{(i)} - x^{(i)} \hat{y}_L^{(i)} \sum_{k=1}^{c} y_{(L)}^{(i)} \right]$$

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

$$\nabla_{w^{(L)}} f_{CE}(W,b) = -\frac{1}{n}\sum_{i=1}^{n} x^{(i)}\left(y_{(L)}^{(i)} - \hat{y}_L^{(i)} \right)$$

Proved

Show that $\nabla_b F_{CE}(W,b) = -\frac{1}{n}\sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})$

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

2. rewrite $\nabla_b \log \hat{y}_k^{(i)}$ as $\left(\frac{\nabla_b \hat{y}_k^{(i)}}{\hat{y}_k^{(i)}}\right)$: $-\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{c} y_k^{(i)} \left[\frac{\nabla_b \hat{y}_k^{(i)}}{\hat{y}_k^{(i)}}\right]$

3. substitute $\hat{y}_k$ for $\frac{\exp(x^T w^{(k)} + b_k)}{\sum_{k'=1}^{c}\exp(x^T w^{(k')} + b_{k'})}$ and simplify $\nabla_b \hat{y}_k$

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

* using $\nabla_b \hat{y}_k^{(i)} = \hat{y}_l^{(i)}(1 - \hat{y}_l^{(i)})$
since $\frac{\partial b}{\partial \hat{y}_k^{(i)}} \cdot \frac{\partial \hat{y}_k^{(i)}}{\partial z_k^{(i)}} \cdot \frac{\partial z_k^{(i)}}{\partial x_k^{(i)}} = \frac{\exp(x+b)}{\exp(x+b)} = 1$

i. if $k=l$: $\nabla_b \hat{y}_{k(k=l)}^{(i)} = \underbrace{\frac{\exp(x^T w^{(l)} + b_l}{\sum_{k'=1}^{c}\exp(x^T w_{k'} + b_{k'})}}_{\hat{y}_l^{(i)}} \cdot \underbrace{\left[\frac{1 - \exp(x^T w_l + b_l)}{\sum_{k'=1}^{c}\exp(x^T w_{k'} + b_{k'})}\right]}_{1 - \hat{y}_k^{(i)}}$

$f(b) = x^T w + b$
$g(y) = \exp(b)$
$h(a) = \frac{1}{b}$

ii. $\nabla_b \hat{y}_{k(k=l)}^{(i)} = \hat{y}_l^{(i)}(1 - \hat{y}_l^{(i)})$

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

* solving with chain/pool rule
$f(b) = f(x^T w + b)$ $f'(b) = 1$ rule
$g(f(b)) = \exp(b)$ $g'(f(b)) = \exp(b)$
$h(g(f(b))) = \frac{1}{b}$ $h'(g(f(b))) = -\frac{1}{b^2}$

$\frac{\partial}{\partial b}(g(F(b)) \cdot h(g(f(b))))$
$= g'(f(b)) \cdot f'(b) \cdot h(g(f(b)))$
$+ g(f(b)) \cdot h'(g(f(b))) \cdot g'(f(b)) \cdot f'(b)$

$\cdot$ $i \neq l = 0 \cdot \frac{\exp(x^T w_l + b_l) \cdot 1}{\exp(x^T w + b)} - \left[\frac{\exp(x^T w_k + b_k) \cdot \exp(x^T w_l + b_l) \cdot 1}{\left(\sum_{k'=1}^{c}(\exp(x^T w + b))\right)^2}\right] \cdot \frac{\partial z_k}{\partial b}$

$V_i = 0 - \frac{\exp(x^T w_k + b_k) \cdot \exp(x^T w_l + b_l)}{1}$

$V_i \cdot \nabla_b \hat{y}_{k(k\neq l)}^{(i)} = -\hat{y}_k^{(i)} \hat{y}_l^{(i)}$

4. Since $\nabla_b \hat{y}_k^{(i)} = \nabla_b \hat{y}_{k(k=l)}^{(i)} + \nabla_b \hat{y}_{k(k\neq l)}^{(i)}$, substitute into eq (2):

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

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

$\therefore \nabla_b F_{CE}(W,b) = -\frac{1}{n}\sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})$

Proved

3.

$$-\log P(D|W,b) = -\log \prod_{i=1}^{n} P(y^{(i)}|x^{(i)},W,b)$$

$$= -\sum_{i=1}^{n} \log P(y^{(i)}|x^{(i)},W,b)$$

$$= -\sum_{i=1}^{n} \log \prod_{k=1}^{c} \hat{y}_k^{(i)\,y_k^{(i)}}$$

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

log power property

$$= -\sum_{i=1}^{n} \sum_{k=1}^{c} y_k^{(i)} \log \hat{y}_k^{(i)}$$

$$\boxed{-\log P(D|W,b) = -\sum_{i=1}^{n} \sum_{k=1}^{c} y_k^{(i)} \log \hat{y}_k^{(i)}}$$

4.

```
Eris-MacBook-Pro:DS541 eri$ python3 homework3_ekim4_ahorn_amevans.py
Optimized hyperparameters:
     Mini-batch size: 200
     Learning rate: 0.01
     # of epochs: 15
     Alpha: 1e-05
     Weights: [[ 0.3028434   0.38584837  0.96888511 ...  0.85339476  0.91509533
   0.86142884]
  [ 0.48925388  0.38130835  0.47666787 ...  0.55585233  0.67352697
   0.82948217]
  [ 0.54769323  0.73651123  0.66853498 ...  0.19282263  0.82017287
   0.18343668]
  ...
  [ 0.25532291  0.39493413  0.88493968 ...  0.70837728 -0.72493666
   0.78279963]
  [ 0.52411611  0.368282    1.46181658 ...  0.15141629  0.47389966
   1.20168119]
  [ 0.45373107  0.87966073  0.65457577 ...  0.51667297  0.72426747
   0.1427419 ]]
          [[ 0.3028434   0.38584837  0.96888511 ...  0.85339476  0.91509533
   0.86142884]
  [ 0.48925388  0.38130835  0.47666787 ...  0.55585233  0.67352697
   0.82948217]
  [ 0.54769323  0.73651123  0.66853498 ...  0.19282263  0.82017287
   0.18343668]
  ...
  [ 0.25532291  0.39493413  0.88493968 ...  0.70837728 -0.72493666
   0.78279963]
  [ 0.52411611  0.368282    1.46181658 ...  0.15141629  0.47389966
   1.20168119]
  [ 0.45373107  0.87966073  0.65457577 ...  0.51667297  0.72426747
   0.1427419 ]]
     Bias:
          [[0.04171745 0.62786595 0.06841105 0.08477843 0.4107784  0.17955993
  0.05757667 0.05119002 0.11661289 0.73491587]]
     min_cost on validation data: 8.718590722681503


---------------TESTING-------------------
Cross Enropy: 8.554876947412696
% of correctly classified images: 81.51%
```