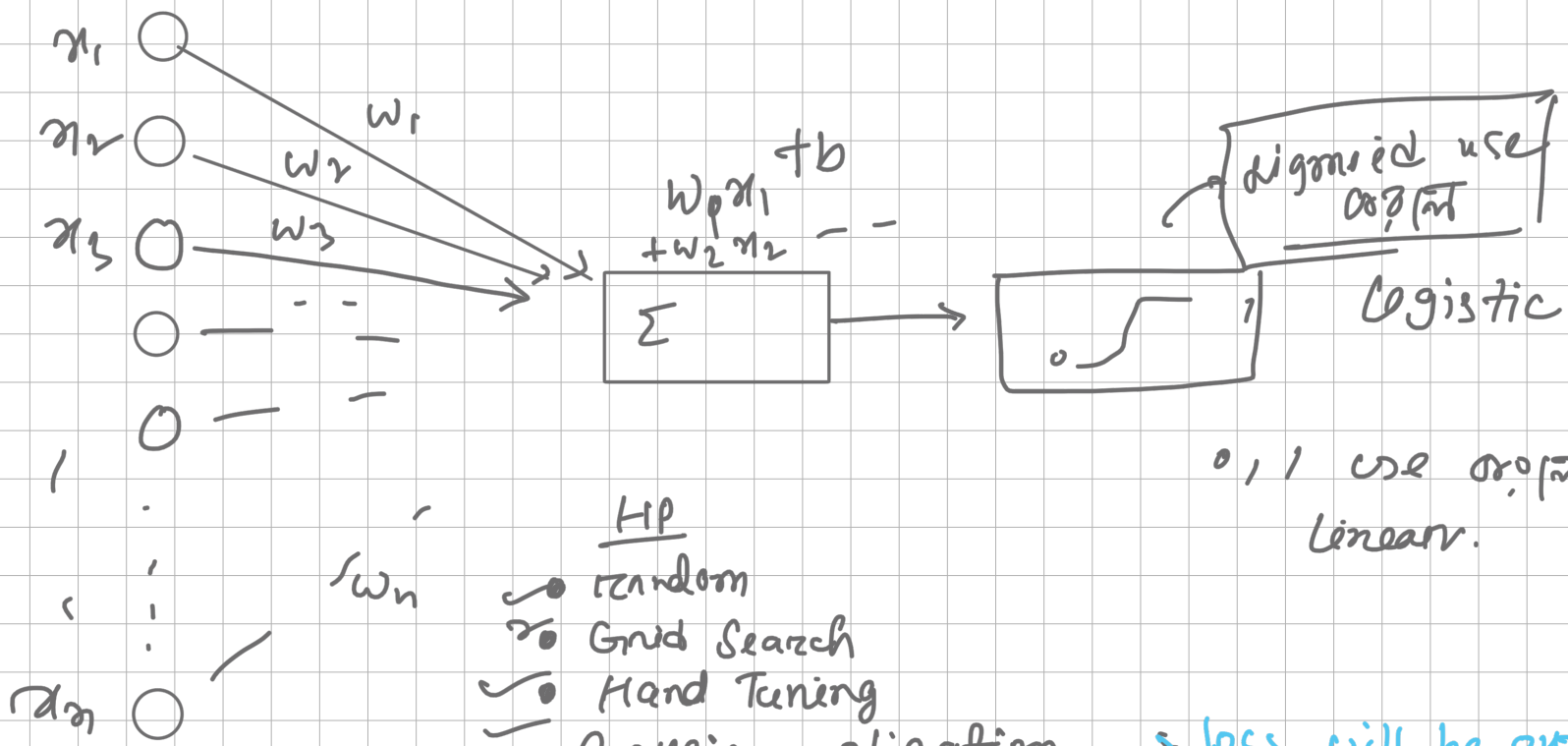
Machine learning



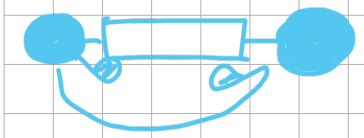
0, 1 use output linear.

