PROBLEM 1

40/40

1 (a) $P(X,Y,Z) = P(X)\,P(Y)\,P(Z/X,Y)$.

(b) let $P(X=1) = 0.2$

$P(Y=1) = 0.8$

This implies

$P(X=0) = 0.8$

$P(Y=0) = 0.2$.

Then, let us define :

~~P(Z)~~ $P(Z=1/X=1, Y=1) = 0.95$

$P(Z=1/X=1, Y=0) = 0.8$

$P(Z=1/X=0, Y=1) = 0.8$

$P(Z=1/X=0, Y=0) = 0.1$

This defines the distribution.

From here, we can calculate

$$P(Z=1, Y=1) = P(Z=1/X=1, Y=1)\,P(X=1)\,P(Y=1) + P(Z=1/X=0, Y=1)\,P(X=0)\,P(Y=1).$$

$$P(Z=1, Y=0) = P(Z=1/X=1, Y=0)\,P(X=1)\,P(Y=0) + P(Z=1/X=0, Y=0)\,P(X=0)\,P(Y=0).$$

$$P(Z=1) = P(Z=1, Y=1) + P(Z=1, Y=0)$$

$$= 0.712$$

$$P(Z=1, X=1) = P(Z=1/X=1, Y=1)\,P(X=1)\,P(Y=1) + P(Z=1/X=1, Y=0)\,P(X=1)\,P(Y=0)$$

$$P(X=1|Z=1) = P(Z=1, X=1)/P(Z=1)$$

$$= 0.258$$

$$P(X=1, Y=1, Z=1) = P(X=1)\,P(Y=1)\,P(Z=1|X=1, Y=1)$$

(from collider)

$$P(X=1|Y=1, Z=1) = P(X=1, Y=1, Z=1)/P(Z=1, Y=1)$$

$$= 0.229.$$

$$\therefore P(X=1) = 0.2 < P(X=1|Y=1, Z=1) = 0.229$$

$$< P(Z=1|X=1) = 0.258.$$

2.
(a) $(A \perp H)$ is NOT independent.
There is a path A - C - F - H which is not blocked. Both C and F have arrows that meet head to tail.

(b) $(B \perp G)$ is independent
E is a collider node and neither E nor its descendents are observed. Therefore E blocks B from G.

(c) $(B \perp G | E)$ is NOT independent.
E is collider node and observed. So E does not block. C has arrows tail - tail and F has head - tail but both are not observed.

(d) $(C \perp D | G, A)$ is NOT independent
One path is C - A - D. This is blocked because A has tail - tail arrows and is observed.
Other path is C - F - D. F is collider node and its descendent G is observed.
So F does not block.

(e) $(C \perp D)$ is NOT independent
A is tail - tail and not observed. So it does not block.
F is collider and not observed. So it blocks.

(b) $(B \perp D \mid C)$ is independent.

E is a collider, ~~not or as~~ its descendent C is observed. So E is not blocking.

C is head-tail and observed. So C will block both ~~A~~ B - E - C - A - D and B - E - CF - D paths, ~~so~~.

3> (a) We can write the joint distribution as:

$$p(x_1, x_2 \ldots x_n) = p(x_1)\, p(x_2 | x_1) \ldots p(x_{i+1} | x_i) \ldots p(x_n | x_{n-1})$$

Therefore,

$$P(X_n = 1 | X_1 = x_1) = P(X_n = 1, X_1 = x_1) / P(X_1 = x_1)$$

$$= \frac{\cancel{P(X_1 = x_1)} \sum P(X_2 | X_1 = x_1)\, P(X_2 | X_3) \ldots P(X_n = 1 | X_{n-1})}{P(X_1 = x_1)}$$

$$= \sum_{X_2, X_3 \ldots X_{n-1}} P(X_2 | X_1 = x_1)\, P(X_2 | X_3) \ldots P(X_n = 1 | X_{n-1})$$

In order to compute this, we write the following algorithm:

```
zero_old = px[n-1][1]
one_old = px[n-1][2]

for (j = n-2; j = 1; j--):
    zero_new = px[j][1] × one_old +
               (1 - px[j][1]) × zero_old.
    one_new = px[j][2] × one_old
              + (1 - px[j][2]) × zero_old.

    zero_old = zero_new
    one_old = one_new

if (x1 == 1):
    P(Xn = 1 | X1 = x1) = one_old.
else:
    P(Xn = 1 | X1 = x1) = zero_old.
```

(e) Proceeding similar to the previous part:

~~if (xn == 1)~~: $P(X_1 = 1 | X_n = x_n) = \dfrac{P(X_1 = 1, X_n = x_n)}{P(X_n = x_n)}$.

