Project 1 : Curve Fitting

Pramod Kumar

February 5, 2020

Abstract

Curve fitting is process of finding a approx function which can mimic actual function behind the dataset. In other words, its about finding function behind the given dataset. if we know about actual/approx function, we can calculate/predict value of a unknown input. To achieve this we will implement 2 methods i.e Linear regression and Bayesian approach and we will see what are the short coming in each approach and how to rectify them with regularization.

Contents

1 Introduction

Basic idea revolve around finding the parameter values in linear equation (w.r.t W) by using randomly generated data-set, which is called training phase. Later use those calculated W^* values to predict new input. In practice, instead of regression mostly such problem are solved using gradient descent approach as dimension of data in real world is very high. But for the scope of this project, we ill look at 4 different approach, namely:
1) Linear regression using error minimisation
2) Linear regression using error minimisation with regularisation
3) Maximum Likelihood Estimator
4) Maximum posterior estimation

2 Approach

2.1 Data

In this project we will be using 2 set of data with size 50 and 100. This data is generated using *generateData.m* script in the project directory. This script generate data by calculating $\sin x$ of random data and add noise to resemble it with real data.

2.2 Methods

2.2.1 Linear regression using error minimization

Error Minimization (non-regularized)
A linear equation (wrt. W) is defined as

$$y(x, w) = \sum_{i=1}^{M} w_i x^i \tag{1}$$

$$W = [\, w_0 \, w_1 \, ... \, w_M] \quad and \quad X = \begin{bmatrix} x_1^0 & x_1 & \dots \\ \vdots & \ddots & \\ x_N^0 & & x_N^M \end{bmatrix}_{N \times M}$$

$$and \quad T = [t_0 \, t_1 \, ... \, t_N]$$

Error associated with this equation (least square error) can be expressed as follows:

$$E(\vec{w}) = \left(\frac{1}{2}\right) \sum (y(\vec{x_n}) - \vec{t_n})^2.$$

A best solution will have LSE as small as possible. In geometric interpretation, we need to determine value of W such that sum of projection of all points on curved formed by W is minimized.

To minimize error function, we can take differentiate and equate to 0.
Differentiating with respect to w, we get

$$
\begin{aligned}
Matrix\, representation\, of\, y: \quad Y &= XW \\
Error\, function\, in\, matrix\, form: & \\
E(W) &= \frac{1}{2}(XW - T)^2 \\
E(W) &= \frac{1}{2}(XW - T)^T(XW - T) \\
E(W) &= \frac{1}{2}(W^TX^T - T^T)(XW - T) \\
E(w) &= \frac{1}{2}(W^TX^TXW - W^TX^TT - T^TXW + T^TT) \\
(T^TXW)^T &== (W^TX^TT) \quad and\, both\, are\, 1 \times 1\, matrix \\
E(w) &= \frac{1}{2}(W^TX^TXW - 2W^TX^TT + T^TT) \\
\frac{\partial}{\partial W}E(W) &= \frac{\partial}{\partial W}\frac{1}{2}(W^TX^TXW - 2W^TX^TT + T^TT) \\
&= \frac{1}{2}(2X^TXW - 2X^TT + 0)
\end{aligned}
$$

$$= X^TXW - X^TT \tag{2}$$

To minimize equation 2, we need to equate to 0

$$
\begin{aligned}
X^TXW - X^TT &= 0 \\
X^TXW &= X^TT \\
W^* &= (X^TX)^{-1}X^TT
\end{aligned}
$$

2.2.2 Linear regression using error minimization with regularization

Here we will use same sum square error function but we will have another parameter λ, which will help in solving problem of over fitting. A linear equation (wrt. W) is defined as

$$E(\vec{w}) = \left(\frac{1}{2}\right)\sum(y(\vec{x_n}) - \vec{t_n})^2 + \frac{\lambda}{2}\vec{W}^2.$$

Similar to minimizing non-regularized error function, we will minimize this function too.

