1a.

Given that $y = w_0 + w_1x_1 + w_2x_2 + \in$

This equation is of the form

$$y = f^*(X) + \in$$

y follows Normal Distribution of the form $y \sim N(XW^*, \sigma^2 I)$

The likelihood probability will follow Gaussian distribution and hence

$$P(y / x_1, x_2) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}(\frac{y - WX}{\sigma})^2}$$

$$P(y / x_1, x_2) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}(\frac{y - (w_0 + w_1x_1 + w_2x_2)}{\sigma})^2}$$

1b.

The Likelihood equation is given by

$$L(Y / X,W) = \prod_{i=1}^{n} P(y^{(i)} / x^{(i)}, w^{(i)})$$

Using the equation of $P(y^{(i)} / x^{(i)}, w^{(i)})$ from 1a we have the equation for likelihood as

$$L(Y / X,W) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}(\frac{y^{(i)} - (w_0 + w_1x_1^{(i)} + w_2x_2^{(i)})}{\sigma})^2}$$

Log Likelihood is given by

$$Log(L(Y / X,W)) = Log(\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}(\frac{y^{(i)} - (w_0 + w_1x_1^{(i)} + w_2x_2^{(i)})}{\sigma})^2})$$

We know the log(ab) = log(a) + log(b) and $\log e^x = x$ using these two we have

$$Log(L(Y / X,W)) = \sum_{i=1}^{n} log(\frac{1}{\sigma\sqrt{2\pi}}) + \frac{-1}{2}(\frac{y^{(i)} - (w_0 + w_1x_1^{(i)} + w_2x_2^{(i)})}{\sigma})^2$$

Ignoring the term $\frac{1}{\sigma\sqrt{2\pi}}$ and $\frac{1}{2\sigma^2}$ since they are independent of w_0, w_1, w_2 we will have

$$Log(L(Y / X,W)) = \sum_{i=1}^{n} -((y^{(i)} - (w_0 + w_1x_1^{(i)} + w_2x_2^{(i)}))^2$$

Conditional Log Likelihood is given by the equation $\sum_{i=1}^{n} -((y^{(i)} - (w_0 + w_1x_1^{(i)} + w_2x_2^{(i)}))^2$

1c.

The MLE is given by the equation

$$W^*_{(MLE)} = \underbrace{argmax}_{W} (\sum_{i=1}^{n} -((y^{(i)} - (w_0 + w_1x_1^{(i)} + w_2x_2^{(i)}))^2)$$

Since we have a - for the equation The max of $\sum_{i=1}^{n} -((y^{(i)} - (w_0 + w_1x_1^{(i)} + w_2x_2^{(i)}))^2$ will be minimum of $\sum_{i=1}^{n} ((y^{(i)} - (w_0 + w_1x_1^{(i)} + w_2x_2^{(i)}))^2$

Multiplying it by 1/2 we have minimum of $\sum_{i=1}^{n}(y^{(i)} - (w_0 + w_1x_1^{(i)} + w_2x_2^{(i)}))^2$ is equal to minimum of $\frac{1}{2}\sum_{i=1}^{n}((y^{(i)} - (w_0 + w_1x_1^{(i)} + w_2x_2^{(i)}))^2$ This is implying that maximizing the likelihood is equal to minimizing the Least Square error.

2a.
The Least Square Error is given by the equation.

LSE = $\frac{1}{2}\sum_{i=1}^{n}\{y^{(i)} - (w_0 + w_1x_1^{(i)} + w_2x_2^{(i)})\}^2 + \frac{\lambda}{2}|||[w1, w2]|||_2^2$

Partial Derivative with respect to w_i is given by

$\frac{\partial(LSE)}{\partial w_j} = \frac{\partial}{\partial w_j}(\frac{1}{2}\sum_{i=1}^{n}\{y^{(i)} - (w_0 + w_1x_1^{(i)} + w_2x_2^{(i)})\}^2 + \frac{\lambda}{2}|||[w1, w2]|||_2^2)$

