2) Grad. for Log loss (wrt w) $= x(\hat{y} - y)$

If all training examples are +ve, $y = 1$

$$\therefore \quad \nabla_w f_{log} = \underline{x(\sigma(x^T w + b) - 1)} \;;\; \nabla_b f_{log} = \hat{y} - y = (\sigma(x^T w + b) - 1)$$

$$f_{log} = -y \log \hat{y} - (1-y) \log(1-\hat{y})$$

If $y = 1$ $\quad f_{log} = -\log(\sigma(x^T w + b))$

The goal of the neural network is $\therefore$ to minimise $-\log(\sigma(x^T w + b))$

i.e. maximise $\log(\sigma(x^T w + b))$

$\log$ is a monotonically increasing functn, $\therefore$ increasing $\log(\sigma(x^T w + b))$ is the same as increasing $\sigma(x^T w + b)$

Sigmoid is also a monotonically increasing function, $\therefore$ increasing $\sigma(x^T w + b)$ is the same as increasing $x^T w + b$

Consider the way in which b changes

$$b_{new} = b_{old} - \epsilon \nabla_b f_{log} = b_{old} - \epsilon(\hat{y} - y) = b_{old} - \epsilon(\sigma(z) - 1)$$

$$\epsilon \cdot (\sigma(z) - 1) < 0 \quad \because \sigma(z) < 1$$

$\therefore$ b will always keep decreasing and its gradient will never be 0

However the value of $\hat{y} - 1$ will also get smaller as z increases

Therefore, the gradient will in fact get smaller but never reach 0.

d) Let us check how a change in b affects the gradient because this will tell us whether the value of grad. getting smaller affects the rate at which b changes

For simplicity, let $W = 0$

$\therefore \quad \hat{y} - 1 = \sigma(b) - 1$

$$\frac{d}{db}(\hat{y} - 1) = \frac{d}{db}(\sigma(b) - 1) = \sigma(b)(1 - \sigma(b))$$

Suppose b is already very large (as we might expect it to be to correctly output a val close to 1 from the network)

then $\sigma(b) \sim 1$

$\therefore$ the rate of change of the grad wrt b tends to 0

Hence b will have to change by a very large amount even if grad is small

$\therefore$ $\boxed{b \text{ will diverge}}$

b) A similar argument holds for w except that grad wrt w depends on x

hence, W can converge if x is 0

otherwise w will also diverge

(c) Since the value of weights and bias converge, the training loss will also converge.

(d) Considering that the ~~traing e~~ training examples were all positive and testing examples can be positive and negative, it will affect the loss function.

Since the ground truth for testing examples will oscillate or change as 0 or 1, the log loss will change ~~from~~.

It will make the log loss oscillate between

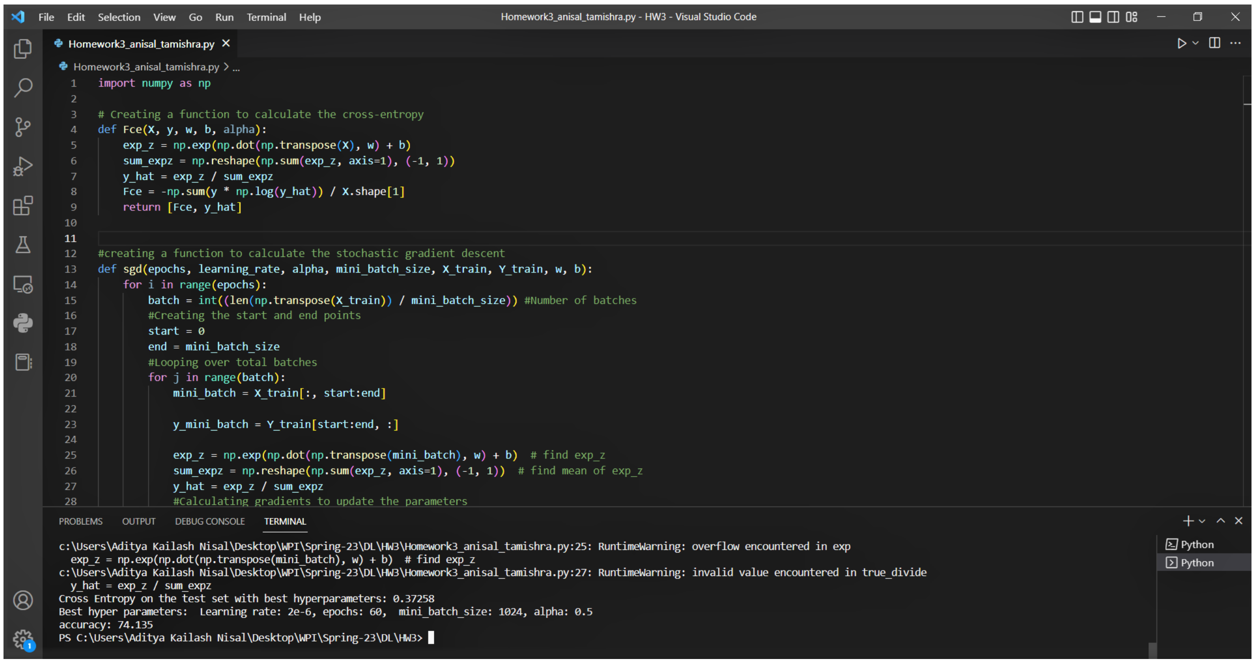
$-\log(\text{yhat})$ and $-(1-y)\log(1-\text{yhat})$.

