Changhao Xu 2103530

1.

Question A: True

Question B: B

Question C: Left: K3, Middle: K1, Right: K2.

Question D: 2

Question E: C

Question F: False

Question G: A

Question H: False

Question I: True

Question J: False

Question K: B

Question L: True

Question M: True

Question N: True

Changhao Xu 2103530

2. Question 1: $P(\text{Grade} = A \mid \text{Happy?} = \text{Yes}) = \frac{1+3}{2+4} = \frac{2}{3}$.

$P(\text{Grade} = A \mid \text{Happy?} = \text{No}) = \frac{1+1}{2+4} = \frac{1}{3}$.

$P(\text{Grade} = C \mid \text{Happy?} = \text{Yes}) = \frac{1+1}{2+4} = \frac{1}{3}$.

$P(\text{Grade} = C \mid \text{Happy?} = \text{No}) = \frac{1+3}{2+4} = \frac{2}{3}$.

$P(\text{Year} = \text{Freshman} \mid \text{Happy?} = \text{Yes}) = \frac{1+1}{2+4} = \frac{1}{3}$.

$P(\text{Year} = \text{Freshman} \mid \text{Happy?} = \text{No}) = \frac{1+2}{2+4} = \frac{1}{2}$.

$P(\text{Year} = \text{Senior} \mid \text{Happy?} = \text{Yes}) = \frac{1+3}{2+4} = \frac{2}{3}$.

$P(\text{Year} = \text{Senior} \mid \text{Happy?} = \text{No}) = \frac{1+2}{2+4} = \frac{1}{2}$.

$P(\text{Happy?} = \text{Yes}) = \frac{4+2\times0.5}{2+8} = \frac{1}{2}$.

$P(\text{Happy?} = \text{No}) = \frac{4+2\times0.5}{2+8} = \frac{1}{2}$.

	Grade = A	Grade = C	Year = Freshman	Year = Senior
Happy? = Yes	$\frac{2}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{2}{3}$
Happy? = No	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{1}{2}$	$\frac{1}{2}$

	P(Happy?)
Happy? = Yes	$\frac{1}{2}$
Happy? = No	$\frac{1}{2}$

Changhao Xu 2103530

2. Question 2:

$$P(\text{Year} = \text{Freshman}, \text{Grade} = C, \text{Happy?} = \text{No}) = P(\text{Happy?} = \text{No})\, P(\text{Grade} = C \mid \text{Happy?} = \text{No})\, P(\text{Year} = \text{Freshman} \mid \text{Happy?} = \text{No})$$

$$= \frac{1}{2} \times \frac{2}{3} \times \frac{1}{2}$$

$$= \frac{1}{6}.$$

Question 3:

```
rand1 = random()
If rand1 < P(Happy? = Yes):
    Happy? = Yes
    rand2 = random()
    If rand2 < P(Year = Freshman | Happy? = Yes):
        Year = Freshman
      Else:
        Year = Senior
    rand3 = random()
    If rand3 < P(Grade = A | Happy? = Yes):
        Grade = A
      Else:
        Grade = C
Else:
    Happy? = No.
    rand2 = random()
    If rand2 < P(Year = Freshman | Happy? = No):
        Year = Freshman
      Else:
        Year = Senior
    rand3 = random()
    If rand3 < P(Grade = A | Happy? = No):
        Grade = A
      Else:
        Grade = C
return (Grade, Year, Happy?)
```

Changhao Xu 2103530

3. Question 1: $w^T x = \tilde{w}^T \tilde{x}$, $\tilde{x} = Ax$.

So $w^T x = \tilde{w}^T A x$

$w^T = \tilde{w}^T A$.

$w = A^T \tilde{w}$.

Question 2: Since $w = A^T \tilde{w}$, $\tilde{w} = (A^T)^{-1} w$

Therefore, $\arg\min_{\tilde{w}} \frac{\lambda}{2} \|\tilde{w}\|^2 + \sum_i (y_i - \tilde{w}^T \tilde{x}_i)^2$ can be rewritten as :

$\arg\min_w \frac{\lambda}{2} \|(A^T)^{-1} w\|^2 + \sum_i (y_i - w^T x_i)^2$.

Question 3: When A is a rescaling transformation, since y_i stays the same,

scaling the data x alone will scale the weight factor to $\tilde{w} = (A^T)^{-1} w$ accordingly,

Therefore, $\sum_i (y_i - w^T x_i)^2$ part remains the same with standard ridge regression.

On the other hand, considering the regularization part, since w has been scaled to $(A^T)^{-1} w$,

we need to replace regularizer $\frac{\lambda}{2} \|w\|^2$ with $\frac{\lambda}{2} \|(A^T)^{-1} w\|^2$,

which considers the scale change of weight factor w.

