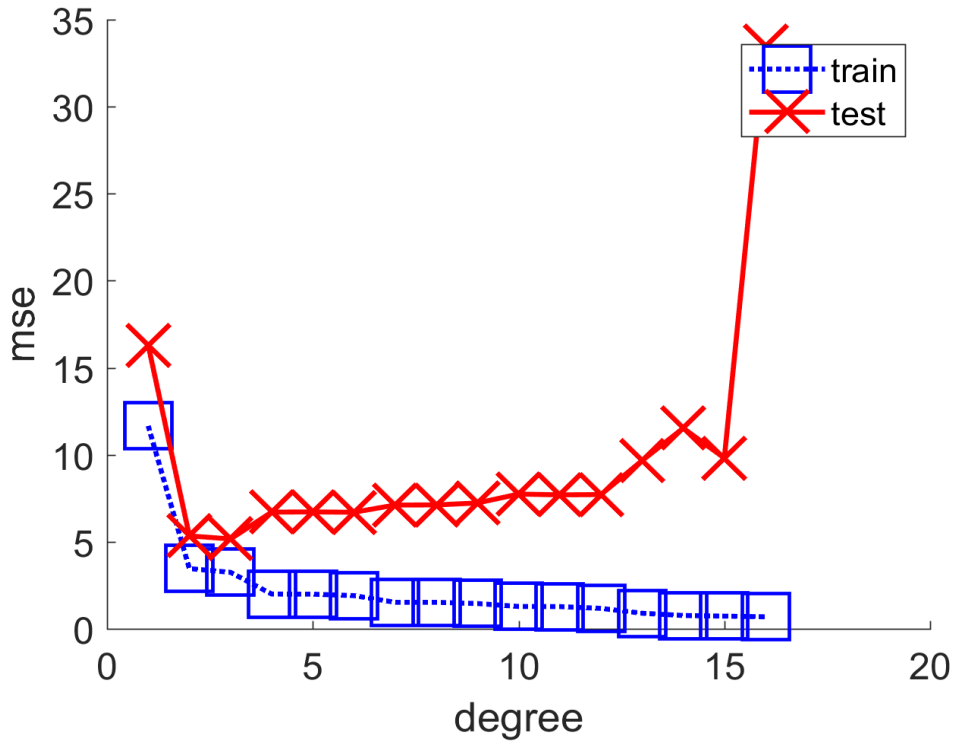
Homework Week 2

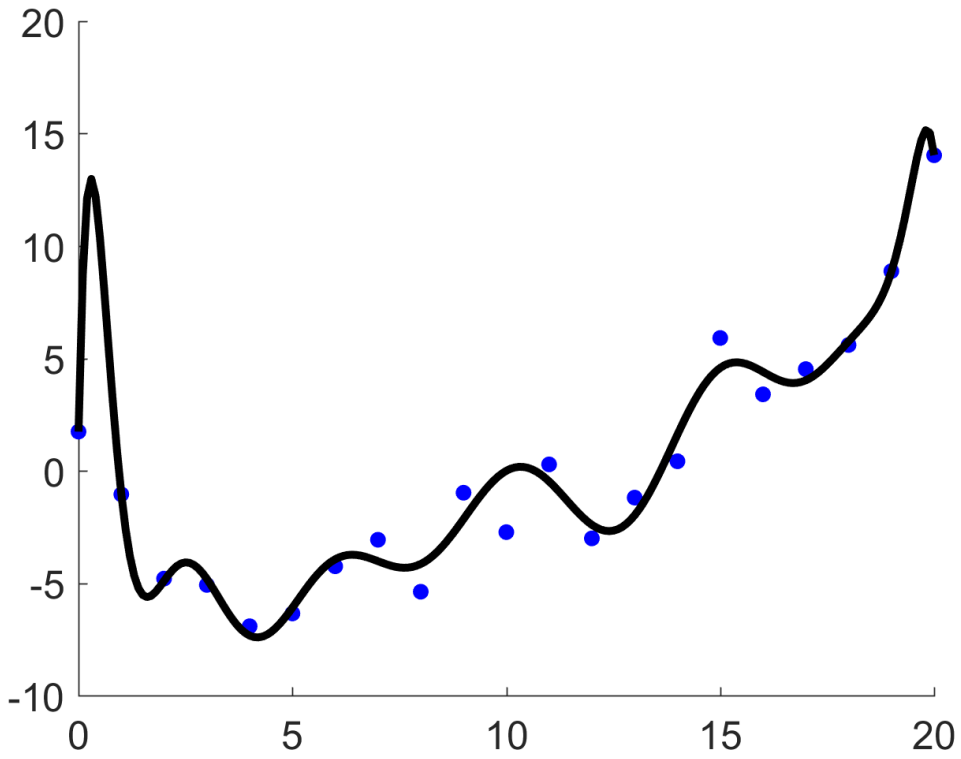
Question 1

b) (i)