HP

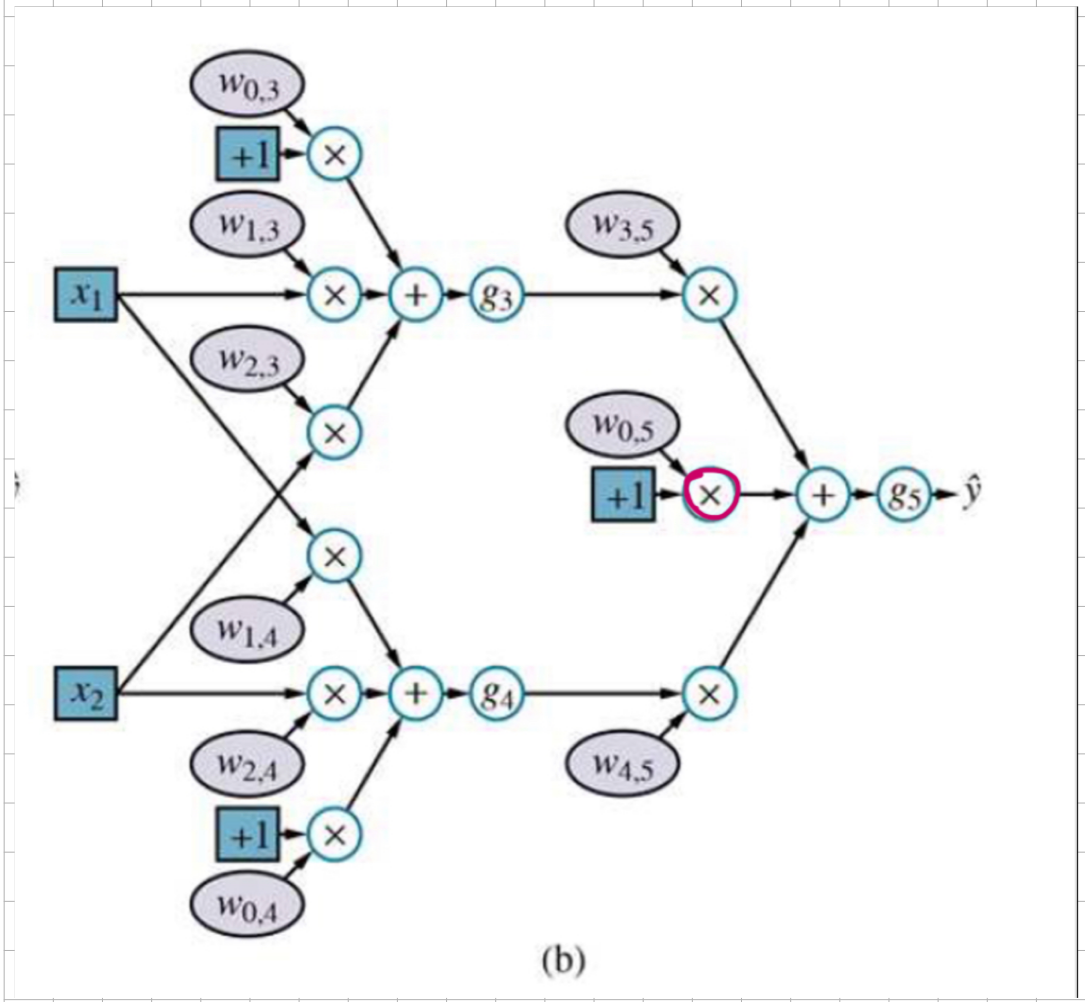
- Random
- Grid Search
- Hand Tuning
- Bayesian optization. → loss will be output
  → λ, α will be feature.
- Population based.

