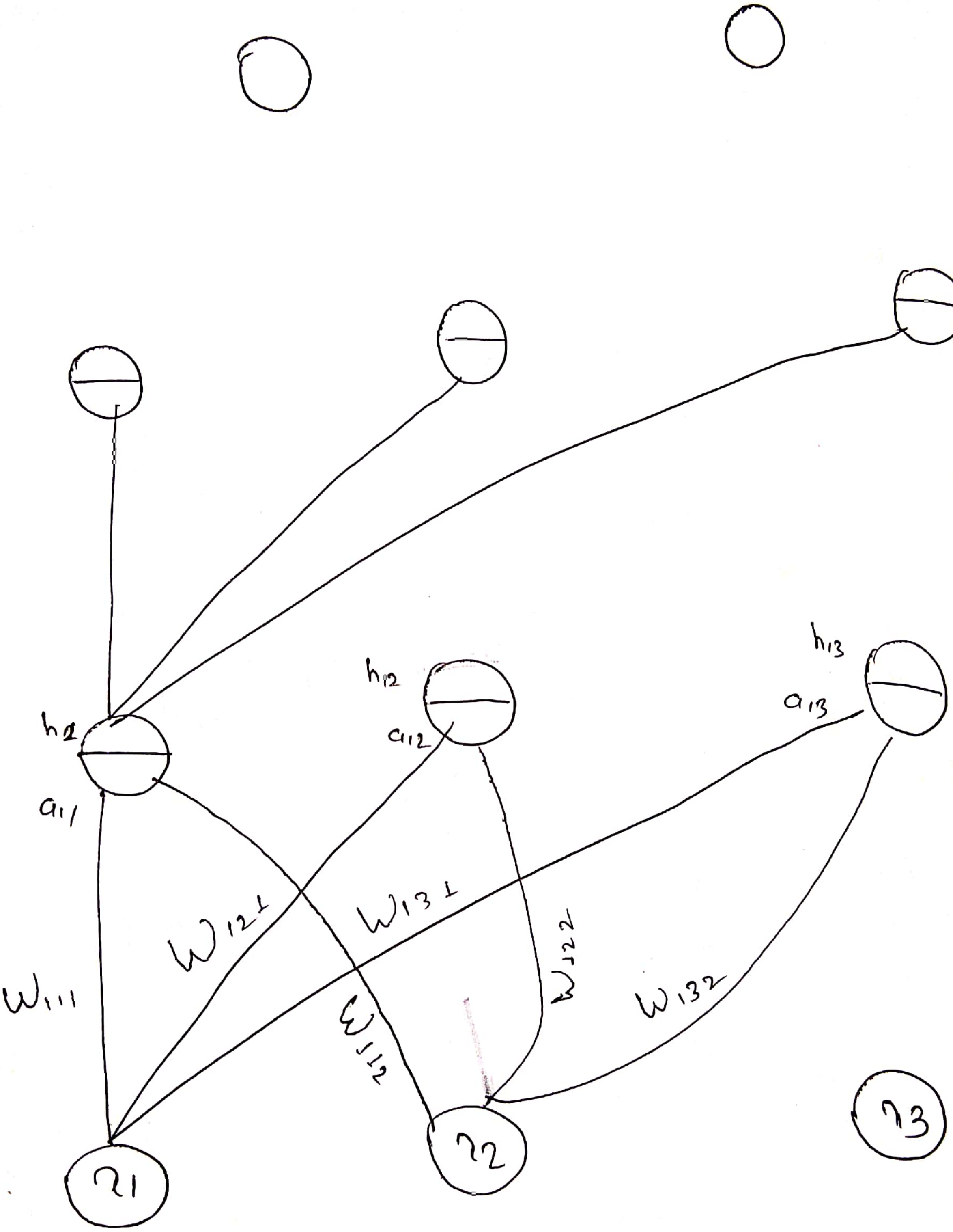


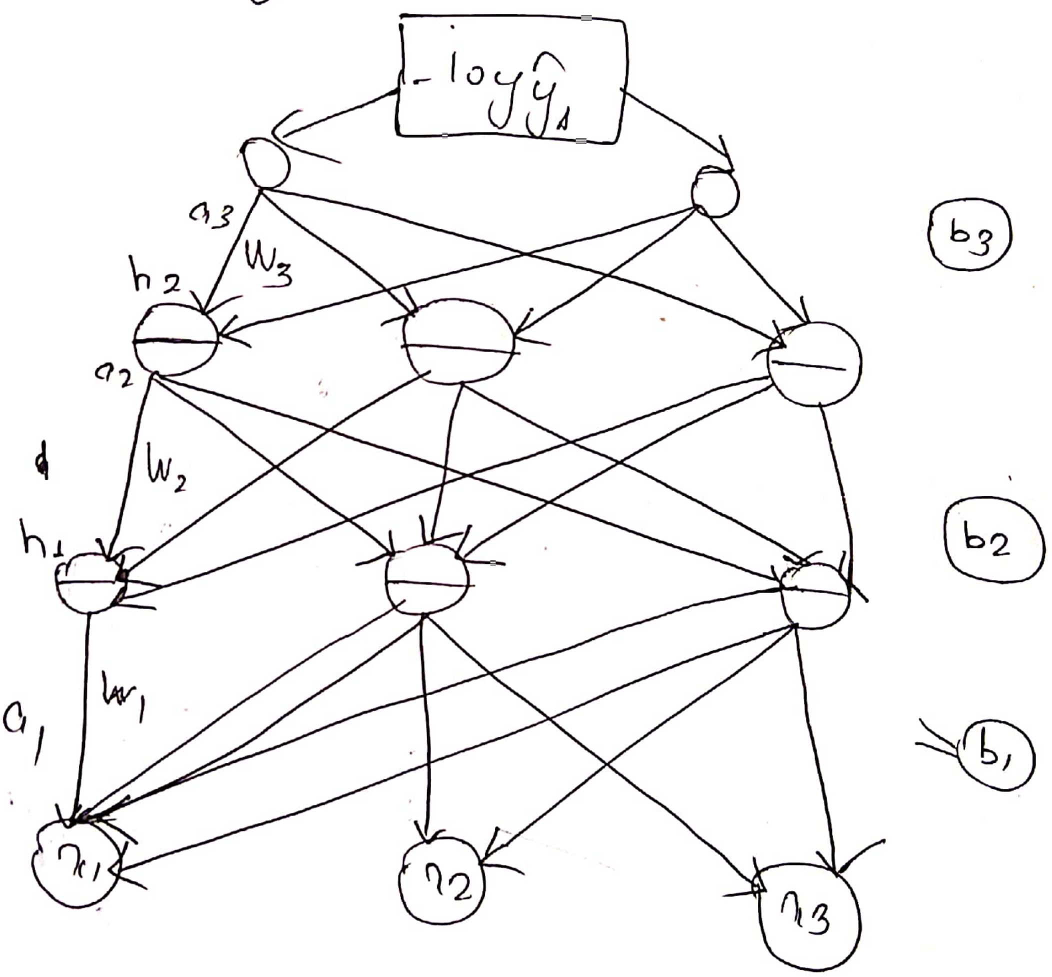
k = Layer
i = current layer neuron no.
j = previous / Input layer neuron no.

$f(x) = x^2$

$\frac{df(x)}{dx}$ $\quad f(x) = x^2 + y^2$

$\frac{\partial}{\partial x}, \frac{\partial}{\partial y}$ } partial derivative

$$\nabla_\theta = \left[\frac{\partial}{\partial x} \quad \frac{\partial}{\partial y} \right]$$



$$\frac{\partial L(\theta)}{\partial W_{111}} = \underbrace{\frac{\partial L(\theta)}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_3}}_{\text{deriv w.r. to output layer}} \underbrace{\frac{\partial a_3}{\partial h_2} \frac{\partial h_2}{\partial a_2} \frac{\partial a_2}{\partial h_1} \frac{\partial h_1}{\partial a_1}}_{\text{der. w.r.t to hidden layer}} \underbrace{\frac{\partial a_1}{\partial W_{111}}}_{\text{wts}}$$

<u>First</u>

we have, $L(\theta) = -y_l \log \hat{y}_l$

Then,

$$\frac{\partial L(\theta)}{\partial a_{Li}} = \frac{\partial(-\log \hat{y}_l)}{\partial a_{Li}}$$

$$= \frac{\partial(-\log \hat{y}_l)}{\partial \hat{y}_l} \times \frac{d\hat{y}_l}{\partial a_{Li}} \quad -(1)$$

Where, L = layer no.
i = neuron no. (1 to k)
l = index of correct output.