Thus, the testing loss doesn't converge or won't converge

Team members:
Aditya Nisal
Tanish Mishra

CS-541 Deep Learning
Assignment - 03

Q.4)



Optimal Hyperparameters after testing on the validation set : [1024, 2e-6, 60, 0.5]

minibatch_size = 1024
learning_rate = 2e-6
Epochs = 60
Alpha = 0.5

CS541 HW3

1) Linear Regression on the XOR functn

$$\mathbb{X} = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix} \qquad y^T = [0 \ 1 \ 1 \ 0] \qquad \text{Loss functn: } f_{MSE}$$

$$\cancel{f_{MSE} = \frac{1}{2n}\sum_{i=1}^{n}(y - y^{(i)})^2 = \frac{1}{2n}\sum_{i=1}^{n}(\hat{X}^T\hat{w} - y)^2} \quad \text{Where} \quad \cancel{\hat{X} = \begin{bmatrix} X \\ 1_{1\times n}\end{bmatrix}}$$

$$\cancel{\nabla_w = \frac{1}{n}\hat{X}(\hat{X}^T w - y)} \qquad \cancel{\hat{w} = \begin{bmatrix} w \\ b \end{bmatrix}}$$

Eq 6.1 from the DL Textbook:

$$J(\theta) = \frac{1}{4}\sum_{x\in\mathbb{X}}\left(f^*(x) - f(x;\theta)\right)^2 \quad \text{Where} \quad f^*(x) = \text{XOR}$$

For a linear model $f(x;\theta) = f(x; w, b) = x^T w + b$

$$J(\theta) = \frac{1}{4}\sum\left(f^*(x) - (x^T w + b)\right)^2$$

$$\nabla_{w,b} J(\theta) = 0 = \nabla_{w,b}\sum\left(f^*(x) - (x^T w + b)\right)^2$$

$$= \nabla_{w,b}\left(\left[0 - [0 \ 0]\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} - b\right]^2 + \left[1 - [0 \ 1]\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} - b\right]^2 + \left[1 - [1 \ 0]\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} - b\right]^2 + \left[0 - [1 \ 1]\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} - b\right]^2\right)$$

$$\nabla_{w,b}\left(b^2 + (1 - w_2 - b)^2 + (1 - w_1 - b)^2 + (-w_1 - w_2 - b)^2\right)$$

$$= \nabla_{w,b}\left(g(w_1, w_2, b)\right)$$

$$\frac{\partial(g)}{\partial w_1} = \left(2(w_1 + b - 1) + 2(w_1 + w_2 + b)\right) = 2(2w_1 + 2b - 1 + w_2) = 4\left(w_1 + b - \tfrac{1}{2} + \tfrac{w_2}{2}\right)$$

$$\frac{\partial(g)}{\partial w_2} = \left(2(w_2 + b - 1) + 2(w_1 + w_2 + b)\right) = 4\left(w_2 + b - \tfrac{1}{2} + \tfrac{w_1}{2}\right)$$

$$\frac{\partial(g)}{\partial b} = \left(2b + 2(w_1 + b - 1) + 2(w_2 + b - 1) + 2(w_1 + w_2 + b)\right) = 4(2b + w_1 + w_2 - 1)$$

$$\nabla_{w,b}(q) = 0 \Rightarrow 4\begin{bmatrix} w_1 + b - 1/2 + w_2/2 \\ w_2 + b - 1/2 + w_1/2 \\ w_1 + w_2 + 2b - 1 \end{bmatrix} = 0$$

