Homework 01

Question 1:

Because the noise is random, the outcomes may change over time. And I choose one of an outcome to discuss. Figure 1 is the required picture of Erms and Figure 2 gives a direct insight of training data, test data, estimated function and true function.

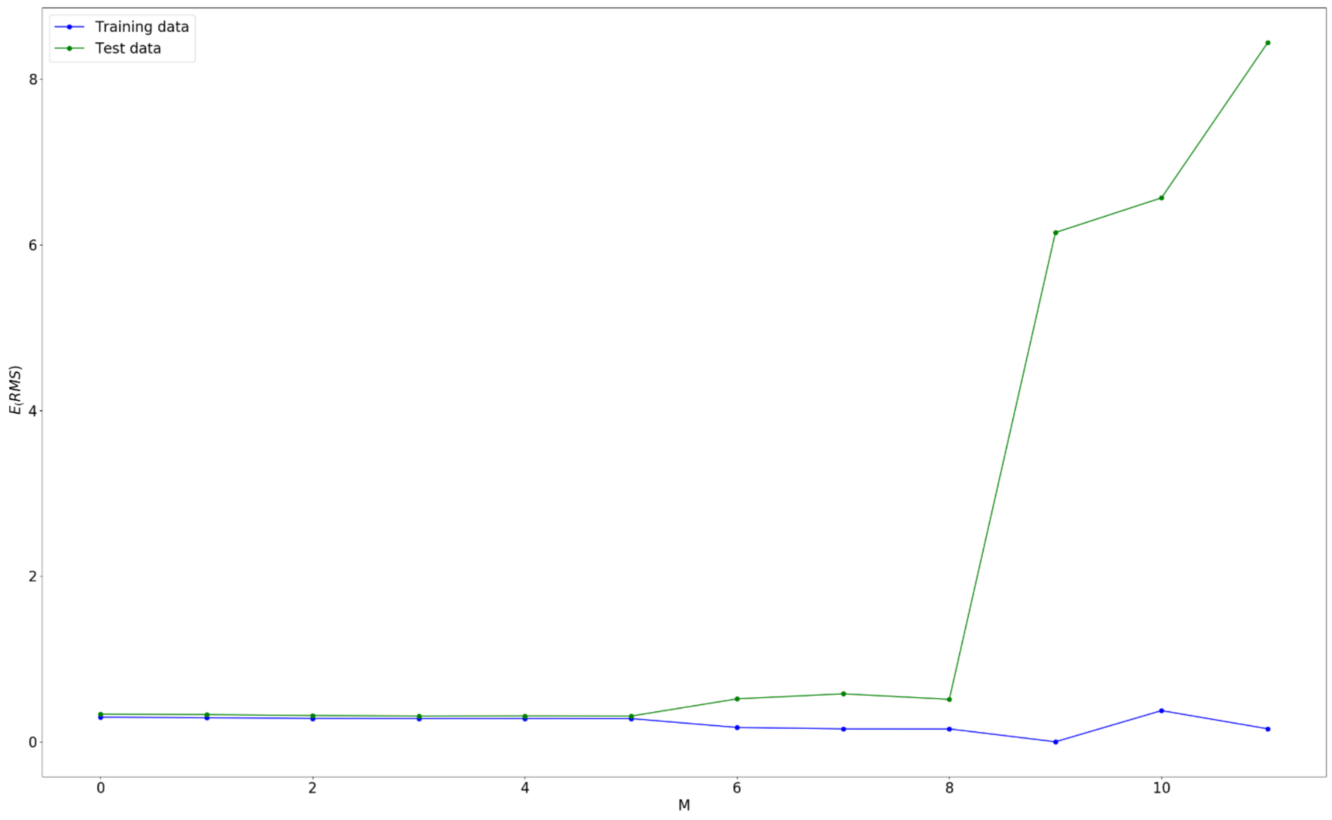

Figure 1

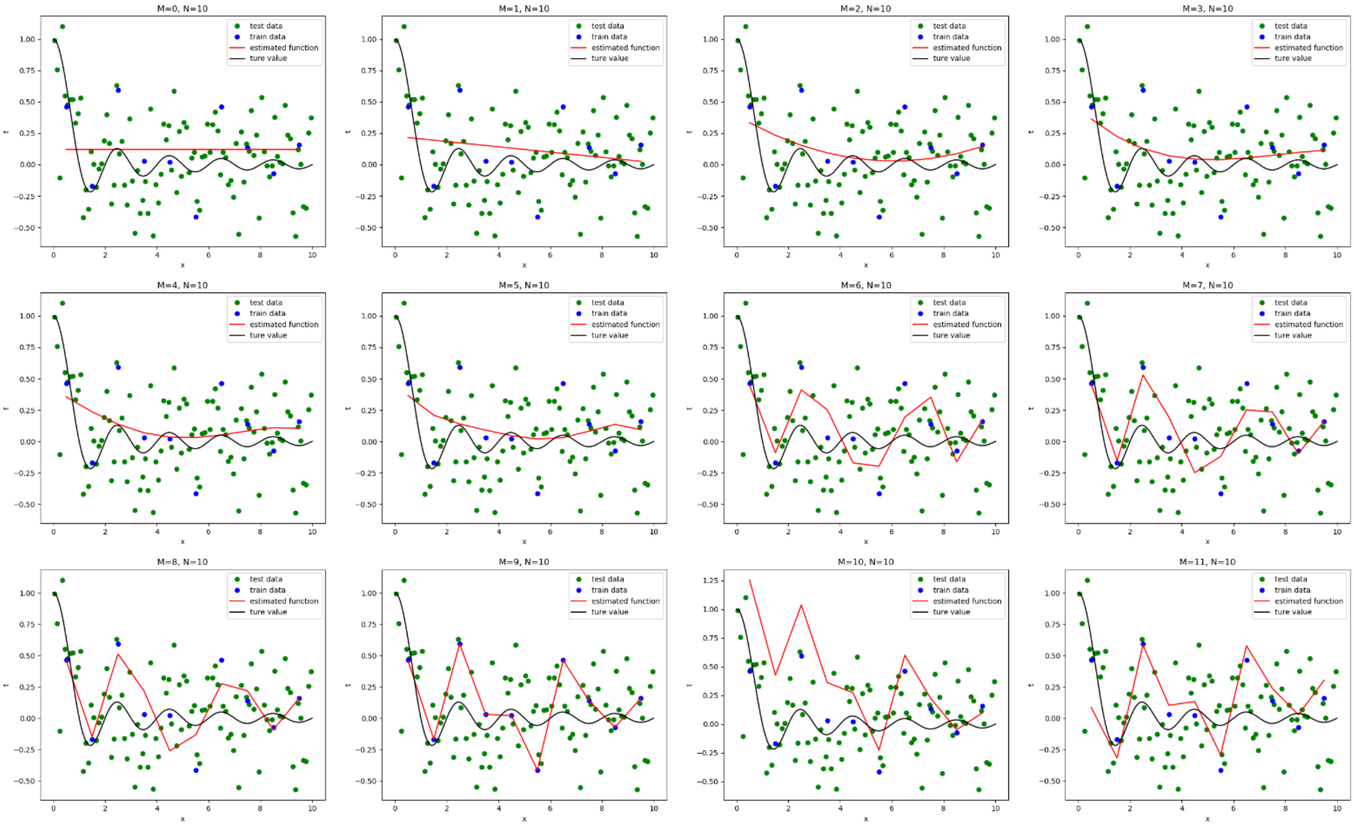

Figure 2

According to Figure 1, we notice that when M=9, Root Mean Square Error of test data Increases dramatically. However, that of training data remains decreasing. That's a obvious diverge point. Then we look at Figure 2 to verify our assumption. When M=9, the estimated function goes through every training data exactly yet it doesn't go through test data. Then we can conclude that overfitting occurs. So the mode order, M, should be less than 9, and 5 or 8 may be the best choice.

