HW1

1. $\dfrac{\partial E(v, h; W, b, c)}{\partial b_i}$

$$E(v, h; W, b, c) = -b^T v - c^T h - b^T W c$$

$\because \; c^T h = \sum_{i=1}^{n} c_i h_i$

$$= -\sum b_i v_i - \sum c_i h_i - [b_1, b_2 \cdots b_n] \begin{bmatrix} W_{11} & \cdots & W_{n1} \\ W_{21} & & \vdots \\ \vdots & & \vdots \\ \vdots & \cdots & W_{nn} \end{bmatrix} \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix}$$

$$= -\sum b_i v_i - \sum c_i h_i - \left[\sum_{1}^{n} b_1 W_{n1}, \sum_{1}^{n} b_2 W_{n2} \cdots \right] \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix}$$

$$= -\sum b_i v_i - \sum c_i h_i -$$

$$b^T W c = \cancel{\left[\sum b_i W_{n1}\right]}$$

$$= \underbrace{\left[\underbrace{\sum_{i=1}^{n} b_i W_{i1}}_{\text{scalar value}} \; \sum_{i=1}^{n} b_i W_{i2} \; \cdots \; \sum_{i=1}^{n} b_i W_{in} \right]}_{\text{row vector}} \times \underbrace{\begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix}}_{\text{column vector}}$$

$1 \times n \times n \times 1 = 1 \rightarrow$ scalar

$$\text{scalar} \rightarrow b^T W c = c_1 \cdot \sum_{i=1}^{n} b_i W_{i1} + c_2 \cdot \sum_{i=1}^{n} b_i W_{i2} + \cdots \; c_n \sum_{i=1}^{n} b_i W_{in}$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} c_j b_i W_{ij}$$

A) $\frac{\partial E(v,h;W,b,c)}{\partial b_i} = \frac{\partial(-\sum b_i v_i - \sum c_i h_i - \sum\sum c_j b_i W_{ij})}{\partial b_i}$

$= -v_i - 0 - c_j W_{ij}$

B) $\frac{\partial E(v,h;W,b,c)}{\partial c_i} = \frac{\partial(-\sum b_i v_i - \sum c_i h_i - \sum\sum c_j b_i W_{ij})}{\partial c_i}$

$= -0 - h_i - b_i W_{ij}$

c) $\frac{\partial E(v,h;W,b,c)}{\partial W_{ij}} = \frac{\partial(-\sum b_i v_i - \sum c_i h_i - \sum\sum c_j b_i W_{ij})}{\partial W_{ij}}$

$= -0 - 0 - c_j b_i$

D) The dimension of

1) $\frac{\partial E(v,h;W,b,c)}{\partial W} \Rightarrow \frac{\partial(-b^T v - c^T h - b^T W c)}{\partial W}$

$\because \frac{\partial a^T X b}{\partial X}$ - - - -> $\Rightarrow -0 - 0 - bc^T$

$\Rightarrow -bc^T$

b → column vector = $n \times 1$ / c ⇒ Transpose of column vector
↳ ⇒ $1 \times n$

$\therefore$ Dimension of $\frac{\partial E}{\partial W} = n\times 1 \times 1 \times n = \underline{\underline{n \times n}}$

2) $\dfrac{\partial E(v, h; W, b, c)}{\partial b} = \dfrac{\partial(-b^T v - c^T h - b^T W c)}{\partial b}$

$\Rightarrow -v - 0 - Wc$

$-v \rightarrow n \times 1$

$Wc \rightarrow n \times n \times n \times 1$

$\Rightarrow n \times 1$

$\therefore$ Dimension of $\dfrac{\partial E}{\partial b} \Rightarrow \underline{\underline{n \times 1}}$

3) $\dfrac{\partial E(v, h; W, b, c)}{\partial c} = \dfrac{\partial(-b^T v - c^T h - b^T W c)}{\partial c}$

$= -0 - h - b^T W$

$h \rightarrow n \times 1$

$b^T W \rightarrow 1 \times n \times n \times n = 1 \times n$

To add 2 vectors of different dimensions, you can take the transpose of one, which will get you same dimension vector. if they have same no. of elements i.e. 'n'

$\therefore$ The dimension of $\dfrac{\partial E}{\partial c} = \underline{\underline{n \times 1}}$ or $\underline{\underline{1 \times n}}$

2. random variable X follows geometric distribution (parameter 'p')

propability of X = k = $(1-p)^{k-1} p$
$\downarrow$
no. of trials needed before getting first success