(ii) As we can see from the figure above, for the training set, the MSE is initially high when the degree of polynomial is one. But, as the degree of the polynomial increases, the MSE initially decreases. This is because the curve starts fitting the data better. But, for the test set, MSE is high for degree 1, then it decreases for degree 2 and 3 polynomials but thereafter, it increases gradually with the increase in degree. . There is a sharp increase in the MSE for degree 15 and greater polynomials which could be because there is possibly an overfit of the curve to the training data after a certain increase in the polynomial degree.

c) (i)



(ii) The degree of the polynomial is 14. It isn't changing during the demo.

(iii) The variable lambda controls the effect of regularization. Increasing the value of the regularizer makes the curve fit more exactly to the training data point. This means that with an increase in the value of lambda, MSE also increases. This can lead to overfitting on training data point and thus decrease in generalization when dealing with new test data points.

Q2.

a) MSE objective function :

$$= \sum_{i=1}^{N} \left(y_i - [w_0 + w_1 x_i + w_2 x_i^2 + w_3 x_i^3] \right)^2$$

b)

We know, $f(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3$

As $f(x)$ is a 3rd order polynomial.

so, $J(w) = \frac{1}{2} RSS(w) = \frac{1}{2} \sum_{i=1}^{N} (y_i - \underline{w}^T \underline{\phi}_i(x))$

$$\underline{\phi}_i(x) = \begin{bmatrix} 1 \\ x_i \\ x_i^2 \\ x_i^3 \end{bmatrix}$$

b) Given $\underline{\underline{\Phi}}$ is the basis set expansion version of $\underline{\underline{X}}$, the i^{th} row is $\underline{\Phi}^T(x_i)$

$$J(w) = \frac{1}{2} \sum_{i=1}^{N} (y_i - \underline{w}^T \underline{\phi}_i(x)) \qquad \underline{\underline{X}} = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{bmatrix}$$

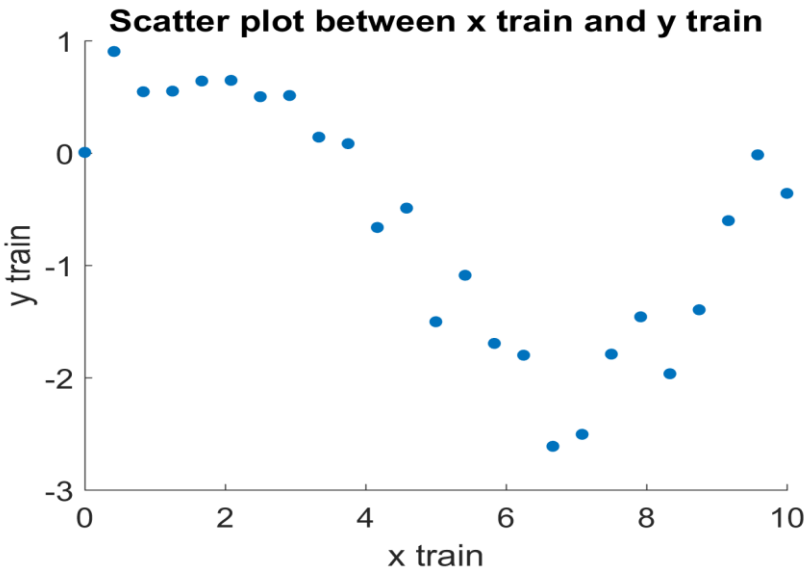
i.e., $\underline{\underline{\Phi}} = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 \\ 1 & x_2 & x_2^2 & x_2^3 \\ \vdots & \vdots & \vdots & \end{bmatrix}$

$$\therefore J(w) = \frac{1}{2} (\underline{y} - \underline{\underline{\Phi}}\, \underline{w})^2$$

$$= \frac{1}{2} (\underline{y} - \underline{\underline{\Phi}}\, \underline{w})(\underline{y} - \underline{\underline{\Phi}}\, \underline{w})^T$$

