P-1

MLE distribution:

~~$\delta_{ML}(y) = \arg\max_{0 \le i \le M-1} P(y|$~~

If likelihood fn is $P(x|\theta)$, MLE for θ is:

$$\theta_{MLE} = \arg\max_{\theta} P(x|\theta)$$

$$= \arg\max_{\theta} \prod_i P(x_i|\theta) \rightarrow (1)$$

for MAP for θ:

$$\theta_{MAP} = \arg\max_{\theta} P(x|\theta).P(\theta)$$

$$= \arg\max_{\theta} \prod_i P(x_i|\theta).P(\theta) \rightarrow (2)$$

if $i = 0, \ldots M-1$.

$P(\theta) = \frac{1}{M}$ in case of equal priors.

from (1),

$$\theta_{MLE} = \arg\max_{\theta} \left(\sum_i \log P(x_i|\theta)\right)$$

from (2),

$$\theta_{MAP} = \arg\max_{\theta} \left(\sum_i \log P(x_i|\theta) + \log P(\theta)\right)$$

$$= \arg\max_{\theta} \left(\sum_i \log P(x_i|\theta) + \log\left(\frac{1}{M}\right)\right)$$

$$= \arg\max_{\theta} \left(\sum_i \log P(x_i|\theta) - \log M\right)$$

Now we see that $\log M$ doesn't vary with θ.

Hence maximising $\left(\sum_i \log P(x_i|\theta) - \log M\right)$ is equivalent to maximising $\sum_i \log P(x_i|$ over θ.

Hence MLE and MAP distributions are same in case of uniform prior.

1.b) Given N observations $x_1, x_2 \dots x_N$,

likelihood of these N observations coming from a Gaussian distribution $N(\mu, \sigma^2)$ is given by:

$$p(x_1, x_2 \dots x_N | \mu, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

Using MLE,

$$p(x_1, x_2 \dots x_N | \mu, \sigma^2) = \sum_{i=1}^{N} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right)$$

$$= \sum_{i=1}^{N} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \log\left(e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right)$$

$$= \sum_{i=1}^{N} -\frac{1}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2}$$

$$= -\frac{N}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2}$$

To get max $p(x|\theta)$ we can just take the gradient of $p(x|\theta)$ w.r.t θ

$$\nabla_\theta p = 0 .$$

$\nabla p(x_1, \dots x_N | \mu, \sigma^2)$ w.r.t μ : (σ^2 known)

$$\frac{\partial}{\partial \mu} \cdot \left(-\frac{N}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2}\right) = 0 .$$

$$\Rightarrow \frac{1}{2\sigma^2} \sum_{i=1}^{N} 2(x_i - \mu) = 0 .$$

$$\sum_{i=1}^{N} (x_i - \mu) = 0 .$$

$$\sum_{i=1}^{N} x_i - \sum_{i=1}^{N} \mu = 0 .$$

$$\sum_{i=1}^{N} x_i - \mu N = 0$$

$$\therefore \mu = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Similarly taking gradient w.r.t σ^2, (μ known)

$$\frac{\partial}{\partial \sigma^2}\left(-\frac{N}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{N}\frac{(x_i-\mu)^2}{2\sigma^2}\right) = 0.$$

$$-\frac{N}{2}\cdot\frac{2\pi}{2\pi\sigma^2} - \sum_{i=1}^{N}\frac{-(x_i-\mu)^2\cdot 2}{(2\sigma^2)^2} = 0.$$

$$-\frac{N}{2\sigma^2} + \sum_{i=1}^{N}\frac{(x_i-\mu)^2}{2\sigma^4} = 0.$$

$$\sum_{i=1}^{N}\frac{(x_i-\mu)^2}{\sigma^4} = \frac{N}{\sigma^2}$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$$

MLE just gives the mean and variance of the distribution as parameters of the Gaussian distribution and doesn't make any assumptions (unbiased)