Dataset $D = \{x_1, x_2 \dots x_N\}$

Given above details, the Likelihood function can be written as;

$$L(D|p) = \prod_{i=1}^{n} (1-p)^{x_i - 1} \times p$$

$$= (1-p)^{x_1 - 1} \times p \times (1-p)^{x_2 - 1} \times p \times \dots (1-p)^{x_n - 1} \times p$$

$$L(D|p) = p^n (1-p)^{\sum_1^n (x_i - n)} \quad \dots (1)$$

As its easier with logarithm :- product $\longrightarrow$ sum
$\therefore$ Taking log on both sides;

$$\ln L(D|p) = \ln\left((1-p)^{\sum_i^n x_i - n} \times p^n \right)$$

$$\ln(L(D|p)) = \sum_{i=1}^{n} (x_i - n) \ln(1-p) + n \ln p \quad \dots (2)$$

Further to get the Maximum likelihood estimation on parameter 'p' we will differentiate the above equation(2) and equate it to '0'.
This is because maximum of a function occurs when its derivative is equals to zero.

$$\frac{d[\ln L(D|p)]}{dp} = \frac{d \sum_1^n (x_i - n) \ln(1-p)}{dp} + \frac{d\, n \ln p}{dp}$$

$$\frac{d\left[\ln L(D|p)\right]}{dp} = -\sum_{i}^{n}(x_i - n) \times \frac{1}{(1-p)} + n \times \frac{1}{p} = 0$$

$$\therefore \quad \frac{n}{p} = \frac{\sum_{i}^{n}(x_i - n)}{1-p}$$

$$n(1-p) = p\left(\sum_{i}^{n}(x_i - n)\right)$$

$$n - \cancel{np} = p\sum_{i}^{n} x_i - \cancel{np}$$

$$n = p\sum_{i=1}^{n} x_i$$

$$\therefore p = \frac{n}{\left(\sum_{i=1}^{n} x_i\right)}$$

The MLE of $p \Rightarrow P = \dfrac{n}{\sum_{i=1}^{n} x_i} \approx \dfrac{1}{x}$

that is 1 success for each set of trial x

3. To perform bayesian parameter estimation on the 2nd Question,

where $X \rightarrow$ follows geometric distribution with 'p' parameter

Dataset $D = \{x_1, x_2 ---- x_n\}$

We need to find a prior distribution along with the likelihood function

As per Bayes theorem;

$$\underbrace{Pr(p|D)}_{\text{posterior}} = \frac{\underbrace{Pr(p)}_{\text{prior}} \underbrace{Pr(D|p)}_{\text{likelihood}}}{\underbrace{Pr(D)}_{\text{Marginal}}}$$

To find the posterior distribution on parameter 'p', we can choose any prior distribution, which might lead to expensive numerical computation. Thus, to avoid the massive computation for Bayesian inference it is preffered to consider a conjugate prior.

Conjugate prior are probablity distribution that belongs to the same family of likelihood distribution or it simply mimics the form of likelihood.

Beta distribution, is known as the conjugate prior of the geometric distribution that is followed by random variable X.

Set of Trials $\Rightarrow$ geometric distribution $\Rightarrow X(k) = (1-p)^{k-1} p$

Probablity of success $\Rightarrow$ Beta distribution $\rightarrow f(p) = \frac{1}{B(\alpha,\beta)} (1-p)^{\beta-1} p^{\alpha-1}$

I would choose the Beta distribution as the prior, as it is a conjugate prior of geometric distribution and thus greatly simplifies the Bayesian Inference calculations.

The posterior distribution of 'p' for dataset 'D' can be written as;

$$\underset{\text{Posterior}}{Pr(p|D)} = \frac{\overbrace{\left(\prod(1-p)^{x_i-1} \times p\right)}^{\text{likelihood}} \times \overbrace{Beta(\alpha,\beta)}^{\text{Prior}}}{Pr(D) \rightarrow \text{Marginal.}}$$

This can be further simplified as; $x_i = k$ & marginal likelihood is discarded → normalization constant $= \int = 1$

$$Pr(p|D) = \left((1-p)^{k-1} \; p \; \frac{1}{B(\alpha,\beta)} \; p^{\alpha-1}(1-p)^{\beta-1}\right) \times \text{constant}$$

$$p \times p^{\alpha-1} = p^{\alpha-1+1} = p^{\alpha}$$

Taking log on both sides;

$$\log(Pr(p|D)) = (k-1)\log(1-p) + \alpha\log p + (\beta-1)\log(1-p) + \text{constant}$$

To find maximum of this function, take derivative w.r.t 'p' and equate it to '0'.

$$\frac{d\log(Pr(p|D))}{dp} = \frac{-1\times(k-1)}{(1-p)} + \frac{\alpha}{p} - 1\frac{(\beta-1)}{(1-p)} = 0$$

∴ to find 'p';

$$\frac{\alpha}{p} = \frac{\beta-1}{(1-p)} + \frac{k-1}{(1-p)}$$

$$\frac{\beta-1+k-1}{1-p} = \frac{\beta+k-2}{1-p}$$

$$\alpha(1-p) = p\beta + pk - 2p$$

$$\alpha - \alpha p = p\beta + pk - 2p$$

$$\alpha = \alpha p + \beta p + kp - 2p$$

$$\alpha = p(\alpha + \beta + k - 2)$$

This is the BPE for the geometric distribution → $\boxed{p = \dfrac{\alpha}{\alpha + \beta + k - 2}}$

given the prior as beta distribution with parameters α & β