Therefore, we have new regularization $\frac{\lambda}{2} \|(A^T)^{-1} w\|^2$ instead of $\frac{\lambda}{2} \|w\|^2$ in standard ridge regression.

In summary, standard ridge regression has the form : $\arg\min_w \frac{\lambda}{2} \|w\|^2 + \sum_i (y_i - w^T x_i)^2$

and our transformed ridge regression scales regularizer, but keeps the minimization part :

$$\arg\min_w \frac{\lambda}{2} \|(A^T)^{-1} w\|^2 + \sum_i (y_i - w^T x_i)^2 .$$

Changhao Xu 2103530.

4. Question 1: Let data likelihood for dual point model be $P_1(S)$, data likelihood for single point model be $P_2(S)$.

$$\log P_1(S) - \log P_2(S) = \log \prod_{p\in S}\prod_{i=1}^{Mp} \frac{e^{-\|U(p^{[i]})-V(p^{[i-1]})\|_2^2}}{Z_1(p^{[i-1]})} - \log \prod_{p\in S}\prod_{i=1}^{Mp} \frac{e^{-\|X(p^{[i]})-X(p^{[i-1]})\|_2^2}}{Z_2(p^{[i-1]})}.$$

$$= \sum_{p\in S}\sum_{i=1}^{Mp} \log \frac{e^{-\|U(p^{[i]}-V(p^{[i-1]})\|_2^2}}{Z_1(p^{[i-1]})} - \sum_{p\in S}\sum_{i=1}^{Mp} \log \frac{e^{-\|X(p^{[i]})-X(p^{[i-1]})\|_2^2}}{Z_2(p^{[i-1]})}.$$

$$= \sum_{p\in S}\sum_{i=1}^{Mp} -\|U(p^{[i]})-V(p^{[i-1]})\|_2^2 - \log Z_1(p^{[i-1]}) + \|X(p^{[i]})-X(p^{[i-1]})\|_2^2 + \log Z_2(p^{[i-1]}).$$

where $Z_1(p^{[i-1]}) = \sum_{p^{[i]}} e^{-\|U(p^{[i]})-V(p^{[i-1]})\|_2^2}$, $Z_2(p^{[i-1]}) = \sum_{p^{[i]}} e^{-\|X(p^{[i]})-X(p^{[i-1]})\|_2^2}$.

$$= \sum_{p\in S}\sum_{i=1}^{Mp}\left\{\log \frac{\sum_{p^{[i]}} e^{-\|X(p^{[i]})-X(p^{[i-1]})\|_2^2}}{\sum_{p^{[i]}} e^{-\|U(p^{[i]})-V(p^{[i-1]})\|_2^2}} + \|X(p^{[i]})-X(p^{[i-1]})\|_2^2 - \|U(p^{[i]})-V(p^{[i-1]})\|_2^2\right\}.$$

Let $V=X$. Then:

$$= \sum_{p\in S}\sum_{i=1}^{Mp}\left\{\log \frac{\sum_{p^{[i]}} e^{-\|X(p^{[i]})-X(p^{[i-1]})\|_2^2}}{\sum_{p^{[i]}} e^{-\|U(p^{[i]})-X(p^{[i]})+X(p^{[i]}-X(p^{[i-1]})\|_2^2}} + \|X(p^{[i]})-X(p^{[i-1]})\|_2^2 - \|U(p^{[i]})-X(p^{[i]})+X(p^{[i]})-X(p^{[i-1]})\|_2^2\right\}$$

$$\geq \sum_{p\in S}\sum_{i=1}^{Mp} \log \frac{\sum_{p^{[i]}} e^{-\|X(p^{[i]})-X(p^{[i-1]})\|_2^2}}{\sum_{p^{[i]}} e^{+\|U(p^{[i]})-X(p^{[i]})\|_2^2}\cdot e^{-\|X(p^{[i]})-X(p^{[i-1]})\|_2^2}} + \|X(p^{[i]})-X(p^{[i-1]})\|_2^2 - \|U(p^{[i]})-X(p^{[i]})\|_2^2 - \|X(p^{[i]})-X(p^{[i-1]})\|_2^2$$

$$= \sum_{p\in S}\sum_{i=1}^{Mp} \log \sum_{p^{[i]}} \frac{1}{e^{\|U(p^{[i]})-X(p^{[i]})\|_2^2}} - \|U(p^{[i]})-X(p^{[i]})\|_2^2$$

$$= \sum_{p\in S}\sum_{i=1}^{Mp} \log \sum_{p^{[i]}} e^{-\|U(p^{[i]})-X(p^{[i]})\|_2^2} - \|U(p^{[i]})-X(p^{[i]})\|_2^2 \geq 0$$, if and only if $U(p^{[i]}) = X(p^{[i]})$ when the "=" holds,

considering that U, V, X are selected that maximize $P(S)$.