1.c) for MAP,

$$\theta_{MAP} = \arg\max_{\theta}\left(\sum_{i} \log P(x_i|\theta) + \log P(\theta)\right)$$

$$P(x_1, x_2 \ldots x_N | \mu, \sigma^2) = \prod_{i=1}^{N}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$\frac{\partial}{\partial\mu}\cdot\left[-\frac{N}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{N}\frac{(x_i-\mu)^2}{2\sigma^2} + \log P(\mu)\right] = 0.$$

$$\frac{1}{\sigma^2}\sum_{i=1}^{N}(x_i-\mu) + \frac{\partial}{\partial\mu}\log P(\mu) = 0.$$

$$\frac{\partial}{\partial\mu}\log P(\mu) = -\frac{1}{\sigma^2}\sum_{i=1}^{N}x_i + \frac{N\mu}{\sigma^2}$$

$$\mu = \frac{\sigma^2}{N}\left[\frac{1}{\sigma^2}\sum_{i=1}^{N}x_i + \frac{\partial}{\partial\mu}\log P(\mu)\right]$$

$$\mu = \frac{1}{N}\sum_{i=1}^{N}x_i + \frac{\sigma^2}{N}\cdot\frac{\partial}{\partial\mu}\log P(\mu)$$

$$P(\mu) = \frac{1}{\sqrt{2\pi\gamma^2}}e^{-\frac{\mu^2}{2\gamma^2}}$$

$$\log P(\mu) = -\frac{1}{2}\log(2\pi\gamma^2) - \frac{\mu^2}{2\gamma^2}$$

$$\frac{\partial}{\partial\mu}\log P(\mu) = -\frac{2\mu}{2\gamma^2} = -\frac{\mu}{\gamma^2}$$

$$\mu = \frac{1}{N}\sum_{i=1}^{N}x_i - \frac{\sigma^2}{N}\cdot\frac{\mu}{\gamma^2}.$$

$$\mu\left[1 \to \frac{\sigma^2}{N\gamma^2}\right] = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\mu = \frac{\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right) N\gamma^2}{(N\gamma^2+\sigma^2)} = \gamma^2 \frac{\left(\sum_{i=1}^{N} x_i\right)}{(N\gamma^2+\sigma^2)}$$

P-2.

a) from directional derivatives we know

$$\nabla_{\vec{v}} f = \lim_{h\to 0} \frac{f(x+h\vec{v}) - f(x)}{h}$$

$$\Rightarrow f(x+h\vec{v}) = f(x) + h\nabla f^T \vec{v}$$

$$h\vec{v} = x' - x$$

$$x' = x + h\vec{v}$$

$$J(w^k) = \frac{1}{2m}\cdot\left[\sum_{i=1}^{m} h_w(x_i) - y\right]^2 = \| h w_k(x) - y \|_2^2$$

$$J(w^{k+1}) = J(w^k) + (w^{k+1} - w^k)\nabla J$$

$$f(x+h\vec{v}) = f(x) + h\nabla f^T\vec{v} + \frac{h^2}{2}\cdot\vec{v}^T H\vec{v} + \dots$$

$$\therefore f(w^{k+1}) = f(w^k) + (w^{k+1}-w^k)\nabla f + \frac{(w^{k+1}-w^k)^T H (w^{k+1}-w^k)}{2} + \dots$$

$$J(w^k) = \frac{1}{2m}\| h w_k(x) - y \|_2^2$$

$$\therefore J(w^{k+1}) = J(w^k) + v^T \nabla_J J + \frac{1}{2}\cdot v^T H v + \dots \quad \rightarrow (1)$$

where

$v = w^{k+1} - w^k$

$\nabla_J J$ = Jacobian of J w.r.t w_k

H = Hessian of J w.r.t w_k

b) Also we know from grad. descent in directional derivatives,

if $\vec{v} = -\lambda \nabla_J J$ $\quad \nabla_J J$ = Jacobian

$\vec{v} = w^{k+1} - w^k$

$\therefore w^{k+1} - w^k = -\lambda \nabla f$

$$J(w^k) = \frac{1}{2m}\| h_{w_k}(x) - y \|_2^2 \quad \Rightarrow w^{k+1} = w^k - \lambda \nabla f$$

From eqn (1), $J(w^{k+1})$

$= J(w^k) - \lambda \nabla J^T \nabla J \dots$

∇J = Jacobian w.r.t w_k

Now we know if ~~a~~ vector $x^T x = 0$ only if x is a null vector, $x^T x = \sum_{i=1}^{N} x_i^2$ which is a +ve quantity.