Question 2

c)



d) The weight vectors are:

Weight_1dim = [0.5579; -0.2349]

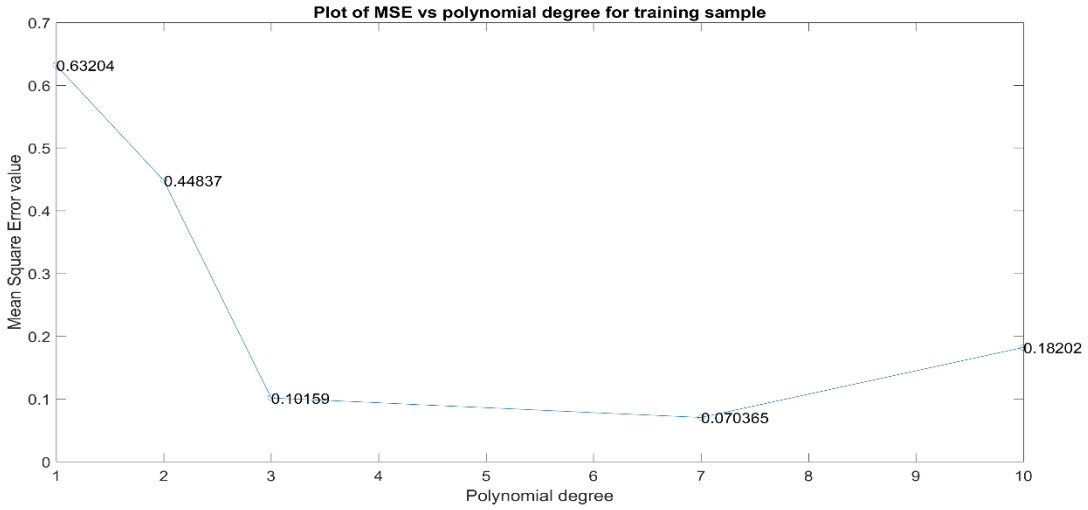
Weight_2dim = [1.4077; -0.7670; 0.0532]

Weight_3dim = [0.1830; 0.8720; -0.3650; 0.0279]

Weight_7dim = [0.2809; 0.3639; 0.1329; -0.1979; 0.0625; -0.0105; 0.0009; -0.0000]

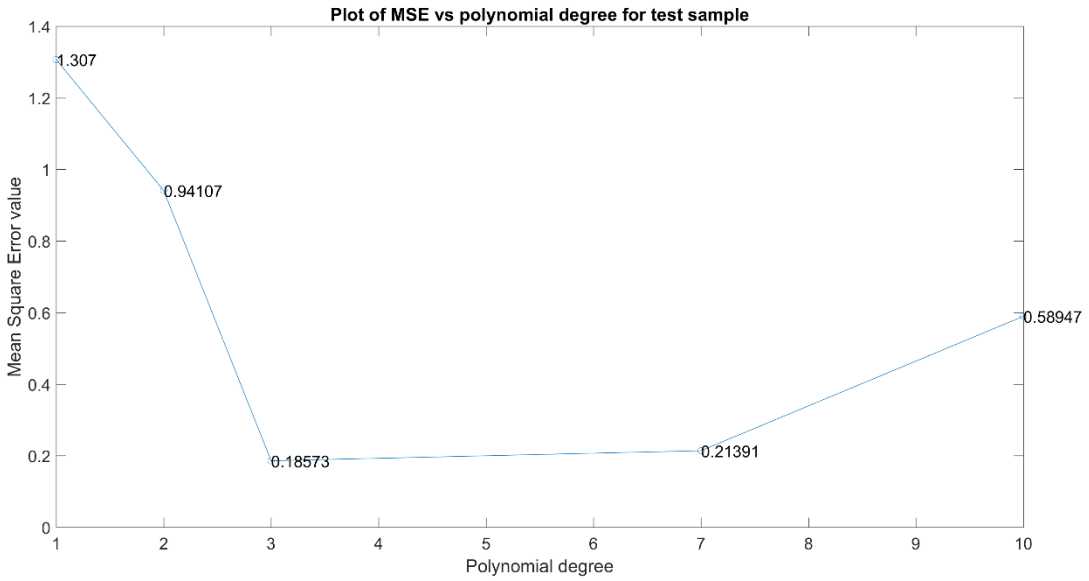
Weight_10dim = 1.0e-03 * [0.0002; 0.0007; 0.0022; 0.0069; 0.0180; 0.0341; 0.0176; -0.1079; 0.0306; -0.0030; 0.0001]

e)