Question 2:

Here $\mu_{MAP} = \mu_N$, we get it through calculation in class. And I use μ_{MAP} below.

*1:

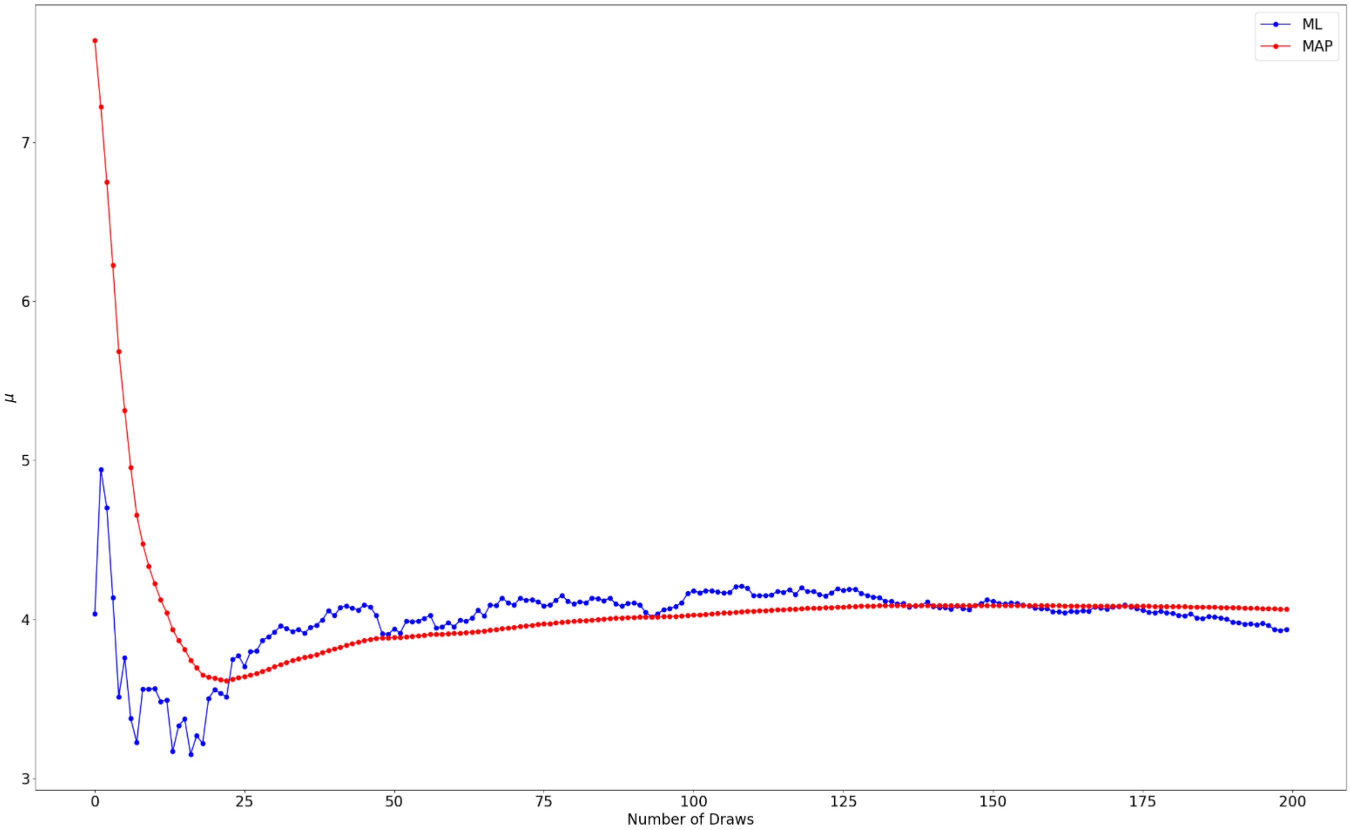

Figure 3

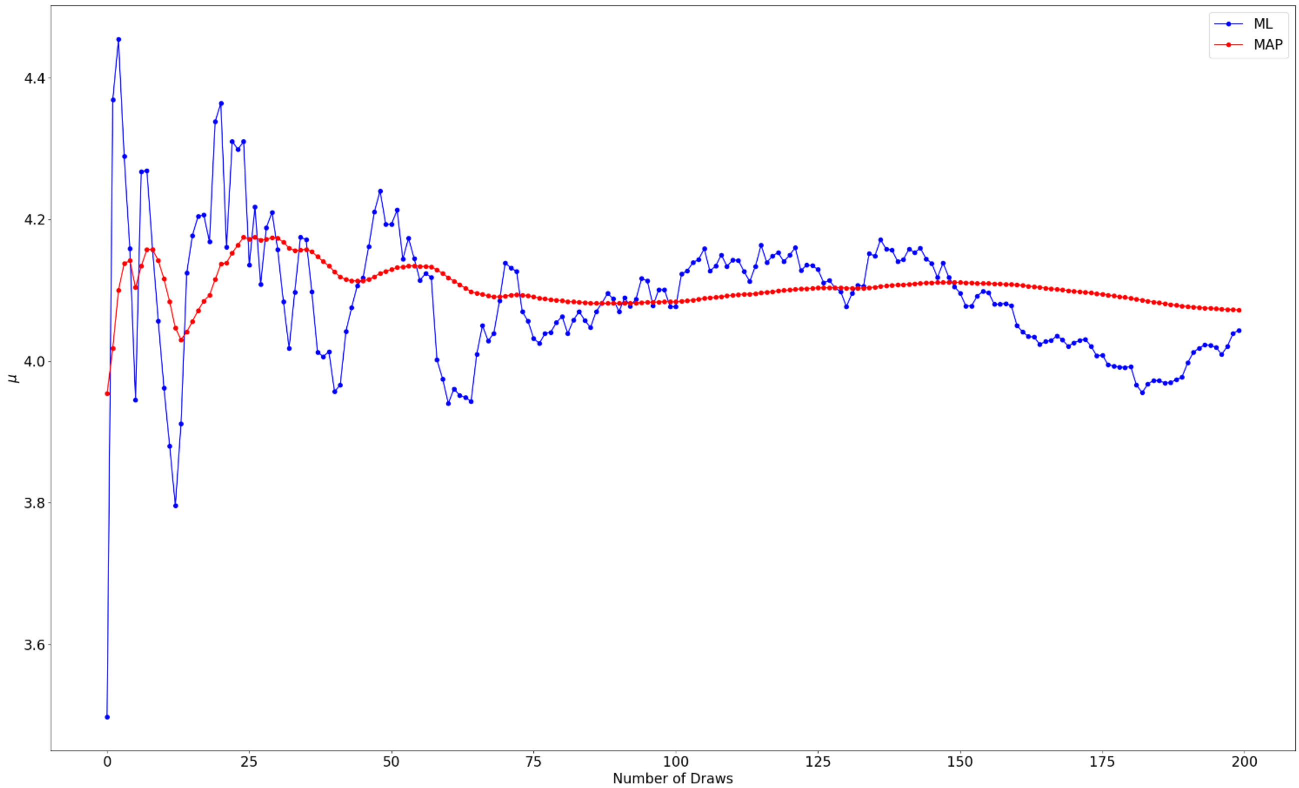

Figure 4

Here true mean=4, wrong prior mean=8.

As Figure 3 shows, if prior mean is initialized to the wrong value, μ_{MAP} will begin at the wrong prior mean and eventually get the true mean. And when prior mean is initialized to the correct value, μ_{MAP} will change around true mean. Meanwhile, no matter prior mean is true or wrong, ML always tends to true mean gradually.