∇J w.r.t w_k is ~~a vector of dimensions same as~~ w_k, so $\nabla J^T \nabla J$ is a +ve quantity. $\nabla J^T \nabla J = \|J\|_2^2$

$J(w^{k+1}) = J(w^k) - \lambda$ (some +ve quantity)

So J decreases in every Iteration of Gradient descent.

c) Optimal learning rate:

$$w^{k+1} = w^k - \lambda \nabla J_{w_k}$$

$$J(w^{k+1}) \leq J(w^k) - \lambda \nabla J_{w_k}^T \nabla J_{w_k}$$

$$J(w^k - \lambda \nabla J_{w_k}) = J(w^k) - \lambda \nabla J_{w_k}^T \nabla J_{w_k}$$

$$J(w^{k+1}) = J(w^k) + v \nabla J_{w_k} + \frac{1}{2} v^T H v + \ldots$$ from eqn ①.

$$v = w^{k+1} - w^k$$

$$v = -\lambda \nabla J$$

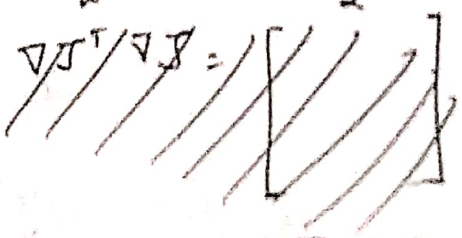
$$J(w^{k+1}) = J(w^k) - \lambda \nabla J_{w_k}^T \nabla J_{w_k} + \frac{\lambda^2}{2} \nabla J_{w_k}^T H \nabla J_{w_k}$$

At optimal learning rate we should have reached the global minima ~~in a single step~~, i.e. cost fn won't change any further.

$$J(w^{k+1}) = J(w^k)$$

$$\nabla J_{w_k}^T \nabla J_{w_k} = \frac{\lambda^2}{2} \nabla J_{w_k}^T H \nabla J_{w_k}$$

$\nabla J^T / \nabla J =$



$$\lambda = \frac{2 \nabla J_{w_k}^T \nabla J_{w_k}}{\nabla J_{w_k}^T H \nabla J_{w_k}}$$

λ = optimal learning rate

d) From eqn (i),

$$J(w^{k+1}) = J(w^k) + v^T \nabla J + \frac{1}{2} v^T H v$$

$$v = w^{k+1} - w^k$$

$$0 = \frac{d}{dv} J(w^k) + v^T H$$

$$\frac{d}{dv} J(w^k) = -v^T H$$

$$v^T = -\frac{d}{dv} J(w^k) \cdot H^{-1}$$

$$\frac{d}{dv} J(w^k) = \nabla J$$

$$0 = \nabla J + v^T \nabla H$$

$$v = -(\nabla H)^{-1} \nabla J$$

$$\therefore w^{k+1} = w^k - (\nabla H)^{-1} \nabla J$$

e) $$p_j \leftarrow \mu p_j - \frac{\alpha}{N} \sum_{i=1}^{N} x_j^{(i)} (y^{(i)} - t^{(i)})$$

$$w_j \leftarrow w_j + p_j$$

x : $N \times D$

y : $N \times 1$ $\quad$ t : $N \times 1$

w : $D \times 1$ $\quad$ p : $D \times 1$

$$y^{(i)} - t^{(i)} = y - t \quad (N \times 1)$$

x_j : $1 \times D$

$j \rightarrow 1$ to N

for each j

$y^{(i)} - t^{(i)}$: $N \times 1$

p_j : 1×1

x_j : $N \times 1$

$$\frac{\alpha}{N} \sum_{i=1}^{N} x_j^{(i)} (y^{(i)} - t^{(i)}) = \frac{\alpha}{N} x_j^T (y - t)$$

$$p = \mu p - \frac{\alpha}{N} x^T (y - t)$$

$$w = w + p$$

P-3.

a) $w^{k+1} = w^k - \alpha \frac{\partial}{\partial w_k} J(w_k)$

$$\frac{\partial}{\partial w_k} J(w_k) = \frac{\partial}{\partial w_k} \cdot \frac{1}{2m} \cdot \sum_{i=1}^{m} \left(h_\theta(x^i) - y^i\right)^2$$

$$= \frac{1}{2m} \cdot \frac{\partial}{\partial w_k} \cdot \sum_{i=1}^{m} \left(\sum_{j=0}^{n} w^j x_j^i - y^i\right)^2$$