Model Selection

```
err ← [ ]
for size 1 to ∞
    err[size] = CrossValidation (training set, K)
    if err increase significantly then,
        best size = (err[size])_argmin
        h = learner (best size, training set)
        Return _ ErrorRate ( h, test set)
```



Now, for Nueral network implementation :



(b)

বসে আঁকি বিরিয়ানি,
তোদের তো জানি না
বন্ধের ফিরিতের জ্বালায়
খায়া মাগাইছে ফিরিতি কিখাইছে
চেওড়ানা বানাইছে
কিক জাদু করিয়া
হারা! বন্দে খায়া মাগাইছে

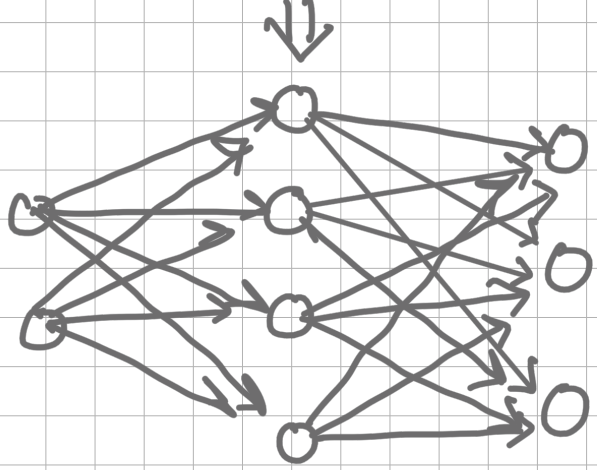
$$h(w) = a_5 = g(in_5)$$

$$= g(w_{0,5} + w_{3,5} a_3 + w_{4,5} a_4)$$

$$= g(w_{0,5} + w_{3,5}\, g(in_3) + w_{4,5} g(in_4))$$

এভাবে চলতে থাকিবে।

formula's :



W (4×3) → next layer neuron.

(4) × (3)

prev layer neuron,

$$W = \begin{bmatrix} w_{0,1} & w_1 \\ w_{1,1} & \\ w_{2,1} & \\ w_{3,1} & \end{bmatrix}$$

4 features.

next layer input should be 3 features.

a

da

যে 3 টা আসবে

$$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial a} \frac{\partial a}{\partial z} \frac{\partial z}{\partial w}$$

$$z = W^T X$$
$$= (3 \times 4)(4 \times 1)$$
$$= (3 \times 1)$$
$$a = g(z)$$

forward.

→ So, in dense layer:

$$\delta^l = \frac{\partial L}{\partial z} = \frac{\partial L}{\partial a} \cdot \frac{\partial a}{\partial z} = \frac{\partial L}{\partial a} \cdot f'(z)$$

↳ gradient in the current layer

↳ derivative of the activation function

$$\boxed{\delta^{l-1} = \delta^l \, W^l}$$

$$\frac{\partial L}{\partial w} = \delta^l \cdot a^{l-1}$$

So, in backprop: (This part)

- grad out = activation.back(grad out)
  (4x1) (4x1) → grad_out
- grad weights = grad_out · self.input^T
  (4x3) = (4x3) (4x1) ↳ 1x3 ↳ from previous (cur) output
- update weights
- grad input = grad out^T · weights = (1x3)
  ↳ (1x4) · (4x3)

dropout

Adam optimizer

$$m_t = \beta_1 * m_{t-1} + (1 - \beta_1) * g_t$$

$$v_t = \beta_2 * v_{t-1} + (1 - \beta_2) * g_t^2$$

$$\hat{m}_t = m_t / (1 - \beta_1^t)$$

$$\hat{v}_t = v_t / (1 - \beta_2^t)$$

$$\theta = \theta - (\alpha * \hat{m}_t / \sqrt{(\hat{v}_t + \varepsilon)})$$

$\beta_1 \rightarrow$ Decay rate of momentum

g_t
↳ grad

$\beta_2 \rightarrow$ Decay rate of squared gradient

$\alpha \rightarrow$ learning rate / step size

Soft-max

↳

$$f(x_i) = \frac{e^{x_i}}{\sum e^x} = \frac{e^{x_i}}{e^{x_1} + e^{x_2} + \cdots + e^{x_i} + \cdots + e^{x_n}}$$