Based on the training sample MSE only, polynomial degree 7 seems to be the best model.

f)



Degrees of polynomial	MSE
1	1.3070
2	0.9411
3	0.1857
7	0.2139
10	0.5895

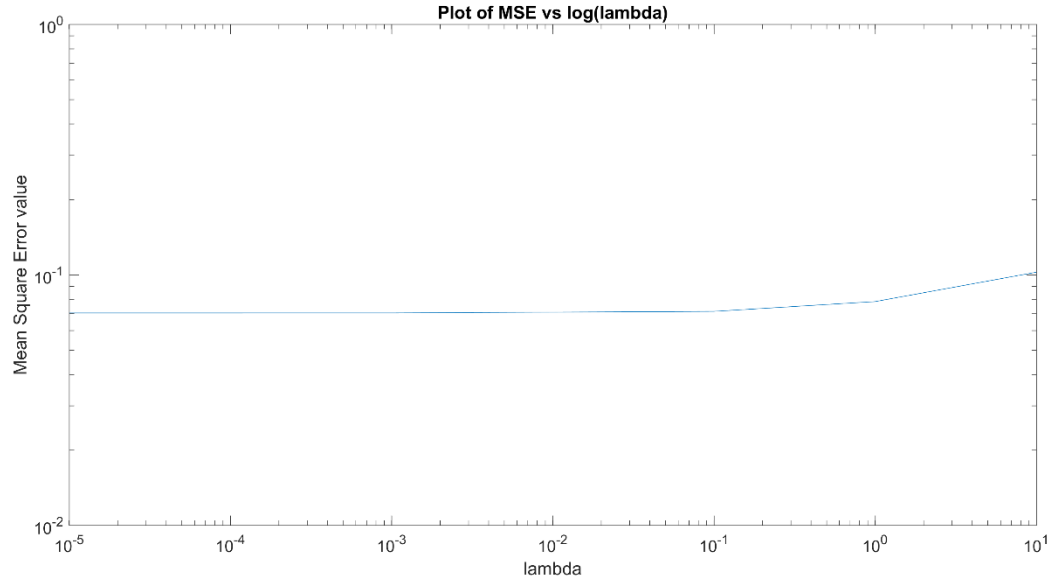
Based on the test sample MSE only, polynomial degree 3 seems to be the best model.

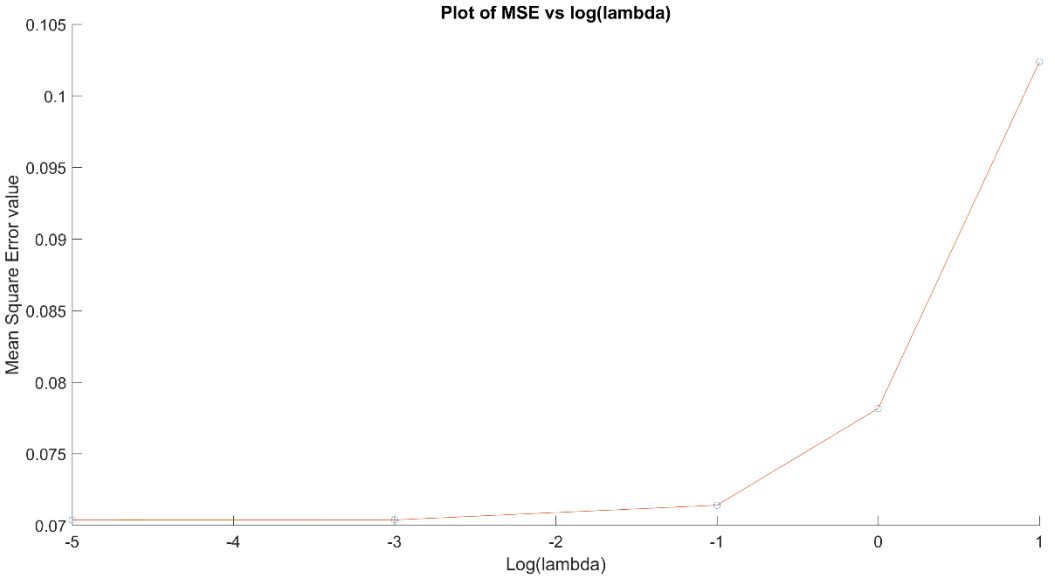
g)