$$
\begin{aligned}
Matrix\, representation\, of\, y: \quad Y &= XW \\
Error\, function\, in\, matrix\, form: & \\
E(W) &= \frac{1}{2}(X\vec{W} - \vec{T})^2 + \frac{\lambda}{2}\vec{W^2} \\
E(W) &= \frac{1}{2}(X\vec{W} - \vec{T})^T(X\vec{W} - \vec{T}) + \frac{\lambda}{2}W\vec{^T}W \\
E(W) &= \frac{1}{2}(\vec{W}^T\vec{X}^T - \vec{T^T})(X\vec{W} - \vec{T}) + \frac{\lambda}{2}W\vec{^T}W \\
E(w) &= \frac{1}{2}(\vec{W}^T\vec{X}^T\vec{X}\vec{W} - \vec{W}^T\vec{X}^T\vec{T} - \vec{T}^TX\vec{W} + \vec{T}^T\vec{T}) + \frac{\lambda}{2}W\vec{^T}W \\
(T^TXW)^T &== (W^TX^TT) \quad and\, both\, are\, 1 \times 1\, matrix \\
E(w) &= \frac{1}{2}(W^TX^TXW - 2W^TX^TT + T^TT) + \frac{\lambda}{2}W^TW \\
\frac{\partial}{\partial W}E(W) &= \frac{\partial}{\partial W}\frac{1}{2}(W^TX^TXW - 2W^TX^TT + T^TT) + \frac{\partial}{\partial W}\frac{\lambda}{2}W^TW \\
&= \frac{1}{2}(2X^TXW + \lambda 2IW - 2X^TT + 0) \\
&= X^TXW + \lambda IW - X^TT
\end{aligned} \tag{3}
$$

To minimize equation 2, we need to equate to 0

$$
\begin{aligned}
X^TXW + \lambda IW - X^TT &= 0 \\
X^TX + \lambda IW &= X^TT \\
\vec{W^*} &= (\vec{X}^T\vec{X} + \lambda\vec{I})^{-1}\vec{X}^T\vec{T}
\end{aligned}
$$

So minimization function becomes :

$$
E(\vec{W^*}) = \frac{1}{2}(X\vec{W^*} - \vec{T})^2 + \frac{\lambda}{2}\vec{W^{*2}} \tag{4}
$$

2.2.3 Maximum Likelihood Estimation

It is defined as

$$
p\left(W \mid X, t, \beta\right) \propto p\left(t \mid X, W, \beta\right)
$$

$p\left(t \mid X, W, \beta\right)$ is called likelihood and left side of above equation is call postprior. Likelihood defined on N data points as:

$$
p\left(t \mid X, W, \beta\right) = \prod_{n=1}^{N} N\left(t_n \mid y\left(x_n, W\right), \beta^{-1}\right) \tag{5}
$$

Where

$$
N\left(x \mid \mu, \sigma^2\right) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} e^{\left\{-\frac{1}{2\sigma^2}\left(x-\mu^2\right)\right\}}
$$

where $\beta^{-1} = \sigma^2$. Taking log on equation 5, we get

$$ln(p(t|x, w, \beta)) = \left(\frac{-\beta}{2}\right) \sum (y(\vec{x_n}) - \vec{t_n})^2 + \left(\frac{N}{2}\right) ln\beta - \left(\frac{N}{2}\right) ln2\pi. \tag{6}$$

To estimate W, Differentiate equation 6 w.r.t W

$$\frac{1}{p(t|x, w, \beta)} = \left(\frac{-\beta}{2}\right) \frac{\partial \left(E(W)\right)}{\partial W} \tag{7}$$

Calculating maximization of likelihood, is equal to calculating minimization of E(W), which is equal to minimization of Task 1.

$$W^* = (\vec{X}^T \vec{X})^{-1} \vec{X}^T \vec{T}$$

Now optimize for β. Differentiate equation 6 W.r.t β and equate it to 0 for maximizing.