Since $\frac{\partial}{\partial x}(a - bx)^2$ = 2(a-bx) * $\frac{\partial}{\partial x}$(a-bx) = 2(a-bx) * -b , since $\frac{\partial}{\partial x}$(a-bx) = b

$\frac{\partial}{\partial w_j}(\frac{1}{2}\sum_{i=1}^{n}\{y^{(i)} - (w_0 + w_1x_1^{(i)} + w_2x_2^{(i)})\}^2) = \frac{2}{2}\sum_{i=1}^{n}\{y^{(i)} - (w_0 + w_1x_1^{(i)} + w_2x_2^{(i)})\}$ *-$x_j^{(i)}$

Since $\frac{\partial}{\partial x}x^2 + y^2$ = 2x

$\frac{\partial}{\partial w_j}(\frac{\lambda}{2}|||[w1, w2]|||_2^2) = \frac{\partial}{\partial w_j}(\frac{\lambda}{2}(w_1^2 + w_2^2)) = \frac{2\lambda}{2}(w_j) = \lambda w_j$

Since w_0 is not present here $\frac{\partial}{\partial w_0}(\frac{\lambda}{2}|||[w1, w2]|||_2^2)$ = 0

Using all the above equations obtained

$\frac{\partial}{\partial w_0}(\frac{1}{2}\sum_{i=1}^{n}\{y^{(i)} - (w_0 + w_1x_1^{(i)} + w_2x_2^{(i)})\}^2) = \sum_{i=1}^{n}\{y^{(i)} - (w_0 + w_1x_1^{(i)} + w_2x_2^{(i)})\}$ (-1)

$\frac{\partial}{\partial w_1}(\frac{1}{2}\sum_{i=1}^{n}\{y^{(i)} - (w_0 + w_1x_1^{(i)} + w_2x_2^{(i)})\}^2) = \sum_{i=1}^{n}\{y^{(i)} - (w_0 + w_1x_1^{(i)} + w_2x_2^{(i)})\}$ *$(-x_1^{(i)})$

$\frac{\partial}{\partial w_2}(\frac{1}{2}\sum_{i=1}^{n}\{y^{(i)} - (w_0 + w_1x_1^{(i)} + w_2x_2^{(i)})\}^2) = \sum_{i=1}^{n}\{y^{(i)} - (w_0 + w_1x_1^{(i)} + w_2x_2^{(i)})\}$ *(-$x_2^{(i)}$)

$\frac{\partial}{\partial w_1}(\frac{\lambda}{2}|||[w1, w2]|||_2^2) = \lambda w_1$

$\frac{\partial}{\partial w_2}(\frac{\lambda}{2}|||[w1, w2]|||_2^2) = \lambda w_2$

$\frac{\partial(LSE)}{\partial w_0} = \sum_{i=1}^{n}\{y^{(i)} - (w_0 + w_1x_1^{(i)} + w_2x_2^{(i)})\} + 0 = \sum_{i=1}^{n}\{y^{(i)} - (w_0 + w_1x_1^{(i)} + w_2x_2^{(i)})\}$ (-1)

$\frac{\partial(LSE)}{\partial w_1} = \sum_{i=1}^{n}\{y^{(i)} - (w_0 + w_1x_1^{(i)} + w_2x_2^{(i)})\}$ *$(-x_1^{(i)})$ + λw_1

$\frac{\partial(LSE)}{\partial w_2} = \sum_{i=1}^{n}\{y^{(i)} - (w_0 + w_1x_1^{(i)} + w_2x_2^{(i)})\}$ *$(-x_2^{(i)})$ + λw_2

Gradient Descent update rules is given by the equation

$$w_j^{(t+1)} = w_j^{(t)} - \eta\left(\frac{\partial(LSE)}{\partial w_j}\right)$$

