10/9/22.

Homework (3)

Page No.:
Date:

Q1. (1) * We know that,

$$w_n = w_{n-1} - \alpha \left.\frac{dL}{dw}\right|_{w_{n-1}} \longrightarrow (1)$$

where w_n, w_{n-1} are consecutive weights & L is the training loss. with $\alpha \rightarrow$ learning rate.

* A very high learning rate causes large error gradients to accumulate which results in very large updates to the weights during training. This is called the exploding gradient problem.

* This causes model training to diverge & the loss increases rapidly as a result.

(2.) * Learning rate scheduler updates the learning rate during the training process, after a certain number of epochs, according to some rules.

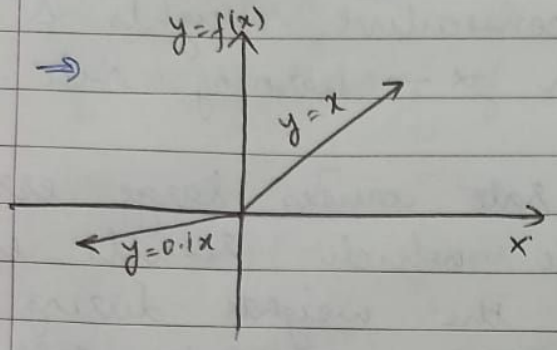
* In the figure, the learning rate decays [decreases] after ~15 epochs each for some part of the training process.

* Such a schedule with decreasing learning rate after some epochs during the training process can lead to improved accuracy, faster convergence, and reduce overfitting of the model.
e.g., Reducing (a very high) learning rate uniformly during the training will cause model to converge faster towards the beginning, and moving reasonably slow in later epochs.

* Also, changing learning rate makes the model training independent of it.

Q2. ① $f(x) = \max(x, 0) + \min(x, 0) * 0.1$

ans: $f(x) = \begin{cases} x, & x \geq 0 \\ 0.1x, & x < 0 \end{cases} \longrightarrow ①$



$\Rightarrow$ ∴ Since $f(x)$ is non-linear, it may be a suitable activation ftⁿ.

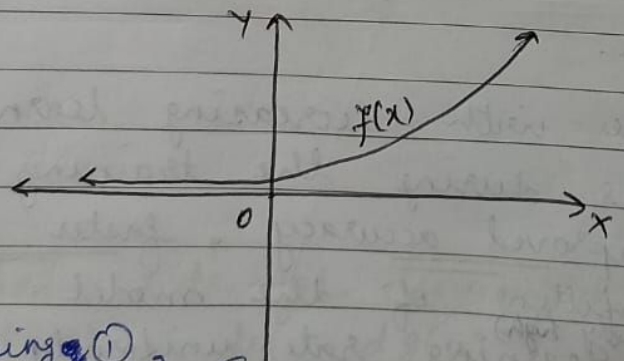
$f'(x) = \begin{cases} 1, & x \geq 0 \\ 0.1, & x < 0 \end{cases} \longrightarrow ②$

∴ Gradient of $f(x)$ is never going to be 0.

② $f(x) = \ln(e^{3x} + 1)$

ans: $f'(x) = \frac{e^{3x}}{e^{3x}+1} \cdot 3 = \frac{3e^{x}}{e^{3x}+1} \longrightarrow ①$

$= \frac{3}{1+e^{-3x}}$



∴ $f(x)$ is a non-linear activation ftⁿ.

Using ①,

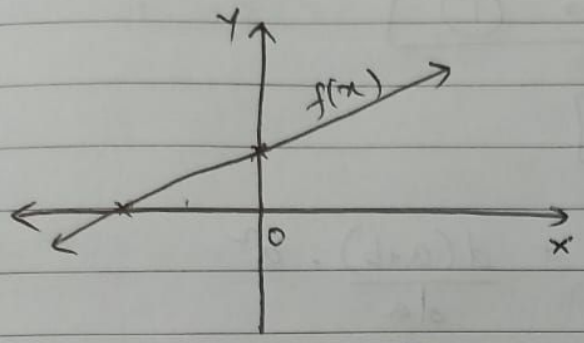
$f'(x) = \begin{cases} 3, & x \to \infty \\ 0, & x \to -\infty \end{cases}$

$\Rightarrow$ $f(x)$ is prone to vanishing gradient for high negative values of x.