The first part of derivative in eq(1) is straight forward $\frac{\partial(-\log \hat{y}_l)}{\partial \hat{y}_l} = -\frac{1}{\hat{y}_l}$

So,

$$\frac{\partial L(\theta)}{\partial a_{Li}} = \frac{\partial(-\log \hat{y}_l)}{\partial \hat{y}_l} \times \frac{\partial \hat{y}_l}{\partial a_{Li}}$$

$$= -\frac{1}{\hat{y}_l} \frac{\partial}{\partial a_{Li}} \text{softmax}(a_L)_l$$

$$= -\frac{1}{\hat{y}_l} \frac{\partial}{\partial a_{Li}} \frac{\exp(a_L)_l}{\sum_{i'} \exp(a_L)_{i'}}$$

(3)

$$\frac{\partial\left(\frac{g(x)}{h(x)}\right)}{\partial x} = \frac{\partial g(x)}{\partial x}\frac{1}{h(x)} - \frac{g(x)}{h(x)^2}\frac{\partial h(x)}{\partial x}$$

Using division rule,

$g(x) = \exp(a_L)_l$

$$= \frac{-1}{\hat{y}_l}\left(\frac{\frac{\partial}{\partial a_{Li}}\exp(a_L)_l}{\sum_{i'}\exp(a_L)_{i'}} - \frac{\exp(a_L)_l\left(\frac{\partial}{\partial a_{Li}}\sum_{i'}\exp(a_L)_{i'}\right)}{\sum_{i'}\left(\exp(a_L)_{i'}\right)^2}\right)$$

Conside,

$$\left(\frac{\partial}{\partial a_{Li}}\exp(a_L)_l,\right)$$

This value is 0 for all i: 0 to k except for l so, an indicator can be used $1_{(l=i)}$ $\exp(a_L)_i$ to denote all the values except i = l, resolve to 0

Now, it is simply derivative of exponent.

$$\frac{\partial L(\theta)}{\partial a_{Li}} = \frac{-1}{\hat{y}_i}\left(\frac{1_{(l=i)}\exp(a_L)_l}{\sum\exp(a_L)_{i'}} - \frac{\exp(a_L)_l}{\sum_{i'}\exp(a_L)_{i'}}\frac{\exp(a_L)_i}{\sum_{i'}\exp(a_L)_{i'}}\right)$$

writing interms of softmax

$$\frac{\partial L(\theta)}{\partial a_{Li}} = \frac{-1}{\hat{y}_i}\left(1_{(l=i)}\text{softmax}(a_L)_i - \text{softmax}(a_L)_i\,\text{softmax}(a_L)_i\right)$$

softmax is $\hat{y}$

$$\frac{\partial L(\theta)}{\partial a_{Li}} = \frac{-1}{\hat{y}_i}\left(1_{(l=i)}\hat{y}_l - \hat{y}_l\hat{y}_i\right)$$

after cancellation,

$$\frac{\partial L(\theta)}{\partial a_{Li}} = -\left(1_{(l=i)} - \hat{y}_i\right)$$

3 So far we have,

$$\frac{\partial L(\theta)}{\partial a_{Li}} = -(1_{(l=i)} - \hat{y}_i)$$

Now, gradient w.r. to vector a_L.