$$\frac{1}{p(t|x, w, \beta)} = \frac{1}{2} \sum_{n=1}^{N} \left(y(x_n, W) - t_n\right)^2 + \frac{N}{2\beta}$$

$$\frac{N}{2\beta} = \frac{1}{2} \sum_{n=1}^{N} \left(y(x_n, W) - t_n\right)^2$$

$$\beta^{-1} = \frac{1}{N} \sum_{n=1}^{N} \left(y(x_n, W) - t_n\right)^2 \tag{8}$$

2.2.4 Maximum Posterior Estimation

$$p\left(W \mid X, t, \alpha, \beta\right) \propto p\left(t \mid X, W, \beta\right) p(W \mid \alpha) \tag{9}$$

[1] In posterior, we need to find optimum (best fit) $\boldsymbol{W}$ for the given data. where

$$p(W \mid \alpha) = \frac{\alpha^{M+1}}{2\pi} e^{\left(\frac{-\alpha}{2} w^T W\right)} \tag{10}$$

[1] and

$$p\left(t \mid X, W, \beta\right) = \prod_{n=1}^{N} N\left(t_n \mid y\left(x_n, W\right), \beta^{-1}\right)$$

[1] Take log :

$$ln\ p\left(W \mid X, t, \alpha, \beta\right) = ln\left(t \mid X, W, \beta\right) + ln\ p(W \mid \alpha)$$

To find minimum, differentiate both side.

$$\ln(p(t|x, w, \beta)) = \left(\frac{-\beta}{2}\right) \sum (y(\vec{x_n}) - \vec{t_n})^2 + \left(\frac{N}{2}\right) \ln \beta - \left(\frac{N}{2}\right) \ln 2\pi + \frac{M+1}{2\pi} \ln \alpha + \left(\frac{-\alpha}{2} w^T W\right) \ln e.$$

In above equation, $\left(\frac{N}{2}\right) ln\ \beta$, $\left(\frac{N}{2}\right) ln 2\pi$ and $\frac{M+1}{2\pi} ln\ \alpha$ are constants W.r.t W. Hence it wont affect our maximization. The remaining terms are as follows

$$\ln(p(t|x,w,\beta)) = \left(\frac{-\beta}{2}\right)\sum(y(\vec{x_n}) - \vec{t_n})^2 + \left(\frac{-\alpha}{2}w^T W\right)\ln e.$$

Differentiate W.r.t $\vec{W}$. and maximize $p(t|x,w,\beta)$, which is equal to minimizing $\frac{1}{p(t|x,w,\beta)}$. Hence, Minimizing equation becomes:

$$E(w) = \left(\frac{-\beta}{2}\right)\sum(y(\vec{x_n}) - \vec{t_n})^2 + \left(\frac{-\alpha}{2}w^T W\right) \tag{11}$$

After solving above equation 11 similar to task 1, we get the following result

$$\vec{W^*_{MAP}} = \left(\vec{X}^T\vec{X} + \frac{\alpha}{\beta}\vec{I}\right)^{-1}\vec{X}^T\vec{T} \tag{12}$$

Now to solve for β^*, since $p(W \mid \alpha)$ is independent of β. we are left with only likelihood function, Hence maximization of β^{-1} posterior is same as Maximization of Likelihood in Task3.

$$\beta^{-1} = \frac{1}{N}\sum\left(\vec{X}^T\vec{W}^*_{MAP} - \vec{T}\right)^2 \tag{13}$$

3 Results

3.1 Linear regression using error minimization

We have 2 parameter to vary, N (Number of observation) and M (Degree of polynomial) M will give more flexibility to fix the data point. with M = 0, Square sum error from all data points will be high. As we increase degree of freedom, curve start fitting better.

At M = 6, curve is almost equal to ground truth. when we increse M to 9, it start over-fitting. Follow below diagram and table for more intuition.