~~zero-old~~

if (x_n == 1):

zero-old = px[n-1][1]
one-old = px[n-1][2].

else:

zero-old = 1-px[n-1][1]
one-old = 1- px[n-1][2].

for (j = n-2; j = 1; j--):
    zero-new = px[j][1] × one-old + ~~px~~
        (1-px[j][1]) × zero-old
    one-new = px[j][2] × one-old
        + (1-px[j][2] × zero-old.
    zero-old = zero-new
    one-old = one-new

~~P(x1 = 1)~~

$P(X_1 = 1, X_n = x_n)$ = px1 × one-old

$P(X_n = x_n)$ = px1 × one-old + (1-px1) × zero-old

$P(X_1 = 1 | X_n = x_n) = \dfrac{px1 \times one\text{-}old}{px1 \times one\text{-}old + (1-px1) \times zero\text{-}old}$.

PROBLEM 2

1> We need to estimate $(K-1)\times d$ parameters.
These are : $W_1, W_2, \cdots W_{K-1}$ vectors.
Each vector has d parameters. So
$W_i = \{w_{i1}, w_{i2} \cdots w_{id}\}$.

2>

$$\ln P(Y = y_i / X = x_i)$$

$$= \ln \left[\frac{\exp(w_k^T x_i)}{1 + \sum_{l=1}^{K-1} \exp(w_l^T x_i)} \right] \quad \text{where } y_i = k$$

$$= w_k^T x_i - \ln\left(1 + \sum_{l=1}^{K-1} \exp(w_l^T x_i)\right) \quad \text{where } y_i = k$$

$$= \sum_{k=1}^{K} \delta_{ky_i} (w_k^T x_i) - \ln\left(1 + \sum_{l=1}^{K-1} \exp(w_l^T x_i)\right)$$

$$\therefore \quad L(W_1, W_2 \cdots W_{K-1}) = \sum_i \ln P(Y = y_i / X = x_i)$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} \delta_{ky_i} (w_k^T x_i) - \sum_{i=1}^{n} \ln\left(1 + \sum_{l=1}^{K-1} \exp(w_l^T x_i)\right)$$

if $\sum_{l=1}^{K-1} \exp(w_l^T x_i)$ is small, we can further simplify it

?? as.

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} \left(\delta_{ky_i} w_k^T x_i - \exp(w_l^T x_i) \right)$$

3) $$\frac{\partial L}{\partial w_{kr}} = \sum_{i=1}^{n} \sum_{i=1}^{K} \delta_{ky_i} x_{ir} - \sum_{i=1}^{n} \frac{x_{ir} \cdot \exp(w_k^T x_i)}{1 + \sum_{k=1}^{K-1} (w_l^T x_i)}.$$

for the r^{th} component of w_k.

So, we may write a vector derivative, w.r.t w_k as:

$$\frac{\partial L}{\partial w_k} = \sum_{i=1}^{n} \sum_{i=1}^{K} \delta_{ky_i} x_i - \sum_{i=1}^{n} \frac{x_i \exp(w_k^T x_i)}{1 + \sum_{k=1}^{K-1} (w_l^T x_i)}$$

$$= \sum_{i=1}^{n} \sum_{i=1}^{K} \delta_{ky_i} x_i - \sum_{i=1}^{n} x_i \; P(Y=k / X=x_i)$$

$$= \sum_{i=1}^{n} x_i \left(\sum_{i=1}^{K} \delta_{ky_i} - P(Y=k / X=x_i) \right).$$

4) $$f(w_1, \cdots w_{K-1}) = L(w_1 \cdots w_{K-1}) - \frac{\lambda}{2} \sum_{k=1}^{K-1} \|w_l\|_2^2$$

$$\frac{\partial f}{\partial w_k} = \frac{\partial L(w_1 \cdots w_{K-1})}{\partial w_k} - \frac{\lambda}{2} \cdot 2 w_k$$

$$= -\lambda w_k + \sum_{i=1}^{n} x_i \left(\sum_{i=1}^{K} \delta_{ky_i} - P(Y=k | X=x_i) \right).$$