$a_L = [a_{L1}, a_{L2} \dots a_{LK}]$

by indictor variable, it resolves to 0, for all values of i except $i=l$.

let $k=4$, $l=2$

$$\frac{\partial L(\theta)}{\partial a_{L1}} = -(0 - \hat{y}_i)$$

$$\frac{\partial L(\theta)}{\partial a_{L2}} = -(1 - \hat{y}_i)$$

$$\frac{\partial L(\theta)}{\partial a_{L3}} = -(0 - \hat{y}_i)$$

$$\frac{\partial L(\theta)}{\partial L_4} = -(0 - \hat{y}_i)$$

The gradient w.r. to a_L is ∇a_L =

$$\nabla a_L = \begin{bmatrix} \frac{\partial L(\theta)}{\partial a_{L1}} \\ \vdots \\ \frac{\partial L(\theta)}{\partial a_{LK}} \end{bmatrix}$$

$$\nabla a_L = \begin{bmatrix} -(1_{(l=1)} - \hat{y}_i) \\ \vdots \\ -(1_{(l=k)} - \hat{y}_i) \end{bmatrix}$$

It is simply the diff. bet" $[0 \ 1 \ 0 \ 0 \dots 0_k]$ and $\hat{y}$

In reality it is the diff bet" true dis(y) and pred dis $\hat{y}$

$$\nabla_{a_L} L(\theta) = -(y - \hat{y}_i)$$

$$\frac{\partial L(\theta)}{\partial W_{111}} = \underbrace{\frac{\partial L(\theta)}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial a_3}}_{-(y-\hat{y}_i)} \quad \frac{\partial a_3}{\partial h_2} \times \frac{\partial h_2}{\partial a_2} \quad \frac{\partial a_2}{\partial h_1} \times \frac{\partial h_1}{\partial a_1} \quad \frac{\partial a_3}{\partial w_{111}}$$

(5)

So, Now,

- deri. w.r. to hidden layers;

$$\frac{\partial L(\theta)}{\partial h_{ij}} = \sum_{m=1}^{k} \underbrace{\frac{\partial L(\theta)}{\partial a_{i+1,m}}} \frac{\partial a_{i+1,m}}{\partial h_{ij}}$$

already computed earlier.

$$\sum_{m=1}^{k} \frac{\partial L(\theta)}{\partial a_{i+1}} W_{i+1,m,j}$$

How $W_{i+1,m,j}$ is coming?

Supp. $\frac{\partial a_{31}}{\partial h_{22}}$

$$\frac{\partial (W_{311} h_{21} + W_{312} h_{22} + W_{313} h_{23})}{\partial h_{22}}$$

$$= W_{312}$$

- $$\sum_{m=1}^{k} \frac{\partial L(\theta)}{\partial a_{i+1}} W_{i+1,m,j}$$

$$\nabla_{a_{i+1}} L(\theta) = \begin{bmatrix} \frac{\partial L(\theta)}{\partial a_{i+1,1}} \\ \vdots \\ \frac{\partial L(\theta)}{\partial a_{i+1,k}} \end{bmatrix} \quad W_{i+1,\cdot j} \begin{bmatrix} W_{i+1,1,j} \\ \vdots \\ W_{i+1,k,j} \end{bmatrix}$$

↓ grad. of Loss func. w.r.t all output neurons from $a_{i+1,1}$ to $a_{i+1,k}$.

↓ It refers all rows of jth column.

dot product of $(W_{i+1,\cdot j})^T \nabla_{a_{i+1}} L(\theta) = \sum_{m=1}^{k} \frac{\partial L(\theta)}{\partial a_{i+1,m}} W_{i+1,m,j}$

6

Here,
derivative of Loss function w.r.t hidden Layer is only the dot product b/t" gradient of Loss w.r.t Output layer and corresponding wts.

so,

$$\frac{\partial L(\theta)}{\partial h_{ij}} = (W_{i+1,\cdot,j})^T \nabla_{a_{i+1}} L(\theta)$$

$$\nabla_{h_i} L(\theta) \quad \begin{bmatrix} (W_{i+1,\cdot,1})^T \nabla_{a_{i+1}} L(\theta) \\ \vdots \\ (W_{i+1,\cdot,n})^T \nabla_{a_{i+1}} L(\theta) \end{bmatrix}$$

also, $(W_{i+1})^T \nabla_{a_{i+1}} L(\theta)$

It is special case for Last hidden Layers.
Now, lets mak it more generic for all hidden Layer.