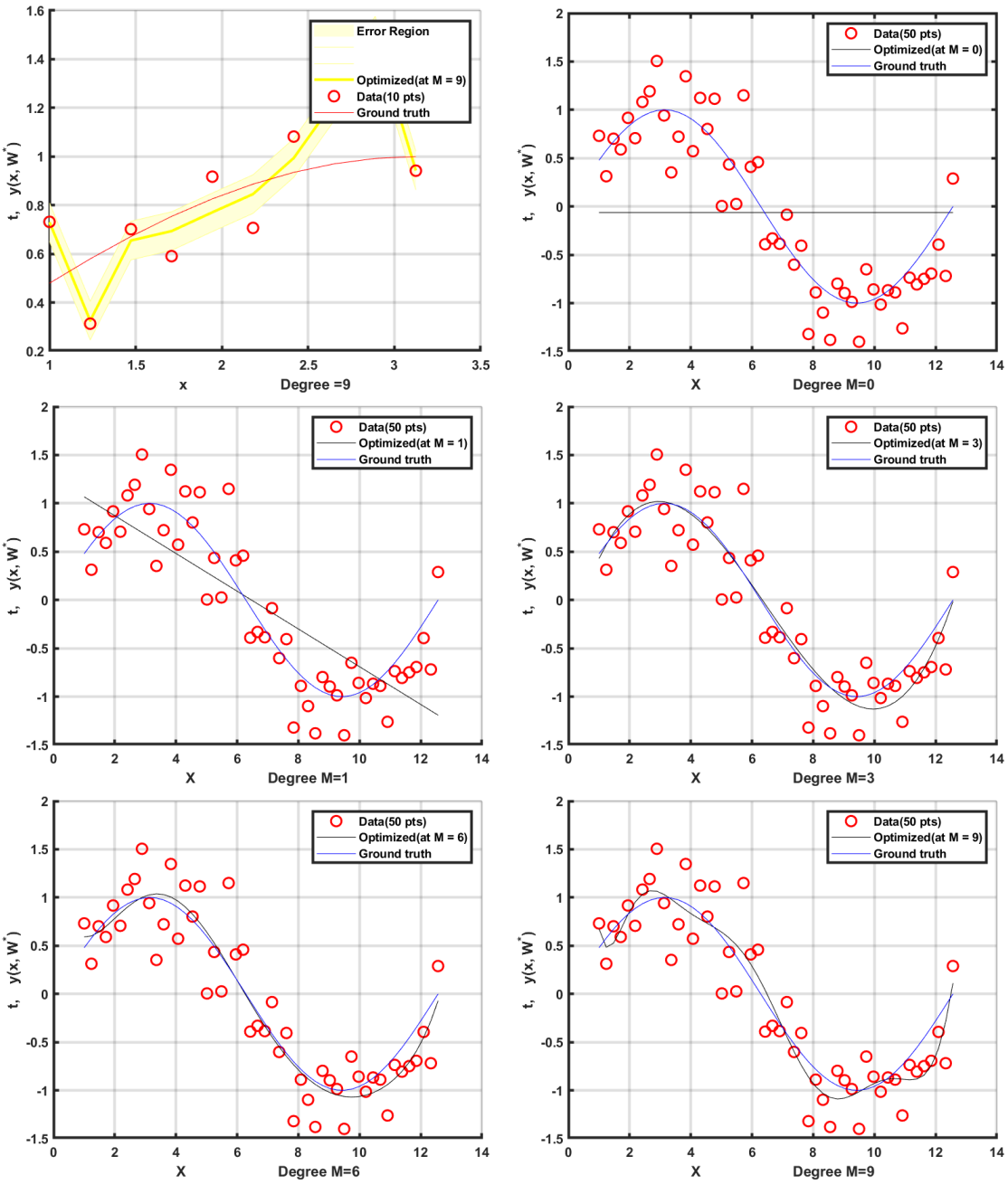

Figure 1: M = [0, 1, 3, 6, 9] with 50 data points

•

W^* value Comparison					
	M = 0	M = 1	M = 3	M = 6	M = 9
w_0^*	-0.0621	1.2625	-0.4407	1.2186	11.0194
w_1^*		-0.1953	1.0991	-1.4575	-26.046
w_2^*			-0.2415	1.1059	25.1370
w_3^*			0.0125	-0.3127	-12.4559
w_4^*				0.0393	3.5895
w_5^*				-0.0023	-0.6346
w_6^*				0.0001	0.0696
w_7^*					-0.0046
w_8^*					0.0002
w_9^*					-0.0000

As we can see in below table, error is reducing along with increase in degree of polynomial

Erms				
M = 0	M = 1	M = 3	M = 6	M = 9
0.8421	0.5163	0.3177	0.3112	0.2929

On increasing N from 10 to 50, Error start getting low. Which means, W gets better and better fit when N value increases.