$i \rightarrow$ iterator for no. of inputs

$j \rightarrow$ iterator over features

$$= \frac{1}{m} \cdot \sum_{i=1}^{m} \left(\sum_{j=0}^{n} w^j x_j^i - y^i\right) \cdot \frac{\partial}{\partial w_k} \cdot \left(\sum_{j=0}^{n} w^j x_j^i - y^i\right)$$

$$= \frac{1}{m} \cdot \sum_{i=1}^{m} \left(\sum_{j=0}^{n} w^j x_j^i - y^i\right) x_k^i$$

$$w^{k+1} = w^k - \frac{\alpha}{m} \cdot \sum_{i=1}^{m} \left(\sum_{j=0}^{n} w^j x_j^i - y^i\right) x_k^i$$

We see that the update of the weights w^k depends on the value of the input x_k. Now suppose that we have feature 1 ranging from $1 \rightarrow 10$ and another feature 2 ranging from $1 \rightarrow 10^6$. Gradient descent would take larger steps for feature 2 and way smaller steps for feature 1, which is undesirable, specially if feature 1 and 2 denote similar features with different units. Hence we need to do feature scaling.

4.a) $y = x^T \varphi x$ Consider $N=3$.

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \qquad \varphi = \begin{bmatrix} q_{11} & q_{12} & q_{13} \\ q_{21} & q_{22} & q_{23} \\ q_{31} & q_{32} & q_{33} \end{bmatrix}$$

$$y = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} q_{11} & q_{12} & q_{13} \\ q_{21} & q_{22} & q_{23} \\ q_{31} & q_{32} & q_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

$$= x_1 (q_{11}x_1 + q_{12}x_2 + q_{13}x_3) + x_2 (q_{21}x_1 + q_{22}x_2 + q_{23}x_3) + x_3 (q_{31}x_1 + q_{32}x_2 + q_{33}x_3)$$

$\frac{\partial y}{\partial \varphi}$ should have dimensions of φ.

$$\frac{\partial y}{\partial \varphi} = \begin{bmatrix} \frac{\partial y}{\partial q_{11}} & \frac{\partial y}{\partial q_{12}} & \frac{\partial y}{\partial q_{13}} \\ \frac{\partial y}{\partial q_{21}} & \frac{\partial y}{\partial q_{22}} & \frac{\partial y}{\partial q_{23}} \\ \frac{\partial y}{\partial q_{31}} & \frac{\partial y}{\partial q_{32}} & \frac{\partial y}{\partial q_{33}} \end{bmatrix} = \begin{bmatrix} x_1^2 & x_1x_2 & x_1x_3 \\ x_2x_1 & x_2^2 & x_2x_3 \\ x_3x_1 & x_3x_2 & x_3^2 \end{bmatrix}$$

$$= \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = x^T x$$

$$= \underset{3\times 1}{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}} \underset{1\times 3}{\begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}} = x x^T$$

$\frac{\partial y}{\partial x}$ is also the same dimensions as x.

$$\frac{\partial y}{\partial x} = \begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \frac{\partial y}{\partial x_2} \\ \frac{\partial y}{\partial x_3} \end{bmatrix} = \begin{bmatrix} 2q_{11}x_1 + q_{12}x_2 + q_{13}x_3 + x_2 q_{21} + x_3 q_{31} \\ x_1 q_{12} + q_{21}x_1 + 2q_{22}x_2 + q_{23}x_3 + q_{32}x_3 \\ x_1 q_{13} + x_2 q_{23} + q_{31}x_1 + q_{32}x_2 + 2q_{33}x_3 \end{bmatrix} = \begin{bmatrix} 2q_{11}x_1 + x_2(q_{12}+q_{21}) + x_3(q_{13}+q_{31}) \\ x_1(q_{12}+q_{21}) + x_2 2q_{22} + x_3(q_{23}+q_{32}) \\ x_1(q_{13}+q_{31}) + x_2(q_{23}+q_{32}) + 2q_{33}x_3 \end{bmatrix}$$

$$Q = \begin{bmatrix} q_{11} & q_{12} & q_{13} \\ q_{21} & q_{22} & q_{23} \\ q_{31} & q_{32} & q_{33} \end{bmatrix} \qquad Q^T = \begin{bmatrix} q_{11} & q_{21} & q_{31} \\ q_{12} & q_{22} & q_{32} \\ q_{13} & q_{23} & q_{33} \end{bmatrix}$$