*2:

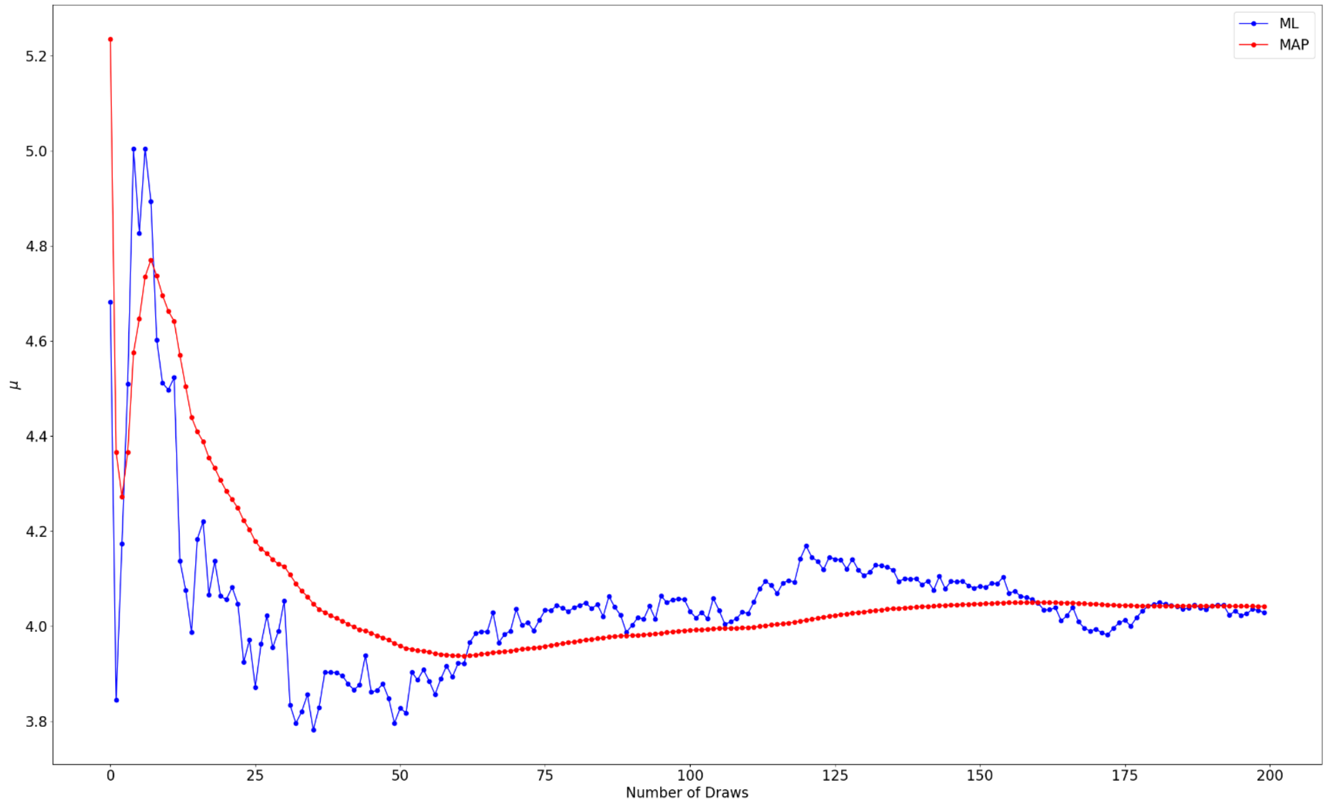

Figure 5

In Figure 5, prior variance=10 compared to 0.2 in Figure 3. We find it takes about 75 draws when μ_{MAP} close to true mean in Figure 5, while in Figure 3, the number is much smaller, about 25. So a large variance leads to a large amount of draws to get relative correct outcome.