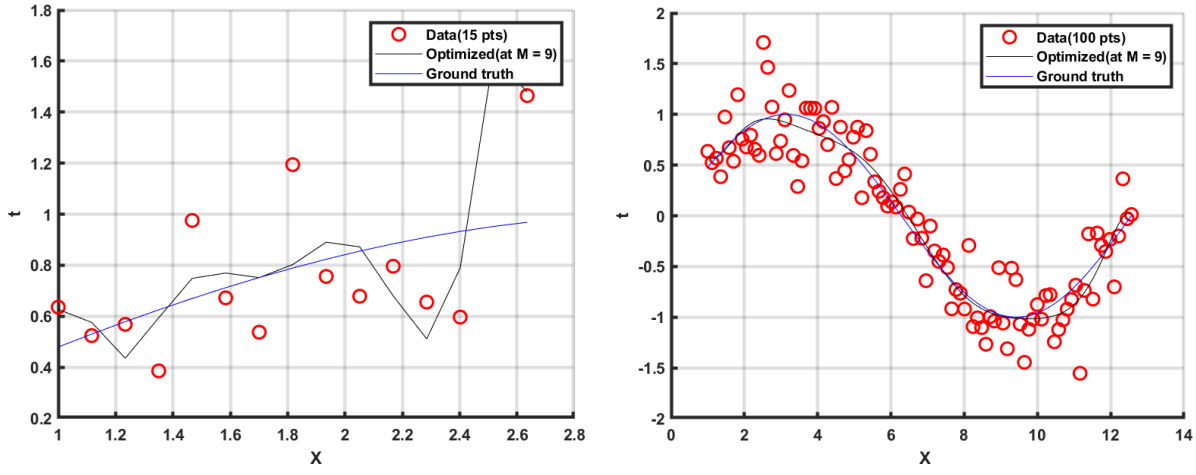

Figure 2: N = [15, 100] with M = 9

•

3.2 Linear regression using error minimization with regularization

Without regularization large weights tries try to overfit when order of equaiton is 9. By adding regularization, we introduce penalty on large weight, which makes the curve more

smooth.
λ here is called hyper-parameter. Changing its value will change the smoothness in the curve fitting.
Below images describe the smoothness in the curve with different values of lambda.

