1. Given that $R_{tr}(B) = \frac{1}{N}\sum_{i=1}^{N}(y_i - B^T x_i)^2$

$R_{te}(B) = \frac{1}{M}\sum_{i=1}^{M}(\tilde{y}_i - B^T\tilde{x}_i)^2$

At regression model with p parameters

$y = XB + \epsilon$, where $\epsilon_i \overset{iid}{\sim} N(0, 6^2)$

solving $y = XB + \epsilon$,

I get the least square estimate of parameter B

$\hat{B} = (X'X)^{-1}X'y$

we know that $E(\hat{B}) = B$

Let $(X_1, y_1) \cdots (X_N, y_N)$ be the training data.

$(\hat{X}_1, \hat{y}_1) \cdots (\hat{X}_M, \hat{y}_M)$ be the test data

$E(y) = E(X\hat{B}) + E(\epsilon)$

$E(y) = XB$

So expectation of y under both training data and test data give the same result.

when $\epsilon_i \overset{iid}{\sim} N(0, 6^2)$

$y_i \sim N(X_i B, 6^2)$

$\hat{y}_i \sim N(\hat{X}_i B, 6^2)$

We know that $s_i^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \hat{B}^T x_i)^2$ is unbiased

s_i^2 is estimation of 6^2.

$E(s_i^2) = E\left(\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \hat{B}^T x_i)^2\right)$

we already know that

$$E(R_{tr}(\hat{B})) = E\left(\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{B}^T x_i)^2\right)$$

$$= \frac{N-1}{N} E\left(\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \hat{B}^T x_i)^2\right)$$

$$= \frac{N-1}{N} E(s_i^2)$$

$$= \frac{N-1}{N} 6^2$$

Similarly, let $s_2^2 = \frac{1}{M-1}\sum_{i=1}^{M}(\hat{y}_i - \hat{B}^T\hat{x}_i)^2$

which is an unbiased estimation of 6^2

$$E(s_2^2) = E\left(\frac{1}{M-1}\sum_{i=1}^{M}(\hat{y}_i - \hat{B}^T\hat{x}_i)^2\right)$$

$$E(R_{te}(\hat{B})) = E\left(\frac{1}{M}\sum_{i=1}^{M}(\hat{y}_i - \hat{B}^T\tilde{x}_i)^2\right)$$

$$= \frac{M-1}{M} E\left(\frac{1}{M-1}\sum_{i=1}^{M}(\hat{y}_i - \hat{B}^T\tilde{x}_i)^2\right)$$

$$= \frac{M-1}{M} 6^2$$

Since $E(R_{tr}(\hat{B})) = \frac{N-1}{N} 6^2$

and $E(R_{te}(\hat{B})) = \frac{M-1}{M} 6^2$

and we have $N \geq M \geq p$

we have $\frac{N-1}{N} \leq \frac{M-1}{M}$

thus $E(R_{tr}(\hat{B})) \leq E(R_{te}(\hat{B}))$

2 (a) Set $h(t) = f(x + t(y - x))$ for $t \in [0, 1]$

$h'(t) = \nabla f(x + t(y - x))^T (y - x)$.

Use equation (1)

$h'(t) - h'(0) = \nabla f(x + t(y-x))^T (y - x) - \nabla f(x)^T (y - x)$

$h(t) - h(0) = \nabla f(x + t(y-x))^T (y-x)$

$$= (\nabla f(x) - \nabla f(x + t(y-x)))^T \cdot (y-x)$$

$$\geq mt \|x-y\|_2^2$$

$$h'(t) \geq h'(0) + mt \|x-y\|_2^2$$

use FTC, I have

$$h(1) - h(0) = \int_0^1 h'(t)\, dt$$

$$f(y) - f(x) \geq \int_0^1 h'(0) - mt\|x-y\|_2^2\, dt$$

$$\leq h'(0) + \frac{m}{2}\|x-y\|_2^2$$

$$\geq \nabla f(x)^T (y-x) + \frac{m}{2}\|x-y\|_2^2$$

$$f(y) \geq f(x) + \nabla f(x)^T (y-x) + \frac{m}{2}\|x-y\|_2^2$$

Thus I proved that (1) is equivalent to

$$f(y) \geq f(x) + \nabla f(x)^T (y-x) + \frac{m}{2}\|y-x\|_2^2$$

(b) assume $g(x) = f(x) - \frac{m}{2}\|x\|_2^2$, where $\nabla g(x) = \nabla f(x) - mx$.