$$Q + Q^T = \begin{bmatrix} 2q_{11} & q_{12}+q_{21} & q_{13}+q_{31} \\ q_{12}+q_{21} & 2q_{22} & q_{23}+q_{32} \\ q_{13}+q_{31} & q_{23}+q_{32} & 2q_{33} \end{bmatrix}$$

$$\frac{\partial y}{\partial x} = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} 2q_{11} & q_{12}+q_{21} & q_{13}+q_{31} \\ q_{12}+q_{21} & 2q_{22} & q_{23}+q_{32} \\ q_{13}+q_{31} & q_{23}+q_{32} & 2q_{33} \end{bmatrix}$$

$$= \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \left(\begin{bmatrix} q_{11} & q_{12} & q_{13} \\ q_{21} & q_{22} & q_{23} \\ q_{31} & q_{32} & q_{33} \end{bmatrix} + \begin{bmatrix} q_{11} & q_{21} & q_{31} \\ q_{12} & q_{22} & q_{32} \\ q_{13} & q_{23} & q_{33} \end{bmatrix} \right)$$

$$= x^T (Q + Q^T)$$

4.b)

Using linear reg. with L2 norm cost fn.

$$h_\theta(x) = \theta^T x$$

$$J(\theta) = \frac{1}{2m} \cdot (X\theta - y)^T (X\theta - y)$$

$$= \frac{1}{2m} \cdot (\theta^T X^T - y^T)(X\theta - y)$$

$$= \frac{1}{2m} \cdot (\theta^T X^T X\theta - y^T X\theta + y^T y - \theta^T X^T y)$$

Using commutativity of dot products

$$y^T (X\theta) = (\theta^T X^T) y$$

$$J(\theta) = \frac{1}{2m}\left(\theta^T X^T X \theta - 2(X\theta)^T y + y^T y\right)$$

$$\frac{\partial J(\theta)}{\partial \theta} = ?$$

Now ~~$X^T X$~~. $X^T X = \varphi$. As shown in part (a)

$$\frac{\partial}{\partial \theta}(\theta^T \varphi \theta) = \theta^T(\varphi + \varphi^T)$$

$$= \theta^T(X^T X + X^T X) = 2\theta^T X^T X$$

Let us consider $\frac{\partial}{\partial \theta}(X\theta)^T$. i.e $\frac{\partial}{\partial \theta} \cdot \theta^T X^T = \frac{\partial}{\partial \theta}(X\theta)$

We know ~~if $\frac{\partial y}{\partial x} = y^T$~~

$$X = [x_1\ x_2\ x_3] \quad X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

$$\theta^T X^T = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \\ \theta_{31} & \theta_{32} & \theta_{33} \end{bmatrix}^T \cdot [x_1\ x_2\ x_3]^T$$

$$= \begin{bmatrix} \theta_{11} & \theta_{21} & \theta_{31} \\ \theta_{12} & \theta_{22} & \theta_{32} \\ \theta_{13} & \theta_{23} & \theta_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

$$= \theta_{11}(x_1) + \theta_{21} x_2 + \theta_{31} x_3 + \theta_{12} x_1 + \theta_{22} x_2 + \theta_{32} x_3 + \theta_{13} x_1 + \theta_{23} x_2 + \theta_{33} x_3$$

$$= \begin{bmatrix} \theta_{11} x_1 + \theta_{21} x_2 + \theta_{31} x_3 \\ \theta_{12} x_1 + \theta_{22} x_2 + \theta_{32} x_3 \\ \theta_{13} x_1 + \theta_{23} x_2 + \theta_{33} x_3 \end{bmatrix}$$

$$\frac{\partial}{\partial \theta}[\theta^T X^T] = x_1\ x_2\ x_3$$

$$\frac{\partial}{\partial \theta}[\theta^T X^T] = \begin{bmatrix} & & \end{bmatrix}$$