I am considering the cost function = LSE = $\frac{1}{2}\sum_{i=1}^{n}\{y^{(i)} - (w_0 + w_1x_1^{(i)} + w_2x_2^{(i)})\}^2 + \frac{\lambda}{2}||[w1, w2]||_2^2$

Using the equations obtained above we get

$$w_0^{(t+1)} = w_0^{(t)} - \eta(\sum_{i=1}^{n}\{y^{(i)} - (w_0 + w_1x_1^{(i)} + w_2x_2^{(i)})\}\ (-1))$$

$$w_1^{(t+1)} = w_1^{(t)} - \eta(\sum_{i=1}^{n}\{y^{(i)} - (w_0 + w_1x_1^{(i)} + w_2x_2^{(i)})\}\ *(-x_1^{(i)}) + \lambda\, w_1)$$

$$w_2^{(t+1)} = w_2^{(t)} - \eta(\sum_{i=1}^{n}\{y^{(i)} - (w_0 + w_1x_1^{(i)} + w_2x_2^{(i)})\}\ *(-x_2^{(i)}) + \lambda\, w_2)$$

2b.

The MAP estimate is given by the equation

$$W^*_{(MAP)} = \underbrace{argmax}_{W}\ (\log(\prod_{i=1}^{n} P(y^{(i)} / x^{(i)}, w^{(i)})\mathrm{P}(w^{(i)} / \tau)))$$

Taking $P(y^{(i)} / x^{(i)}, w^{(i)})$ from question 1 and we are given that w1,w2 ~ $\mathrm{N}(0, \tau^2\mathrm{I})$ hence equation for

$$\mathrm{P}(w / \tau) = \frac{1}{\tau\sqrt{2\pi}}\, e^{\frac{-1}{2}(\frac{(||w_1, w_2||) - 0)}{\tau})^2}$$, using these equations we have

$$\log(\prod_{i=1}^{n} P(y^{(i)} / x^{(i)}, w^{(i)})\mathrm{P}(w^{(i)} / \tau)) = \log(\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}}\, e^{\frac{-1}{2}(\frac{y^{(i)} - (w_0 + w_1x_1^{(i)} + w_2x_2^{(i)})}{\sigma})^2} * \frac{1}{\tau\sqrt{2\pi}}\, e^{\frac{-1}{2}(\frac{(||w_1, w_2||) - 0)}{\tau})^2})$$

We know the log(ab) = log(a) + log(b) and $\log e^x$ = x using these two we have

$$\log(\prod_{i=1}^{n} P(y^{(i)} / x^{(i)}, w^{(i)})\mathrm{P}(w^{(i)} / \tau)) = \sum_{i=1}^{n} log(\frac{1}{\sigma\sqrt{2\pi}}) + \frac{-1}{2}(\frac{y^{(i)} - (w_0 + w_1x_1^{(i)} + w_2x_2^{(i)})}{\sigma})^2 + \log(\frac{1}{\tau\sqrt{2\pi}}) + \frac{-1}{2}(\frac{(||w_1, w_2||) - 0)}{\tau})^2$$

Neglecting all the terms independent of w, we get

$$W^*_{(MAP)} = \underbrace{argmax}_{W}\ (\sum_{i=1}^{n} \frac{-1}{2}(\frac{y^{(i)} - (w_0 + w_1x_1^{(i)} + w_2x_2^{(i)})}{\sigma})^2 + \frac{-1}{2}(\frac{(||w_1, w_2||) - 0)}{\tau})^2)$$ neglecting the constants σ and τ

$$W^*_{(MAP)} = \underbrace{argmax}_{W}\ (\sum_{i=1}^{n} \frac{-1}{2}(y^{(i)} - (w_0 + w_1x_1^{(i)} + w_2x_2^{(i)}))^2 + \frac{-1}{2}((||w_1, w_2||))^2)$$

Since we have a minus at the start of each terms this is same as

$$W^*_{(MAP)} = \underbrace{argmin}_{W} \left(\sum_{i=1}^{n} \frac{1}{2}(y^{(i)} - (w_0 + w_1 x_1^{(i)} + w_2 x_2^{(i)}))^2 + \frac{1}{2}((||w_1, w_2||))^2 \right)$$