Value of lambda	**Value of weight**
10^-5	[0.2809, 0.3639, 0.1329, -0.1979, 0.0625, -0.0105, 0.0009, 0.0000]
10^-3	[0.2810,0.3636,0.1328,-0.1976,0.0624,-0.01050, 0009,0.0000]
10^-1	[0.2911,0.3350,0.1230,-0.1756,0.0536,-0.0090,0.0008,0.0000]
1	[0.2833, 0.2512, 0.0684, -0.0768, 0.0141, -0.0021, 0.0003,0.0000]
10	[0.1306,0.1073, 0.0777, 0.0203,-0.0310, 0.0060,-0.0004,0.0000]

h)

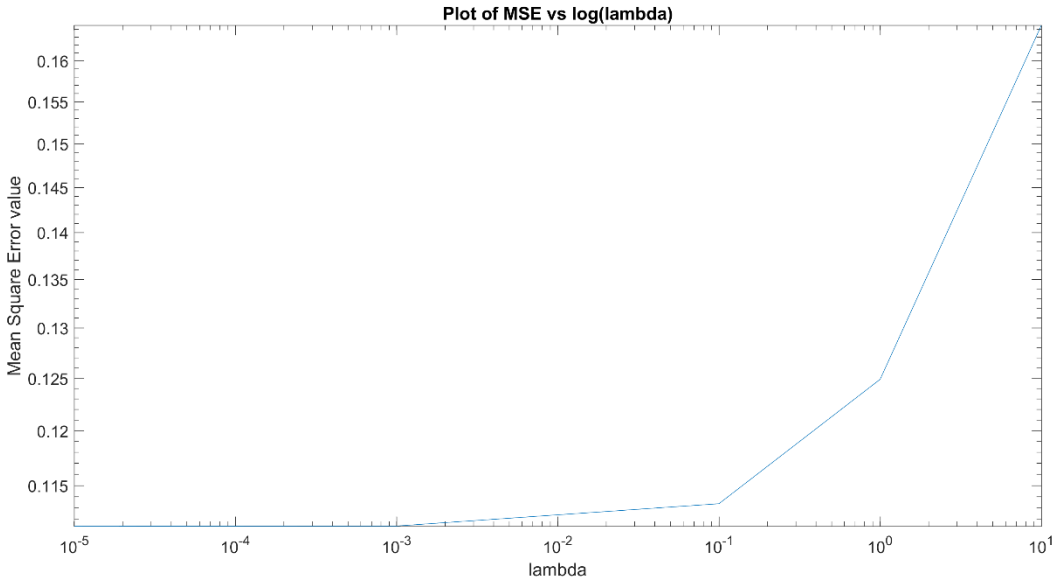
Train MSE = [0.0704, 0.0704, 0.0714, 0.0782, 0.1024]