1. consider for next layer a_i

$$\frac{\partial L(\theta)}{\partial a_{ij}} = \underbrace{\frac{\partial L(\theta)}{\partial h_{ij}}}_{\text{already computed}} \frac{\partial h_{ij}}{\partial a_{ij}}$$

{ h_{ij} is application of activation

so, $\frac{\partial L(\theta)}{\partial a_{ij}} = \frac{\partial L(\theta)}{\partial h_{ij}} \underbrace{g'(a_{ij})}_{\text{activation}}$

$$\nabla_{a_i} L(\theta) \quad \begin{bmatrix} \frac{\partial L(\theta)}{\partial h_{i1}} g'(a_{i1}) \\ \vdots \\ \frac{\partial L(\theta)}{\partial h_{in}} g'(a_{in}) \end{bmatrix}$$

$$\nabla_{a_i} L(\theta) = \nabla_{h_i} L(\theta) \odot [\ldots g'(a_{ik}) \ldots]$$

element wise vector Mult.

Till now

$$\frac{\partial L(\theta)}{\partial w_{111}} = \frac{\partial L(\theta)}{\partial \hat{y}\, \partial a_3} \frac{\partial a_3}{\partial h_2} \frac{\partial h}{\partial a_2} \frac{\partial a_2}{\partial h_1} \frac{\partial h_1}{\partial a} \frac{\partial a_3}{\partial w_{111}}$$

Uptil now,

$$\nabla h_i\, L(\theta) = (W_{i+1})^T \nabla a_{i+1}\, L(\theta)$$

$$\nabla a_i\, L(\theta) = \nabla h_i\, L(\theta) \odot [\ldots g'(a_{ik}) \ldots]$$

Now, deri of loss wrt wt and bias

$$a_k = b_k + W_k h_{k-1}$$

$$\frac{\partial a_{ki}}{\partial W_{kij}} = h_{k-1,j}$$

- k: layer
- i = current layer neron no
- j = prev/input layer neuron no.

$$\frac{\partial L(\theta)}{\partial W_{kij}} = \frac{\partial L(\theta)}{\partial a_{ki}} h_{k-1,j}$$

$$W_{kij} = W_{kij} - \eta \frac{\partial L(\theta)}{\partial W_{kij}}$$

$$\nabla_{W_k} L(\theta) \begin{bmatrix} \frac{\partial L(\theta)}{\partial W_{k11}} & \frac{\partial L(\theta)}{\partial W_{k12}} & \frac{\partial L(\theta)}{\partial W_{k13}} \\ \frac{\partial L(\theta)}{\partial W_{k21}} & \frac{\partial L(\theta)}{\partial W_{k22}} & \frac{\partial L(\theta)}{\partial W_{k23}} \\ \frac{\partial L(\theta)}{\partial W_{k31}} & \frac{\partial L(\theta)}{\partial W_{k32}} & \frac{\partial L(\theta)}{\partial W_{k33}} \end{bmatrix}$$

$$\nabla W_{kij}\, L(\theta) = \frac{\partial L(\theta)}{\partial a_{ki}}$$

$$\nabla_{W_k} L(\theta) \begin{bmatrix} \frac{\partial L(\theta)}{\partial a_{k1}} h_{k-} & \frac{\partial L(\theta)}{\partial a_{k1}} h_{k-1,2} & \frac{\partial L(\theta)}{\partial a_{k1}} h_{k-1,3} \\ \frac{\partial L(\theta)}{\partial a_{k2}} \cdot h_{k-} & \frac{\partial L(\theta)}{\partial a_{k2}} h_{k-1,2} & \frac{\partial L(\theta)}{\partial k_2} h_{k-1,3} \\ \frac{\partial L(\theta)}{\partial k_3} h_{k-} & \frac{\partial L(\theta)}{\partial a_{k3}} h_{k-1,2} & \frac{\partial L(\theta)}{\partial a_{k3}} k_{-1,3} \end{bmatrix}$$