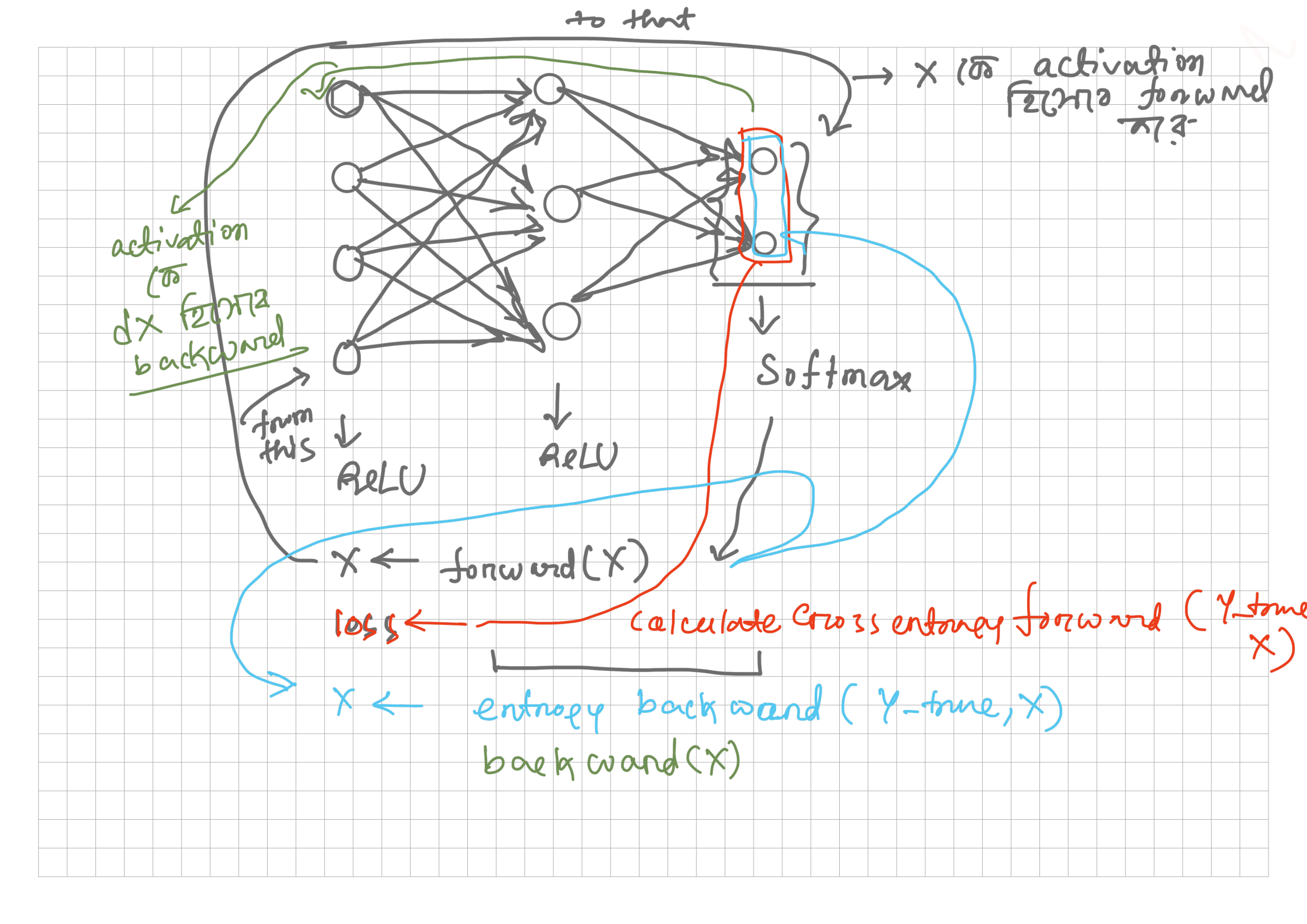
need to check for overflow

and,

$$f'(x_i) = \frac{\frac{d}{dx} e^{x_i} \cdot \sum e^x - e^{x_i} \frac{d}{dx_i} \sum e^x}{\left(\sum e^x\right)^2}$$

$$\frac{\partial}{\partial x_i} f(x_i) = \frac{e^{x_i}\left(\sum e^x - e^{x_i}\right)}{\left(\sum e^x\right)^2}$$