Equating each element to 0

$$w_1 + b - 1/2 + \frac{w_2}{2} = 0 \Rightarrow 2w_1 + 2b + w_2 - 1 = 0 \quad \cdots \mathrm{I}$$

similarly

$$2w_2 + 2b + w_1 - 1 = 0 \quad \cdots (\mathrm{II})$$

$$w_1 + w_2 + 2b - 1 = 0 \quad \cdots \mathrm{III}$$

$$(\mathrm{I}) - (\mathrm{II}) \Rightarrow \quad w_1 - w_2 = 0 \Rightarrow \boxed{w_1 = w_2 = w}$$

Replacing in (II)

$$3w + 2b - 1 = 0 \quad \cdots \mathrm{IV}$$

and (III)

$$2w + 2b - 1 = 0 \quad \cdots \mathrm{V}$$

$$(\mathrm{IV}) - (\mathrm{V}) \Rightarrow \boxed{w = 0}$$

Replacing in III

$$2b - 1 = 0$$

$$\Rightarrow \boxed{b = \frac{1}{2}}$$

Hence Proved

(Q3) $W = [w^{(1)} \dots w^{(C)}]$ ①

$$\hat{y}_K = \frac{\exp z_K}{\sum_{k'=1}^{C} \exp z_{k'}}$$

$$z_K = x^T W^{(K)} + b_K$$

$$\nabla_{w^{(l)}} f_{CE}(W, b) = -\frac{1}{n} \sum_{i=1}^{n} \sum_{K=1}^{C} y_K^{(i)} \nabla_{w^{(l)}} \log \hat{y}_K^{(i)}.$$

$$= -\frac{1}{n} \sum_{i=1}^{n} \sum_{K=1}^{C} y_K^{(i)} \left(\frac{\nabla \hat{y}_K^{(i)}}{\hat{y}_K^{(i)}} \right).$$

For $l = K$:

$$\nabla_{w^{(l)}} \hat{y}_K^{(i)} = ?$$

since $z_K = x^{(i)T} W^{(K)} + b_K$

$$\hat{y}_K^{(i)} = \frac{\exp z_K}{\sum_{k'=1}^{C} \exp z_{k'}} = \frac{\exp(x^{(i)T} W^{(K)} + b_K)}{\sum_{k'=1}^{C} \exp(x^{(i)T} w^{(k')} + b_{k'})}.$$

since $K = l$.

$$\hat{y}_l^{(i)} = \frac{\exp(x^{(i)T} w^{(l)} + b_l)}{\sum_{k'=1}^{C} \exp(x^{(i)T} w^{(k')} + b_{k'})}. \Rightarrow \frac{\exp(x^{(i)T} W^{(l)} + b_l)}{\text{den}}.$$

$$\nabla_{w^{(l)}} \hat{y}_l^{(i)} = \frac{\text{den} \cdot \frac{d}{dw^l}\left(\exp(x^{(i)T} W^{(l)} + b_l)\right) - \exp(x^{(i)T} w^{(l)} + b_l) \cdot \frac{d}{dw_l} \text{den}}{(\text{den})^2}$$

$$= \frac{x^{(i)}\left(den \cdot \exp(x^{(i)T}w^{l}+b_l) - \exp(x^{(i)T}w^{l}+b_l)(\cdot \exp(x^{(i)T}w^{l}+b_l))\right)}{den^2}$$

$=$ let $\exp(x^{(i)T}w^{l}+b_l) = \underline{num}$

$$= \frac{x^{(i)}(\cdot den \cdot num - num \cdot num)}{den^2}$$

$$= x^{(i)}\left(\frac{num}{den} - \frac{num^2}{den^2}\right)$$

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

$$= x^{(i)}\left(\hat{y}_l^{(i)} - \hat{y}_l^{(i)2}\right)$$

$$\boxed{= x^{(i)}\hat{y}_l^{(i)}\left(1-\hat{y}_l^{(i)}\right)} \quad — (1)$$

for $l \neq k$: $y_k^{(i)} = \frac{\exp z_k}{den} = \frac{\exp(x^{(i)T}W^{(k)}+b_k)}{den}$.