Page No.:
Date:

(3.) $f(x) = \ln(e^{3x+1})$

ans: $f(x) = (3x+1) \longrightarrow (1)$



$\therefore$ $f(x)$ is a linear activation ft^{n}.

$f'(x) = 3 \longrightarrow (2)$

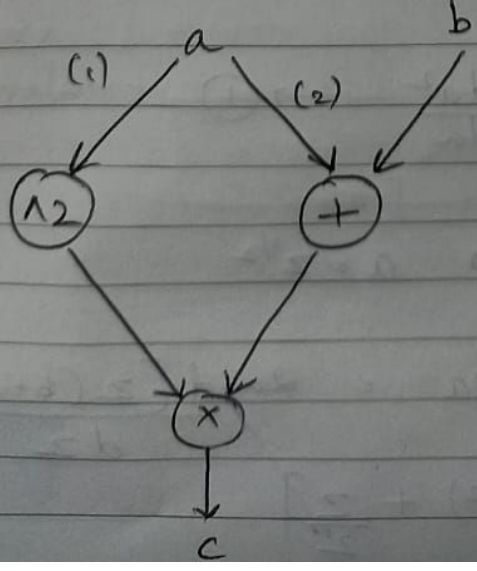
$\Rightarrow f'(x)$ is not prone to gradient vanishing.

$\therefore$ Combining (1), (2), (3) parts,

(i) $f(x) = \ln(e^{3x+1})$ is not a suitable activation ft^{n} since its linear & won't be able to fit a variety of data.

(ii) $f(x) = \ln(e^{3x}+1)$ is prone to gradient vanishing.

Q3.



$a = -1$, $b = 4$

(1.) $dc/da = ?$

ans: $c = (a^2) \times (a+b)$

$$\boxed{c = a^2 \cdot (a+b)} \longrightarrow (1)$$

$$\frac{dc}{da} = \frac{d[a^2 \cdot (a+b)]}{da}$$

$$= \left(\frac{d(a^2)}{da}\right)(a+b) + \frac{d(a+b) \cdot a^2}{da}$$

$$= 2a(a+b) + a^2 \longrightarrow (2)$$

* Putting values in eq. (2),

$$\left(\frac{dc}{da}\right) = 2(-1)[-1+4] + (-1)^2$$

$$= -6+1 = \underline{\underline{-5}}.$$

(2.) $\left(\frac{dc}{da}\right)$ component at (1) & (2) ?

ans: $$\left(\frac{dc}{da}\right) = \left[\frac{\partial c}{\partial v_1} \cdot \frac{\partial v_1}{\partial a}\right] + \left[\frac{\partial c}{\partial v_2} \cdot \frac{\partial v_2}{\partial a}\right] \longrightarrow (1)$$

where $v_1 = a^2$ & $v_2 = (a+b)$

& $c = v_1 + v_2$

(v_1 = location (1)
v_2 = location (2))

At (1) :→

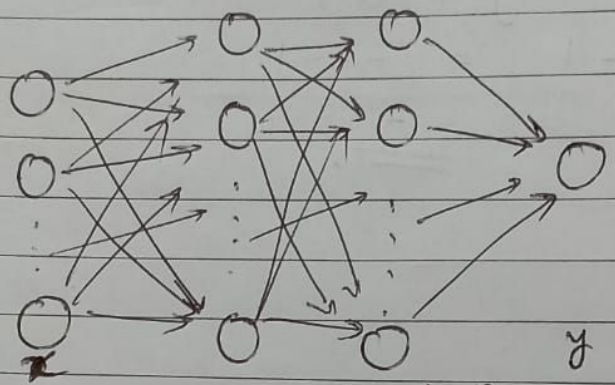
$$\frac{\partial c}{\partial v_1} \cdot \frac{\partial v_1}{\partial a} = (a+b) \cdot (1) \cdot (2a) = (-1+4) \cdot 2(-1)$$

$$= \underline{\underline{-6}}.$$

At (2) :→

$$\frac{\partial c}{\partial v_2} \cdot \frac{\partial v_2}{\partial a} = a^2 \cdot (1) \cdot (1) = (-1)^2 = \underline{\underline{1}}.$$