So, done yahoo }}}.


to this
→ X को activation
फिर्ता forward
गर्ने
activation
को
dX फिर्ता
backward
गर्ने
from this
ReLU
ReLU
Softmax
X ← forward(X)
loss ← calculate cross entropy forward (Y_true, X)
X ← entropy backward (Y_true, X)
backward(X)

for Softmax – details:

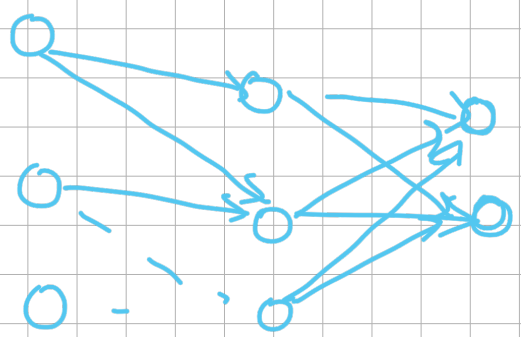
$$\frac{\partial L}{\partial z_i} = \frac{\partial L}{\partial f(z_i)} \frac{\partial f(z_i)}{\partial z_i} = \Delta_i f(z_i)\left(1 - f(z_i)\right)$$

$$\frac{\partial L}{\partial z_i} = \frac{\partial L}{\partial f(z_j)} \cdot \frac{\partial f(z_j)}{\partial z_i} = \Delta_j f(z_i) f(z_j) \quad (-1)$$
$$= -\Delta_j f(z_i) f(z_j)$$

$$\frac{\partial L}{\partial z_i} = f(z_i)\left\{\Delta_i - \sum \Delta_j f(z_j)\right\}$$

Now it holds. If $i = j$; $\Delta_i f(z_i)\{1 - f(z_i)\}$

If $i \neq j$; $f(z_i)\{0 - \sum \Delta_j f(z_j)\}$



x input

ReLu

$z = xw$

$a = \sigma(z)$

loss function

$-p \log p - (1-p) \log(1-p)$

x, z, a

x

$z = xw$

$a = \sigma(z)$

e are
ing activation
again in
l-1 th layer
no need of
computing
here

$$z^l = w^l x^l = w^l a^{l-1}$$

$$a = \sigma(z^l)$$

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

$$L = -\hat{y} \log \sigma(z) - (1-\hat{y}) \log (1-\sigma(z))$$

$$L = -\hat{y} \log \sigma(wx) - (1-\hat{y}) \log (1-\sigma(wx))$$

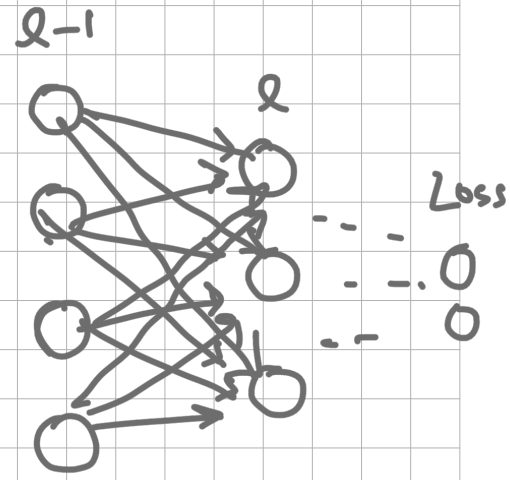
$$\frac{dL}{dw} = \frac{dL}{dz} \frac{dz}{dw}$$