*3:

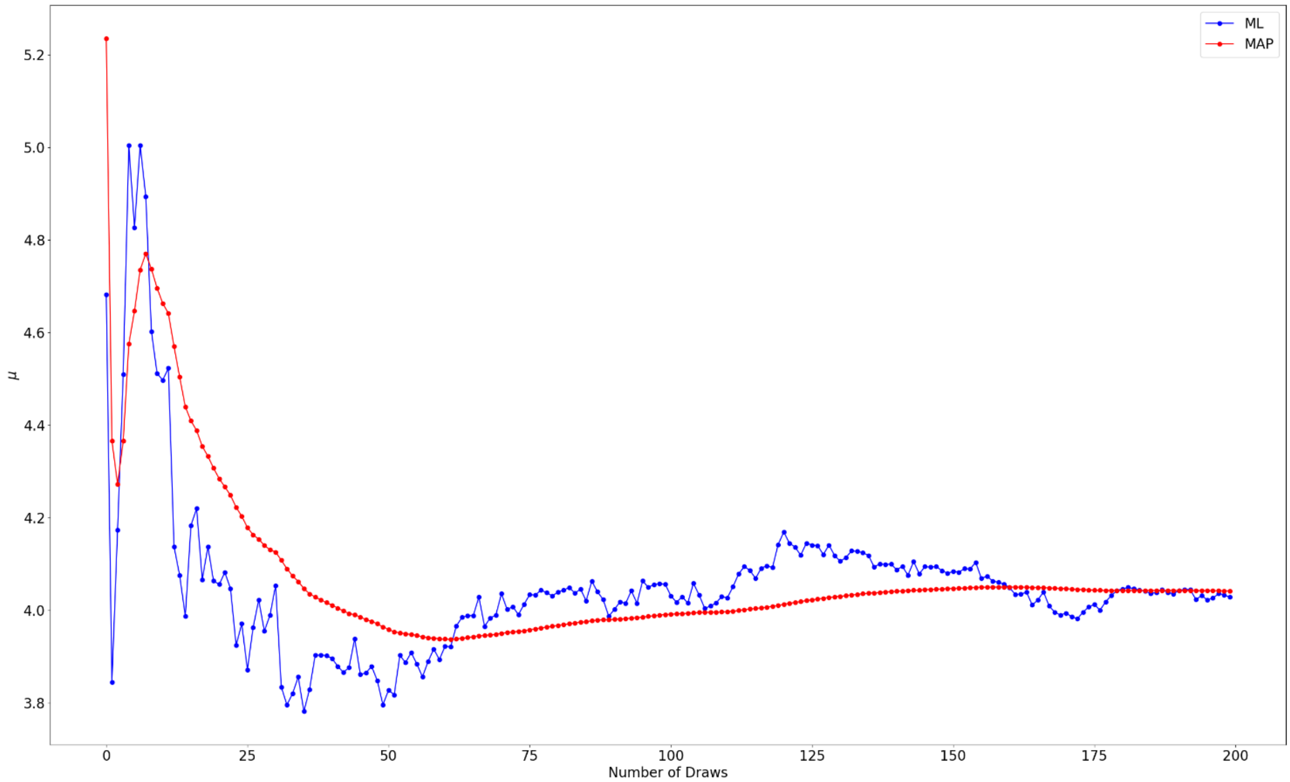

Figure 6

In Figure 6, we change likelihood variance to a large value compared to Figure 3. It is obvious that we need larger number of draws to get relative stable mean of ML. Also, the number of draws increases for μ_{MAP} to get a stable mean.

*4:

Prior mean and variance are our assumed value of true function, if prior mean is wrong, we may get wrong mean at the beginning but as number of draws increases, the weight of prior mean decreases, so we eventually get true mean. Variance reflects the density of data, if variance is small, no matter prior or likelihood, we can get relatively accurate mean with less draws, and if variance is large, the number of draws increases. Meanwhile, the variance of likelihood can change the number of draws to get true mean not only for μ_{ML} but also for μ_{MAP} . But the variance of prior just influences μ_{MAP}

Question 3-6:

Q_3 $f(\vec{x}) = 3\vec{x}^T\vec{x} + 4\vec{y}^T\vec{x} - 1$

$= 3[x_1, x_2 \cdots x_d][x_1, x_2 \cdots x_d]^T + 4[y_1, y_2 \cdots y_d][x_1, x_2 \cdots x_d]^T - 1$

$\frac{\partial f(\vec{x})}{\partial \vec{x}} = [\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2} \cdots \frac{\partial f}{\partial x_d}]^T$

$= \begin{bmatrix} 3([1,0\cdots 0][x_1,x_2\cdots x_d]^T + [x_1\cdots x_d][1,\cdots 0]^T) + 4[y_1\cdots y_d][1,0\cdots 0]^T \\ 3([0,1,\cdots 0][x_1,x_2\cdots x_d]^T + [x_1\cdots x_d][0,1\cdots 0]^T) + 4[y_1\cdots y_d][0,1\cdots 0]^T \\ \vdots \\ 3([0,0,\cdots 1][x_1\cdots x_d]^T + [x_1\cdots x_d][0,\cdots 1]^T) + 4[y_1\cdots y_d][0,\cdots 1]^T \end{bmatrix}$

$= \begin{bmatrix} 3(x_1+x_1) + 4y_1 \\ 3(x_2+x_2) + 4y_2 \\ \vdots \\ 3(x_d+x_d) + 4y_d \end{bmatrix} = \cancel{6x^T + 4y^T} = 6x + 4y$

Q_4 $f(\vec{x}) = -10\vec{x}^TQ\vec{x} + 4\vec{y}^T\vec{x} + 2$ $\quad \frac{\partial 4\vec{y}^T\vec{x}}{\partial \vec{x}} = 4\vec{y}$ we get in Q_3

$\frac{\partial f}{\partial \vec{x}} = [\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2} \cdots \frac{\partial f}{\partial x_d}]^T$

$= \begin{bmatrix} -10([1,0,\cdots 0]\begin{bmatrix} a_{11} \cdots a_{1d} \\ a_{d1} \cdots a_{dd}\end{bmatrix}[x_1,x_2\cdots x_d]^T + [x_1\cdots x_d]\begin{bmatrix} a_{11} \cdots a_{1d} \\ a_{d1} \cdots a_{dd}\end{bmatrix}[1,0,\cdots 0]^T) \\ -10([0,1,\cdots 0]\begin{bmatrix} a_{11} \cdots a_{1d} \\ a_{d1} \cdots a_{dd}\end{bmatrix}[x_1,x_2\cdots x_d]^T + [x_1\cdots x_d]\begin{bmatrix} a_{11} \cdots a_{1d} \\ a_{d1} \cdots a_{dd}\end{bmatrix}[0,1,\cdots 0]^T) \\ \vdots \\ -10([0,0,\cdots 1]\begin{bmatrix} a_{11} \cdots a_{1d} \\ a_{d1} \cdots a_{dd}\end{bmatrix}[x_1,x\cdots x_d]^T + [x_1\cdots x_d]\begin{bmatrix} a_{11} \cdots a_{1d} \\ a_{d1} \cdots a_{dd}\end{bmatrix}[0,\cdots 0,1]^T) \end{bmatrix} + 4\vec{y}$

$= \begin{bmatrix} -10([a_{11}\cdots a_{1d}][x_1\cdots x_d]^T + [x_1\cdots x_d][a_{11}\cdots a_{d1}]^T) \\ \vdots \\ -10([a_{d1}\cdots a_{dd}][x_1\cdots x_d]^T + [x_1\cdots x_d][a_{1d}\cdots a_{dd}]^T) \end{bmatrix} + 4\vec{y}$