We know $\frac{\partial}{\partial x} x^T A x = x^T(A + A^T)$

$$\frac{\partial}{\partial x} x^T Z = x^T\left(Z x^{-1} + x^{-T} Z^T\right)$$

$Z = Ax \Rightarrow A = Z x^{-1}$

$$= x^T Z x^{-1} + Z^T$$

$$= x^T A + Z^T$$

$$\frac{\partial}{\partial \theta}(\theta^T X^T) = \frac{\partial}{\partial \theta}(X\theta)$$ (by Commutativity of dot product)

$$\frac{\partial}{\partial \theta}(X\theta) = X^T$$

$$\frac{\partial J}{\partial \theta} = \frac{1}{2m} \cdot 2\,\theta^T X^T X - \frac{1}{2m} 2 \frac{\partial}{\partial \theta} (X\theta)^T y$$

$$= \frac{1}{m} \cdot \left[\theta^T X^T X - X^T y\right]$$

$$\frac{\partial J}{\partial \theta} = 0$$

$$\theta^T X^T X - X^T y = 0$$

$$X^T y = \theta^T X^T X \quad \Rightarrow X^T y = (X^T X)^T \theta$$

(using commutativity of dot prod).

~~$(X^T X)^T (\theta^T X^T X)$~~

~~$\theta^T = (X^T X)^{-1} X^T y$~~

$$X^T y = X^T X . \theta$$

$$\therefore \theta = (X^T X)^{-1} X^T y$$

c). $f_i = f(s_{y_i}) = \dfrac{e^{s_{y_i}}}{\sum_j e^{s_{y_j}}}$ (softmax fn).

~~$\frac{\partial F}{\partial s_{y_i}}$~~ $\dfrac{\partial f_i}{\partial s_{y_j}} = \dfrac{\partial}{\partial s_{y_j}} \left(\dfrac{e^{s_{y_i}}}{\sum_k e^{s_{y_k}}} \right) = \dfrac{h_i \frac{\partial}{\partial s_{y_j}} e^{s_{y_i}} - e^{s_{y_i}} \frac{\partial h_i}{\partial s_{y_j}}}{h_i^2}$

$= \dfrac{\partial}{\partial s_{y_j}} / h$

Let ~~$g_i = e^{s_{y_i}}$~~ $h = \sum_k e^{s_{y_k}}$

$\dfrac{\partial g_i}{\partial s_{y_j}} = \dfrac{\partial}{\partial s_{y_j}} e^{s_{y_i}} = e^{s_{y_i}} \cdot \dfrac{\partial}{\partial s_{y_j}} s_{y_i}$

$$\frac{\partial f_i}{\partial s_{y_j}} = \frac{h e^{s_{y_i}} \frac{\partial}{\partial s_{y_j}} s_{y_i} - e^{s_{y_i}} \cdot \frac{\partial}{\partial s_{y_j}} \sum_k e^{s_{y_k}}}{h^2}$$

$$\frac{\partial s_{y_i}}{\partial s_{y_j}} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \qquad \frac{\partial}{\partial s_{y_j}} \sum_k e^{s_{y_k}} = e^{s_{y_j}}$$

$$= -\frac{e^{s_{y_i}} e^{s_{y_j}}}{h \cdot h} = -F_i F_j \quad \text{if } i \neq j$$

$$\frac{h e^{s_{y_i}} - e^{s_{y_i}} e^{s_{y_j}}}{h^2} = \frac{e^{s_{y_i}}}{h} \left[\frac{h - e^{s_{y_j}}}{h}\right]$$

if $i=j$.

$$f_i \cdot \left[1 - \frac{e^{s_{y_j}}}{h\theta}\right] = f_i[1-f_j]$$

$$\frac{\partial f_i}{\partial s_{y_j}} = \begin{cases} f_i[1-f_j] & i=j \\ -f_i f_j & i \neq j \end{cases}$$

P-5.

a) $h_H = J$ $f_H = M$ stride = S, padding = P.

$h_W = K$ $f_W = N$

We know that

$$n_H' = \frac{n_H + 2p - f_H}{S} + 1$$

$$n_W' = \frac{n_W + 2p - f_W}{S} + 1$$

$$\therefore n_H' = \frac{J + 2p - M}{S} + 1$$

$$n_W' = \frac{K + 2p - N}{S} + 1$$

Size of resulting layer: $\left(\frac{J + 2p - M}{S} + 1\right), \left(\frac{K + 2p - N}{S} + 1\right)$

5b) Filter size = $F \times F$

stride = S

$n_H = J$ $n_W = K$

$$n_H' = \frac{n_H + 2p - F}{S} + 1 \qquad n_W' = \frac{n_W + 2p - F}{S} + 1$$

If all the dimensions are valid for the convolution, $p=0$ (Assumption)

$$n_H' = \frac{n_H - F}{S} + 1 \qquad n_W' = \frac{n_W - F}{S} + 1$$

5c) Yes fully connected layers can be represented as convolutional layers.