$$= \frac{dL}{da^l} \cdot \frac{da^l}{dz^l} \cdot \frac{dz^l}{dw^l}$$

$$= \left(-\frac{\hat{y}}{a} + \frac{1-\hat{y}}{1-a}\right) a(1-a) \frac{dz}{dw}$$

$$= \underbrace{\left(-\frac{\hat{y}}{a} + \frac{1-\hat{y}}{1-a}\right)}_{\text{for loss}} \underbrace{a(1-a)}_{\text{for activation}} \underbrace{x}_{\text{input of layer } l}$$

we
ll
ude it
rately
l-1 th
layer
as
layer



Now, again $\rightarrow \frac{dz^l}{dz^{l-1}} = w^l$ {as c
copu

$z^l = w^l a^{l-1}$ and

So,

$$\frac{dL}{dz^{l-1}} = \frac{dL}{dz^l} \cdot \frac{dz^l}{dz^{l-1}}$$

$$= \frac{dL}{dz^l} \cdot w^l$$

actually,

$$\frac{dz^l}{dz^{l-1}} = \frac{dz^l}{da^{l-1}} \left(\frac{da^{l-}}{dz^{l-}} \right.$$

sea
w
com
sep
in

Sqn
1 t

* ignoring β for batch norm.

$$\mu = \frac{1}{m}\sum x_i$$

$$\sigma^2 = \frac{1}{m}\sum (x_i - \mu)^2$$

$$\hat{x}_i = \frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}}$$

$$y_i = \gamma \hat{x}_i \Rightarrow y_i = \gamma \frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}}$$

suppose l is a function of y.

$$\frac{\partial l}{\partial \gamma} = \frac{\partial l}{\partial y_i} \cdot \frac{\partial y_i}{\partial \hat{x}_i} = \frac{\partial l}{\partial y_i} \cdot \hat{x}_i = \Delta \hat{x}_i$$

$$\frac{\partial L}{\partial x_i} = \frac{\partial L}{\partial y_i} \cdot \frac{\partial y_i}{\partial \hat{x}_i} = \Delta \gamma$$

$$\hat{x}_i = \frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}}$$

$$\frac{\partial}{\partial \mu}\left(\sqrt{\sigma^2+\epsilon}\right)$$

and,

$$\frac{\partial L}{\partial \sigma^2} = \sum \frac{\partial L}{\partial \hat{x}_i} \cdot \frac{\partial \hat{x}_i}{\partial \sigma^2}$$

$$\left(\sqrt{\frac{1}{m}\sum (x_i-\mu)^2 + \epsilon}\right) \cdot \frac{\partial L}{\partial \hat{x}_i}$$

$$= \sum \frac{\partial L}{\partial \hat{x}_i} \frac{\partial}{\partial \sigma^2}\left(\frac{x_i - \mu}{\sqrt{\sigma^2+\epsilon}}\right)$$

$$= \sum \frac{\partial L}{\partial \hat{x}_i} \frac{\frac{\partial}{\partial \sigma^2}(x_i-\mu)\cdot\sqrt{\sigma^2+\epsilon} - \frac{1}{2}(x_i-\mu)(\sigma^2+\epsilon)^{-1/2}}{\sigma^2+\epsilon}$$

$$1/2 \cdot 2\frac{1}{m}\sum(x_i-\mu)$$

$$= \sum -\frac{\partial L}{\partial \hat{x}_i} \cdot \frac{1}{2}(x_i-\mu)(\sigma^2+\epsilon)^{-3/2}$$

$$\frac{\partial L}{\partial \mu} = \frac{\partial L}{\partial x_i'} \frac{\partial x_i'}{\partial \mu}$$

$$= \frac{\partial L}{\partial x_i'} \frac{\partial}{\partial \mu}\left(\frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}}\right)$$

$$= \frac{\partial L}{\partial x_i'} \frac{\sqrt{\sigma^2+\epsilon}\,(-1) - (x_i-\mu)\frac{1}{2\sqrt{\sigma^2+\epsilon}}\cdot 2\mu}{\sigma^2+\epsilon}$$