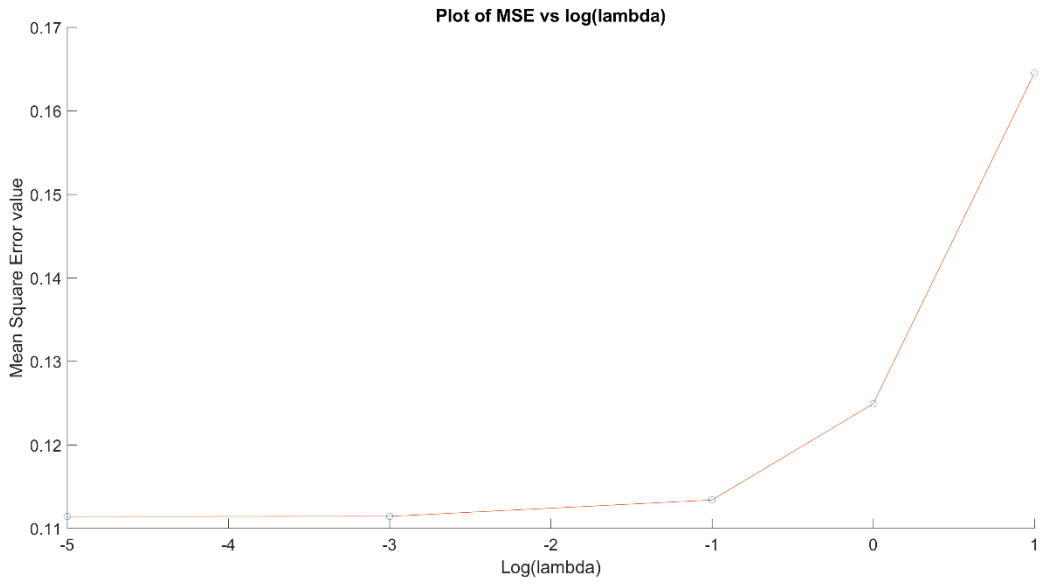




From the Graph, it can be seen that as the value of lambda increases, the MSE increases too for the training data. The change is more rapid when lambda changes from 1 to 10.

Test MSE = [0.1114, 0.1114, 0.1134, 0.1249, 0.1645]





From the Graph, as the value of lambda increases, the MSE increases too for the test data. The change is more rapid when lambda changes from 1 to 10.

93.

$$l(w, \sigma^2) = \frac{-1}{2\sigma^2} RSS(w) - \frac{N}{2} \log(2\pi\sigma^2)$$

This is the full log likelihood for the least squares. To drive the log likelihood with respect to new variable and equate to 0,

$$\frac{\partial l(w, \sigma^2)}{\partial \sigma} = \frac{1}{\sigma^3} RSS(w) - \frac{N}{2} \times \frac{1}{2\pi\sigma^2} (4\pi\sigma)$$

$$= \frac{RSS(w)}{\sigma^3} - \frac{N}{\sigma}$$

$$= 0$$

$$\therefore \quad \frac{RSS(w)}{\sigma^3} - \frac{N}{\sigma} = 0$$

$$\therefore \quad \sigma^2 = \frac{RSS(w)}{N}$$

$$\therefore \quad \sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (y_i - x_i^T \hat{w})^2$$

As solution for $\nabla_w l(w, \sigma^2)$ remains

$$\hat{w} = \text{argmin } RSS(w)$$

Thus, MLE for variance becomes,

$$\hat{\sigma}^2 = \frac{RSS(\hat{w})}{N}$$

Q5. Relation between $p(\underline{w} \mid \underline{\underline{X}}, \underline{y}, \sigma^2)$ and $p(\underline{y} \mid \underline{\underline{X}}, \underline{w}, \sigma^2)$ and prior term can be given by,

We know,

$$\boxed{p(\underline{\theta} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \underline{\theta}) \cdot p(\underline{\theta})}{p(\mathcal{D})}}$$

Here,

$$p(\underline{\theta} \mid \mathcal{D}) = p(\underline{w} \mid \mathcal{D}, \sigma^2) = p(\underline{w} \mid \underline{\underline{X}}, \underline{y}, \sigma^2)$$

So,

$$p(\underline{w} \mid \mathcal{D}, \sigma^2)$$

$$= p(\underline{w} \mid \underline{\underline{X}}, \underline{y}, \sigma^2)$$

$$= \frac{p(\underline{y} \mid \underline{w}, \underline{\underline{X}}, \sigma^2) \cdot p(\underline{w} \mid \underline{\underline{X}}, \sigma^2)}{\int p(\underline{y} \mid \underline{w}, \underline{\underline{X}}, \sigma^2) \cdot p(\underline{w} \mid \underline{\underline{X}}, \sigma^2) d\underline{w}}$$

So,

$$\boxed{p(\underline{w} \mid \underline{\underline{X}}, \underline{y}, \sigma^2) = \frac{p(\underline{y} \mid \underline{w}, \underline{\underline{X}}, \sigma^2) \cdot p(\underline{w} \mid \underline{\underline{X}}, \sigma^2)}{\int p(\underline{y} \mid \underline{w}, \underline{\underline{X}}, \sigma^2) \cdot p(\underline{w} \mid \underline{\underline{X}}, \sigma^2) d\underline{w}}}$$

POSTERIOR TERM

LIKELIHOOD TERM

PRIOR TERM