Suppose a FC layer takes an input of size $n_H \times n_W \times n_c$ and outputs n_f outputs.

We can use a CONV layer of filter size $(n_H \times n_W)$ and n_f filters of such shape with stride = 1, no padding

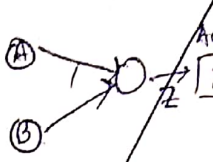
$n_H' = n_H - n_H + 1 = 1$

$n_W' = n_W - n_W + 1 = 1$

∴ size of output of conv layer = $1 \times 1 \times n_f$

P-5. Neural Nets & Backpropagation

a)



AND

A	B	X
0	0	0
0	1	0
1	0	0
1	1	1

$Z = w_0 A + w_1 B$

$X = \max(0, Z)$

$X = \max(0, w_0) = 0$

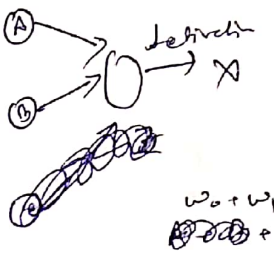
$w_0 \leq 0$. Similarly $w_1 \leq 0$ ..

$w_0 + w_1 = 1 \rightarrow (1)$

$w_0 = 1 - w_1$

P-5

a).



$Z = w_0 A + w_1 B + b$.

$X = \text{sigmoid}(Z)$

A	B	X
0	0	0
0	1	0
1	0	0
1	1	1

$w_0 + w_1 + b = Z \rightarrow (1)$

$Z > 4.6$.

$b = Z$.

$b < -4.6 \rightarrow (2)$

$w_0 + b < -4.6 \rightarrow (3)$

$w_1 + b < -4.6 \rightarrow (4)$.

15·4.

-4.6, no.

If $b = -20$, $w_0 < 15.4$

$w_1 < 15.4$.

$w_0 + w_1 > 24.6$

$w_0 = 14$

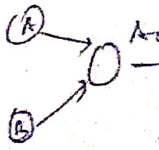
$w_1 = 15$

$b = -20$

if $b =$ -20

$w_0 =$ 8

$w_1 = 8$



$Z = w_0 A + w_1 B + b$

A	B	X
0	0	0
0	1	1
1	0	1
1	1	1

$b < -4.6$

$w_1 + b > 4.6$

$w_0 + b > 4.6$

$w_0 + w_1 + b > 4.6$

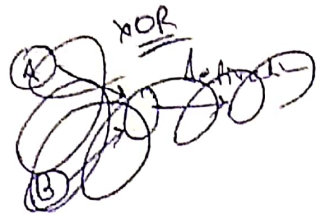
If $b = -10$

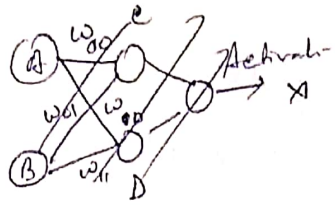
$w_1 = 15$

$w_2 = 20$

satisfies this condition.

b)



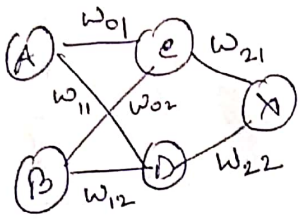


A	B	X
0	0	0
0	1	1
1	0	1
1	1	0

Using sigmoid activation fn.

$X = \bar{A} \cdot B + A \cdot \bar{B}$

$C = \bar{A} \cdot B$ $\quad D = A \cdot \bar{B}$



A	B	C
0	0	0
0	1	1
1	0	0
1	1	0

A	B	D
0	0	0
0	1	0
1	0	1
1	1	0

$C = w_{01} A + w_{02} B + b_0$

$w_{02} + b_0 > 4.6$

$b_0 < -4.6$

$w_{01} + b_0 < -4.6$

$w_{01} + w_{02} + b < -4.6$

if $b_0 = -10$

$w_{02} = 20$

$w_{01} = -25$

$D = w_{11} A + w_{12} B + b_1$

$w_{11} + b_1 > 4.6$

$w_{12} + b_1 < -4.6$

$b_1 < -4.6$

$w_{11} + w_{12} + b_1 < -4.6$

$b_1 = -15$

$w_{11} = 20$

$w_{12} = -5$

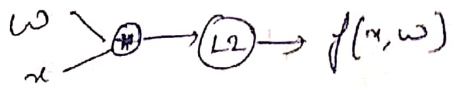
$X = w_{21} C + w_{22} D + b_2$

$w_{21} = 15$ $\quad b_2 = -10$ (using OR as above)