Question 2: When $P_1(S) = P_2(S)$, then $U = V = X$.

This means that the likelihood for dual point model is generally better than that of single point model, since the optimized choices of (U, V) has more tuning capability than X.

In the worst case, $U=V=X$, and dual point model generates the same symmetric probability with the single point model, i.e. any likelihood of single point model can be achieved by dual point model.

Changhao Xu 2103530

5. Question 1: $\frac{\partial}{\partial w_{11}} L(y, f(x)) = \frac{\partial}{\partial w_{11}} (y - f(x))^2 = 2(y - f(x)) \cdot \frac{\partial}{\partial w_{11}} (y - f(x)) = -2(y - f(x)) \frac{\partial f(x)}{\partial w_{11}}$.

$$= -2(y - f(x)) \frac{\partial}{\partial w_{11}} \sigma\left(\sum_{i=1}^{2} u_i h_i(x)\right)$$

$$= -2(y - f(x)) \cdot \sigma\left(\sum_{i=1}^{2} u_i h_i(x)\right)\left[1 - \sigma\left(\sum_{i=1}^{2} u_i h_i(x)\right)\right] \cdot \frac{\partial}{\partial w_{11}} \cdot \sum_{i=1}^{2} u_i h_i(x).$$

$$= -2(y - f(x)) \cdot \sigma\left(\sum_{i=1}^{2} u_i h_i(x)\right)\left[1 - \sigma\left(\sum_{i=1}^{2} u_i h_i(x)\right)\right] \cdot u_1 \frac{\partial h_1}{\partial w_{11}}$$

$$= -2(y - f(x)) \cdot \sigma\left(\sum_{i=1}^{2} u_i h_i(x)\right)\left[1 - \sigma\left(\sum_{i=1}^{2} u_i h_i(x)\right)\right] \cdot u_1 \cdot \frac{\partial}{\partial w_{11}} \sigma\left(\sum_{j=1}^{2} w_{j1} x_j\right)$$

$$= -2(y - f(x)) \cdot \sigma\left(\sum_{i=1}^{2} u_i h_i(x)\right)\left[1 - \sigma\left(\sum_{i=1}^{2} u_i h_i(x)\right)\right] u_1 \sigma\left(\sum_{j=1}^{2} w_{j1} x_j\right)\left[1 - \sigma\left(\sum_{j=1}^{2} w_{j1} x_j\right)\right] \frac{\partial}{\partial w_{11}} \left(\sum_{j=1}^{2} w_{j1} x_j\right)$$

$$= \frac{\partial}{\partial w_{11}} \cdot w_{11} x_1$$

$$= x_1$$

$$= -2(y - f(x)) \cdot \sigma\left(\sum_{i=1}^{2} u_i h_i(x)\right)\left[1 - \sigma\left(\sum_{i=1}^{2} u_i h_i(x)\right)\right] u_1 \sigma\left(\sum_{j=1}^{2} w_{j1} x_j\right)\left[1 - \sigma\left(\sum_{j=1}^{2} w_{j1} x_j\right)\right] x_1 .$$

Question 2: $h_1 = \sigma(w_{11} x_1 + w_{21} x_2) = \frac{e^{0.05}}{1 + e^{0.05}}$. $h_2 = \sigma(w_{12} x_1 + w_{22} x_2) = \frac{e^{-0.115}}{1 + e^{-0.115}}$. $f(x) = \sigma(h_1 u_1 + h_2 u_2) = 0.552097$.

$$\frac{\partial}{\partial w_{11}} L(y, f(x)) = -2(0.75 - 0.552097) \cdot 0.552097 \cdot (1 - 0.552097) \cdot 0.5 \cdot \frac{e^{0.05}}{1 + e^{0.05}} \left(1 - \frac{e^{0.05}}{1 + e^{0.05}}\right) \cdot 0.1$$

$$= -0.00122275.$$

Question 3: $\sigma\left(\sum_{i=1}^{2} u_i h_i(x)\right)\left[1 - \sigma\left(\sum_{i=1}^{2} u_i h_i(x)\right)\right]$ results in the vanishing gradient.

This term is the derivative of the activation function, and can be further decomposed into product of sereral partial derivatives. With more layers in neural networks, we will then have more products, and due to the activation function $\sigma(s) < 1$, each element of the products is less than 1.

Multiplying many terms less than 1 will exacerbate the vanishing gradient problem.