Since σ and τ are constants, If we add the $\frac{\sigma^2}{\tau^2}$ for the second term we get,

$$W^*_{(MAP)} = \underbrace{argmin}_{W} \left(\sum_{i=1}^{n} \frac{1}{2}(y^{(i)} - (w_0 + w_1 x_1^{(i)} + w_2 x_2^{(i)}))^2 + \frac{\sigma^2}{2\tau^2}((||w_1, w_2||))^2 \right)$$

This is equivalent to minimizing the $\frac{1}{2}\sum_{i=1}^{n}\{y^{(i)} - (w_0 + w_1 x_1^{(i)} + w_2 x_2^{(i)})\}^2 + \frac{\lambda}{2}|||[w1, w2]|||_2^2$

Where $\lambda = \frac{\sigma^2}{\tau^2}$

3a.

The code for the same is submitted.

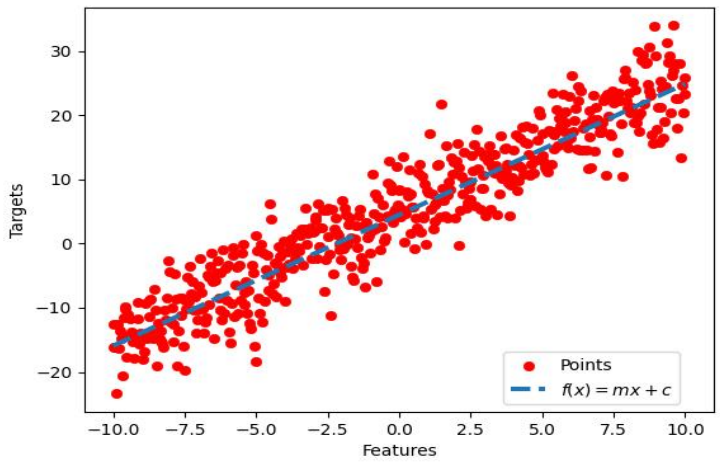
3b.

The code for the same is submitted

3c.

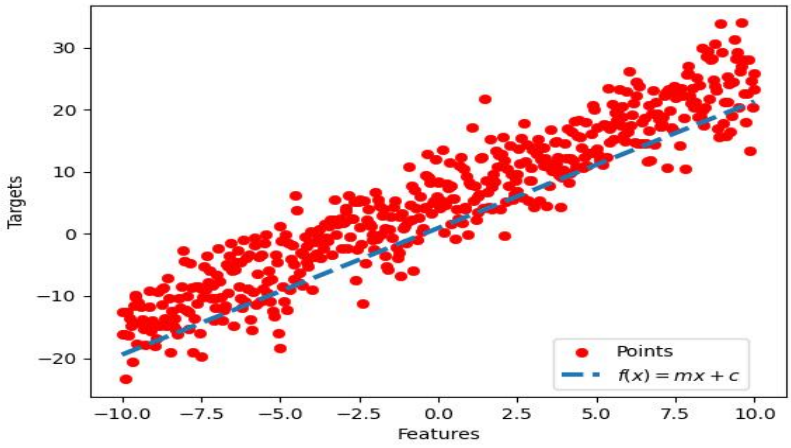
The Plots and data obtained are as follows

For Learning Rate 0.01 :



The value of Loss obtained at the end of 100th epoch is = 17.469

For Learning rate 0.001:



The value of Loss obtained at the end of 100th epoch = 34.815

From the data above we can say that the Learning rate 0.01 is working better compared to 0.001. The Learning rate 0.01 is neither too low nor to high hence within 100 epochs it almost got the optimal values and if we observe the curve obtained for 0.01 it fitted well with the data points and hence MSE is less for 0.01. For the case of 0.001 it is very slow and it actually need more epochs to converge and give the optimal values.