$w_{22} = 20$

Similarly we can do so for XNOR using a 2-layer network.

c) $f(x, w) = \| w \cdot x \|^2 = \sum_{i=1}^{n} (w \cdot x)_i^2 = \|y\|^2 = \sum_{i=1}^{n} y_i^2$



f is a scalar fn. $\quad y = wx$

$\frac{\partial y}{\partial x} = w^T$

$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial y} \cdot \frac{\partial y}{\partial x} \frac{\partial f}{\partial y}$

$= w^T \frac{\partial f}{\partial y} = 2 w^T y = 2 w^T w x$ (Ans)

$\frac{\partial}{\partial y_i} \sum_{i=1}^{n} y_i^2 = 2 y_i \quad \therefore \frac{\partial f}{\partial y} = 2y$

$$\frac{\partial f}{\partial w} = \frac{\partial f}{\partial y} \cdot \frac{\partial y}{\partial w} \qquad y = wx$$

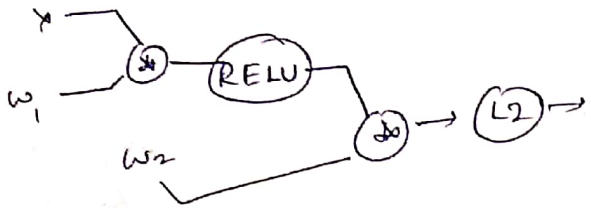
$$\frac{\partial f}{\partial w} = \frac{\partial f}{\partial y} \cdot x^T$$
$$= 2y \cdot x^T$$
$$= 2wx \cdot x^T$$

d) $z_1 = xw_1$

$h_1 = \max(0, z_1)$

$\hat{y} = h_1 w_2$

$L = \| \hat{y} \|_2^2$



$$\frac{\partial L}{\partial w_2} = \frac{\partial \hat{y}}{\partial w_2} \cdot \frac{\partial L}{\partial y}$$
$$= h_1^T \cdot \frac{\partial L}{\partial y}$$
$$= h_1^T \cdot 2y$$
$$= 2h_1^T y$$
$$= 2\max(0, xw_1)^T y$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial h_1} \cdot \frac{\partial h_1}{\partial z_1} \cdot \frac{\partial z_1}{\partial w_1}$$

$$\frac{\partial L}{\partial \hat{y}} = 2\hat{y}$$

$$\frac{\partial \hat{y}}{\partial h_1} = w_2^T$$

$$\frac{\partial h_1}{\partial z_1} = \begin{cases} 1 & z_1 > 0 \\ 0 & z_1 \leq 0 \end{cases}$$

$$\frac{\partial z_1}{\partial w_1} = x^T$$

if $z_1 > 0$, $\frac{\partial L}{\partial w_1} = 2\hat{y} \cdot w_2^T \cdot x^T$

if $z_1 \leq 0$, $\frac{\partial L}{\partial w_1} = 0$

where $z_1 = xw_1$

where $\hat{y} = \max(0,$

5.e)

$z = w_0 + w_1 x + w_2 x^2$

$y = 1 + e^z$

$L = \frac{1}{2}(\log y - \log t)^2$

$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial y} \cdot \frac{\partial y}{\partial z} \cdot \frac{\partial z}{\partial w_2}$

$\frac{\partial L}{\partial y_i} = \frac{\partial}{\partial y_i} \frac{1}{2}(\log y_i - \log t)^2$

$= \frac{1}{2} \cdot 2 (\log y_i - \log t) \cdot \frac{1}{y_i}$

$= (\log y_i - \log t) \cdot \frac{1}{y_i}$

$\therefore \frac{\partial L}{\partial y} = (\log y - \log t) \frac{1}{y}$

$\frac{\partial z}{\partial w_2} = 2 w_2 x$

$\frac{\partial y}{\partial z} = e^z$

$\frac{\partial L}{\partial w_2} = (\log y - \log t) \frac{1}{y} \cdot e^z \cdot 2 w_2 x$

where $z = w_0 + w_1 x + w_2 x^2$

$y = 1 + e^z$