$$= \frac{\partial L}{\partial x_i'} \frac{-\sqrt{\sigma^2+\epsilon} - \frac{(x_i-\mu)^2}{\sqrt{\sigma^2+\epsilon}}}{\sigma^2+\epsilon}$$

$$= -\frac{\partial L}{\partial x_i'} \frac{1}{\sqrt{\sigma^2+\epsilon}} - \frac{\frac{\partial L}{\partial x_i'}(x_i-\mu)^2}{(\sigma^2+\epsilon)^{3/2}}$$

$$= -\frac{\partial L}{\partial x_i'} \frac{1}{\sqrt{\sigma^2+\epsilon}} - \frac{\frac{\partial L}{\partial x_i'}(x_i-\mu)\frac{1}{m}\sum(x_i-\mu)^2}{(\sigma^2+\epsilon)^{3/2}}$$

$$= -\frac{\partial L}{\partial x_i'} \frac{1}{\sqrt{\sigma^2+\epsilon}} + \frac{\partial L}{\partial x_i'}(x_i-\mu)\left\{\sigma^2+\epsilon\right\}^{-1/2} \frac{1}{m}\sum(x_i-\mu)$$

$$= -\frac{\partial L}{\partial x_i'} \frac{1}{\sqrt{\sigma^2+\epsilon}} + \frac{\partial L}{\partial x_i'}(x_i-\mu)(\sigma^2+\epsilon)^{-3/2}\cdot\frac{1}{m}\sum(x_i-\mu)$$

$$= -\frac{\partial L}{\partial x_i'} \frac{1}{\sqrt{\sigma^2+\epsilon}} + \sum -\frac{1}{2}\frac{\partial L}{\partial x_i'}(x_i-\mu)(\sigma^2+\epsilon)^{-3/2}(-2)\frac{1}{m}(x_i-\mu)$$

$$\frac{\partial L}{\partial \sigma^2} = \frac{\partial L}{\partial x_i'}\frac{\partial x_i'}{\partial \sigma^2}$$

$$= \frac{\partial L}{\partial x_i'}\frac{\partial}{\partial \sigma^2}\left(\frac{x_i-\mu}{\sqrt{\sigma^2+\epsilon}}\right)$$

$$= \frac{\partial L}{\partial x_i'}\frac{0 - (x_i-\mu)\frac{1}{2\sqrt{\sigma^2+\epsilon}}}{\sigma^2+\epsilon}$$

$$= -\frac{\partial L}{\partial x_i'}\frac{(x_i-\mu)}{2(\sigma^2+\epsilon)^{3/2}}$$

$$= -\frac{\partial L}{\partial x_i'}\frac{(x_i-\mu)}{2}\left\{\sigma^2+\epsilon\right\}^{-3/2}$$

$$= -\frac{\partial L}{\partial x_i'}\frac{(x_i-\mu)}{2}\left\{\frac{1}{m}\sum(x_i-\mu)^2+\epsilon\right\}^{-3/2}$$

$$\frac{\partial L}{\partial x_i} = \frac{\partial L}{\partial x_i'}\frac{\partial x_i'}{\partial x_i} + \frac{\partial L}{\partial \mu}\frac{\partial \mu}{\partial x_i} + \frac{\partial L}{\partial \sigma^2}\frac{\partial \sigma^2}{\partial x_i}$$

$$= \frac{\partial L}{\partial x_i'}\frac{1}{\sqrt{\sigma^2+\epsilon}} + \frac{\partial L}{\partial \mu}\frac{1}{m} + \frac{\partial L}{\partial \sigma^2}\frac{2(x_i-\mu)}{m}$$

$= - \frac{\partial y_i}{\sqrt{\sigma^2 + \epsilon}} + \frac{}{\partial \sigma^2} \quad -2 \quad \frac{}{m} (x_i - \mu$