Therefore I get

$$(\nabla g(x) - \nabla g(y))^T \cdot (x-y) = \left(\nabla f(x) - \nabla f(y) - \frac{m}{2}(x-y)\right)^T (x-y)$$

$$= (\nabla f(x) - \nabla f(y))^T (x-y) - (m(x-y))^T (x-y)$$

$$\geq 0$$

Thus $g(x)$ is convex in domain of f.

Therefore $\nabla^2 g(x) \geq 0$ and I get $\nabla^2 f(x) - mI \geq 0$

which means that I proved that (1) is equivalent to $\nabla^2 f(x) \geq mI$

3 (a) $Pr(q_2 = \text{Happy}) = 0.8$

(b) $Pr(O_2 = \text{frown}) = 0.8 \times 0.1 + 0.2 \times 0.5$

$$= 0.08 + 0.1 = 0.18$$

(c) $Pr(q_2 = \text{Happy} \mid O_2 = \text{frown}) = \frac{0.8 \times 0.1}{0.18} = \frac{4}{9}$

(d) $Pr(O_{100} = \text{yell}) = Pr(O_{100} = \text{yell} \cap q_{100} = \text{happy}) + Pr(O_{100} = \text{yell} \cap q_{100} = \text{angry})$

$$= 0.2$$

(e) $q_1 = \text{happy}$

$p(f_2 \mid h_2) = 0.8 \times 0.1 = 0.08$

$p(f_2 \mid a_2) = 0.2 \times 0.5 = 0.1$ so $q_2 = \text{angry}$

$p(f_3 \mid h_3) = 0.2 \times 0.1 = 0.02$

$p(f_3 \mid a_3) = 0.8 \times 0.5 = 0.4$ so $q_3 = \text{angry}$.

the most likely sequence is happy, angry, angry.

4. (a) The optimization problem now becomes $R(\theta) + \lambda J(\theta)$

which equals $\sum_{i=1}^{n} \sum_{k=1}^{K} (y_{ik} - f_k(x_i))^2 + \lambda\left(\sum_{k,m} B_{km}^2 + \sum_{m,l} \alpha_{ml}^2\right)$

taking the gradient of $R(\theta) + \lambda J(\theta)$

I get $\frac{\partial R_i + \partial \lambda J(\theta)}{\partial B_{km}} = \frac{\partial R_i}{\partial B_{km}} + \frac{\partial \lambda J(\theta)}{\partial B_{km}}$

$$= \delta_{ki} z_{mi} + 2\lambda \sum_{k,m} B_{km}$$

and $$\frac{\partial R_i + \partial\lambda J(\theta)}{\partial a_{ml}} = S_{mi} X_{il} + 2\lambda \sum_{ml} a_{ml}$$

According to the problem, a gradient update at the $(r+1)$st iteration has the form

$$B_{km}^{(r+1)} = B^{(r)}_{km} - \frac{\partial R}{\partial B^{(r)}_{km}}, \quad a^{(r+1)}_{lm} = a^{(r)}_{ml} - \frac{\partial R}{\partial a^{(r)}_{ml}}$$

taking the new gradient update. $\frac{\partial R + \partial\lambda J(\theta)}{\partial B_{km}}$ and $\frac{\partial R + \partial\lambda J(\theta)}{\partial a_{ml}}$ into the gradient update.

I get $$B^{(r+1)}_{km} = B^{(r)}_{km} - \frac{\partial R}{\partial B^{(r)}_{km}} - \frac{\partial\lambda J(\theta)}{\partial B^{(r)}_{km}}$$

$$= B^{(r)}_{km} + 2(y_{ik} - f_k(x_i))\, g_k'(B_k^{(r)T} z_i) z_{mi} - 2\lambda \sum_{km} B^{(r)}_{km}$$

$$a^{(r+1)}_{ml} = a^{(r)}_{ml} + \sum_{k=1}^{K} 2(y_{ik} - f_k(x_i))\, g_k'(B_k^T z_i) B_{km}\, \sigma'(a_m^{(r)T} x_i) x_{il} - 2\lambda \sum_{ml} a^{(r)}_{ml}$$

Above are the gradient update for this regularized problem

(b) Stochastic gradient descent replaces the actual gradient, which is calculated by the entire dataset, with an estimate. The estimate is calculated from a randomly subset of the data. This could largely reduce the computational burden when n is large, achieving faster iterations in trade for a lower convergence rate