$\cancel{-10(Q^Tx + x^TQ^T) + 10}$

$\cancel{-10(Qx + x^TQ^T) + 4y^T}$

Because Q is symmetric $[a_{11}\cdots a_{d1}] = [a_{11}\cdots a_{1d}]$

$\frac{\partial f}{\partial \vec{x}} = \begin{bmatrix} -10([a_{11}\cdots a_{1d}][x_1\cdots x_d]^T + [x_1\cdots x_d][a_{11}\cdots a_{1d}]^T) \\ \vdots \\ -10([a_{d1}\cdots a_{dd}][x_1\cdots x_d] + [x_1\cdots x_d][a_{d1}\cdots a_{dd}]) \end{bmatrix} + 4\vec{y}$

$= \begin{bmatrix} -10 \times 2(a_{11}x_1 + \cdots a_{1d}x_d) \\ \vdots \\ -10 \times 2(a_{d1}x_1 + \cdots a_{dd}x_d) \end{bmatrix} + 4y^T = -20\vec{Q}\vec{x} + 4\vec{y}$

Q5 1. $f(x) = 8\vec{x}^T\vec{Q}\vec{x} - 2\vec{y}^T\vec{Q}^T\vec{x} + 6$

$$\frac{\partial f}{\partial \vec{x}^T} = [\frac{\partial f}{\partial \vec{x}}]^T$$

According to the outcome of Q3, Q4

$$\frac{\partial 8\vec{x}^T\vec{Q}\vec{x}}{\partial \vec{x}} = 16\vec{Q}\vec{x} \qquad \frac{2\vec{y}^T\vec{Q}^T\vec{x}}{\partial \vec{x}} = 2\vec{Q}\vec{y}$$

$$\frac{\partial f}{\partial \vec{x}^T} = [16\vec{Q}\vec{x} - 2\vec{Q}\vec{y}]^T = 16\vec{x}^T\vec{Q}^T - 2\vec{y}^T\vec{Q}^T$$

2. $$\frac{\partial f}{\partial Q} = \begin{bmatrix} \frac{\partial f}{\partial Q_{11}} & \cdots & \frac{\partial f}{\partial Q_{1d}} \\ \vdots & & \\ \frac{\partial f}{\partial Q_{d1}} & \cdots & \frac{\partial f}{\partial Q_{dd}} \end{bmatrix}$$ Because $\vec{Q}$ is symmetric $\vec{Q} = \vec{Q}^T$

$$= \begin{bmatrix} 8\vec{x}^T\begin{bmatrix} 1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{bmatrix}\vec{x} - 2\vec{y}^T\begin{bmatrix} 1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{bmatrix}\vec{x} & \cdots & 8\vec{x}^T\begin{bmatrix} 0 & \cdots & 1 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{bmatrix}\vec{x} - 2\vec{y}\begin{bmatrix} 0 & \cdots & 1 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{bmatrix}\vec{x} \\ \vdots & & \vdots \\ 8\vec{x}^T\begin{bmatrix} 0 & \cdots & 0 \\ \vdots & & \vdots \\ 1 & \cdots & 0 \end{bmatrix}\vec{x} - 2\vec{y}\begin{bmatrix} 0 & \cdots & 0 \\ \vdots & & \vdots \\ 1 & \cdots & 0 \end{bmatrix}\vec{x} & \cdots & 8\vec{x}^T\begin{bmatrix} 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 1 \end{bmatrix}\vec{x} - 2\vec{y}^T\begin{bmatrix} 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 1 \end{bmatrix}\vec{x} \end{bmatrix}$$

$$= \begin{bmatrix} 8x_1^2 - 2y_1x_1 & \cdots & 8x_1x_d - 2y_1\cancel{x}_d \\ \vdots & & \\ 8x_dx_1 - 2y_dx_1 & \cdots & 8x_d^2 - 2y_dx_d \end{bmatrix}$$

$$= 8\vec{x}^T\vec{x} - 2\vec{y}^T\vec{x}$$

Q6 $f(x) = ||4\vec{x}||_2^2$

$= (4\vec{x})^T 4\vec{x}$

$= 16\vec{x}^T\vec{x}$

According to the outcome of Q3

$$\frac{\partial f}{\partial \vec{x}} = 32\vec{x}$$