Q4. Binary Classification Problem



$x \in R^{100 \times 1}$

$y \in R^{1 \times 1}$

(1.) H.L-1 = 100 units

H.L-2 = 20 units

No. of parameters = $(100 \times 100) + (100 \times 20) + (20 \times 1)$

$= 10000 + 2000 + 20$

$= 12020$.

(2.) H.L-1 = 20 units

H.L-2 = 100 units

No. of parameters = $(100 \times 20) + (20 \times 100) + (100 \times 1)$

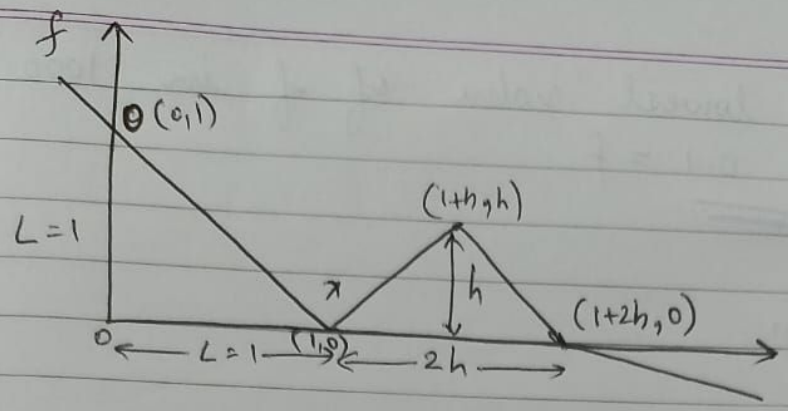
$= 2000 + 2000 + 100$

$= 4100$.

P.T.O.

Page No.:
Date:

Q.5.



(1.) Lowest value of f in 1000 steps ?

ans. $$f(x) = \begin{cases} (1-x) & , x \leq 1 \\ (x-1) & , 1 < x \leq (h+1) \\ (1-x) & , (h+1) < x \leq (1+2h) \\ \beta(1-x) & , (1+2h) < x \quad \forall \beta > 0 \end{cases}$$

$$f'(x) = \begin{cases} -1 & , x \leq 1 \\ 1 & , 1 < x \leq (h+1) \\ -1 & , (h+1) < x \leq (1+2h) \\ -\beta & , x > (1+2h) \end{cases}$$

* At point 0, $x_0 = 0$

$\alpha = 0.3$

$\Rightarrow x_1 = x_0 - \alpha f'(x_0) = 0 - 0.3(-1) = 0.3$

$x_2 = 0.3 - 0.3(-1) = 0.6$

$x_3 = 0.6 - 0.3(-1) = 0.9$

$x_4 = 0.9 - 0.3(-1) = 1.2$

$x_5 = 1.2 - 0.3(1) = 0.9$

$x_6 = 0.9 - (0.3)(-1) = 1.2$

$\Rightarrow$ We see that after $x_3 = 0.9$, x starts to oscillate b/w 0.9 & 1.2.

↳ We are oscillating around a local minima.

$\Rightarrow y_3 = (1-0.9) = 0.1$ & $y_4 = (1.2-1) = 0.2$

Date:

∴ The lowest value of f in 1000 steps is 0.1 = f

(2) $\alpha = 0.001$

$\beta_1 = 0.9$

$\beta_2 = 0.999$

$\epsilon = 0$

ans ☆ From the graph, we know that Adam optimizer will escape the local minima at x if

$$x > (1+h) \longrightarrow (1)$$

$$\Rightarrow (x-1) > h$$

$$\Rightarrow h_{max} = (x_{max} - 1) \longrightarrow (2)$$

☆ ∴ We need to find the max. value of x such that it doesn't start descending towards local minima again.

NOTE: Please find attached code for finding h_{max}.

$$h_{max} = 0.002434$$