$$\nabla_{w^{(l)}} y_k^{(i)} = \frac{den \times \cancel{\frac{d}{dw^{(l)}}(x^{(i)T}W^{(k)}+b_k)}^{\,0} - (\exp x^{(i)T}W^{(k)}+b_k^{*})\frac{d}{dw^{(l)}}den}{den^2}$$

$$= \frac{\cancel{den}\ 0 - x^{(i)}(\exp x^{(i)T}W^{(k)}+b_k^{*})(\exp(x^{(i)T}W^{(l)}+b_l)}{den^2}$$

$$\rightarrow \boxed{\nabla_{w^{(\ell)}} \hat{y}^{(i)}_{k} = -x^{(i)} \hat{y}_k^{(i)} \cdot \hat{y}_\ell^{(i)}} \text{ — } (2)$$

Now, to find:

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

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

$$= \frac{1}{n} \sum_{i=1}^{n} \left(y_1^{(i)} \cdot \left(\frac{\nabla_{w^\ell} \hat{y}_1^{(i)}}{\hat{y}_1^{(i)}} \right) + y_2^{(i)} \left(\frac{\nabla_{w^\ell} \hat{y}_2^{(i)}}{\hat{y}_2^{(i)}} \right) \cdots\cdots + y_c^{(i)} \cdot \left(\frac{\nabla_{w^\ell} \hat{y}_c^{(i)}}{\hat{y}_c^{(i)}} \right) \right)$$

now, using (1) and (2) we get:

$$= \frac{1}{n} \sum_{i=1}^{n} \left[\left(y_1^{(i)} \frac{(-x \cancel{\hat{y}_1^{(i)}} \cdot \hat{y}_\ell^{(i)})}{\cancel{\hat{y}_1}} \right) + \left(y_2^{(i)} \frac{(-x \cancel{\hat{y}_2^{(i)}} \cdot \hat{y}_\ell^{(i)})}{\cancel{\hat{y}_2}} \right) \cdots + \left(y_\ell^{(i)} \frac{(x \cancel{\hat{y}_\ell^{(i)}} - x \hat{y}_\ell^{(i)2})}{\cancel{\hat{y}_\ell^{(i)}}} \right) + \cdots \right.$$

$$\left. \cdots y_c^{(i)} \left(\frac{-x \cancel{\hat{y}_c^{(i)}} \hat{y}_\ell^{(i)}}{\cancel{\hat{y}_c^{(i)}}} \right) \right]$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left(-x^{(i)} y_1^{(i)} \hat{y}_\ell - x^{(i)} y_2^{(i)} \hat{y}_\ell \cdots + x^{(i)} y_\ell^{(i)} - x^{(i)} \hat{y}_\ell^{(i)} y_\ell^{(i)} \cdots - x^{(i)} y_c^{(i)} \hat{y}_\ell^{(i)} \right)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left[\left(x^{(i)} \hat{y}_\ell^{(i)} \left(-y_1 - y_2^{(i)} \cdots - y_c^{(i)} + \cdots \right) \right) + x^{(i)} y_\ell^{(i)} \right] \quad \because \sum_{k=1}^{c} y_k^{(i)} = 1$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left(x^{(i)} \hat{y}_\ell^{(i)} (-1) + x^{(i)} y_\ell^{(i)} \right).$$

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

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

Similarly, to compute the gradient of f_{CE} wrt b_K, we follow the same procedure.

Since, $x^{(i)}$ was a coefficient of $W_K^{(i)}$ so we had it ~~outside~~ multiplied with the derivative.

Now, $b_K^{(i)}$ here doesn't have any coeff but 1.

Hence,

for $k = l$: $\nabla_{b_K} f_{CE} = \hat{y}_l^{(i)}(1 - \hat{y}_l^{(i)})$.

$k \neq l$: $\nabla_{b_K} f_{CE} = -\hat{y}_K^{(i)} \cdot \hat{y}_l^{(i)}$.

Thus we get:

$$\nabla_{b_K} f_{CE} = -\frac{1}{n}\sum_{i=1}^{n} [y_l^{(i)} - \hat{y}_l^{(i)}]$$

Combining all these scalars into one vector, we get:

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