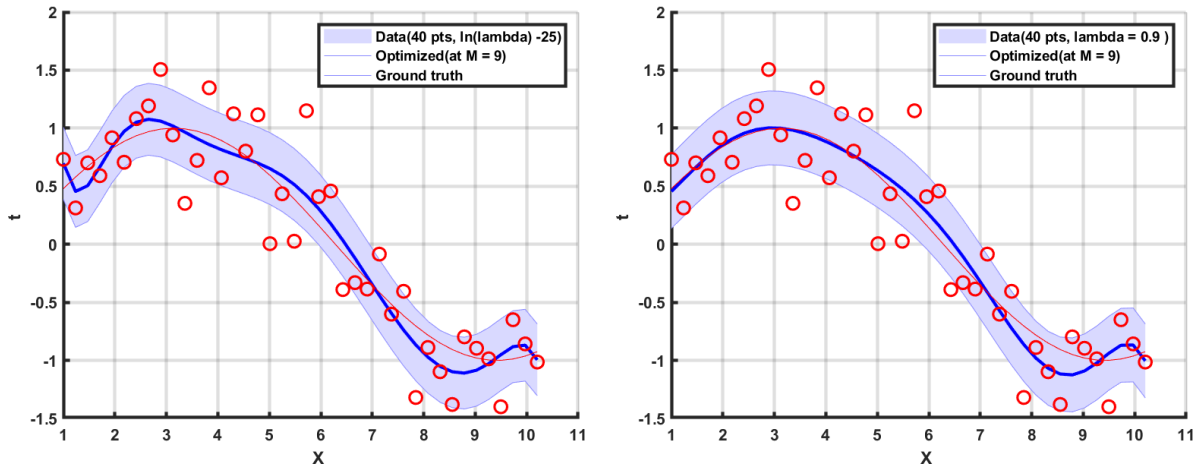

Figure 3: $\lambda = [exp^{-25}, 0.9]$ with M = 9

- Similarly, we can see error rate varying with varying λ

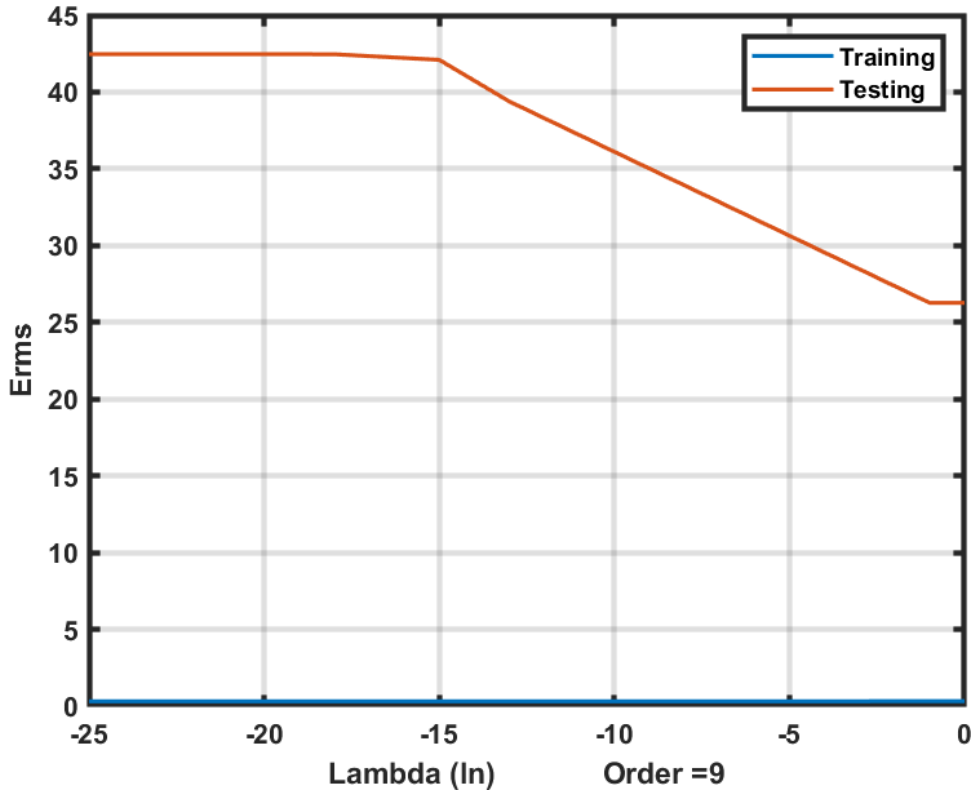

Figure 4: $\lambda = [\lambda = e^{-25}, \lambda = e^{-18}, \lambda = e^{-15}, \lambda = e^{-13}, \lambda = e^{-1}, \lambda = 1]$ with M = 9

Error w.r.t λ						
	$\lambda = e^{-25}$	$\lambda = e^{-18}$	$\lambda = e^{-15}$	$\lambda = e^{-13}$	$\lambda = e^{-1}$	$\lambda = 1$
Train error	0.3093	0.3093	0.3093	0.3096	0.3165	0.3186
Test error	42.4763	42.4650	42.1000	39.3892	26.2731	26.2731

3.3 Maximum Likelihood Estimation

MLE has same results as Task 1, Hence avoiding writing duplication.

Below is figure with 10 data points and we can see over-fitting with degree 9

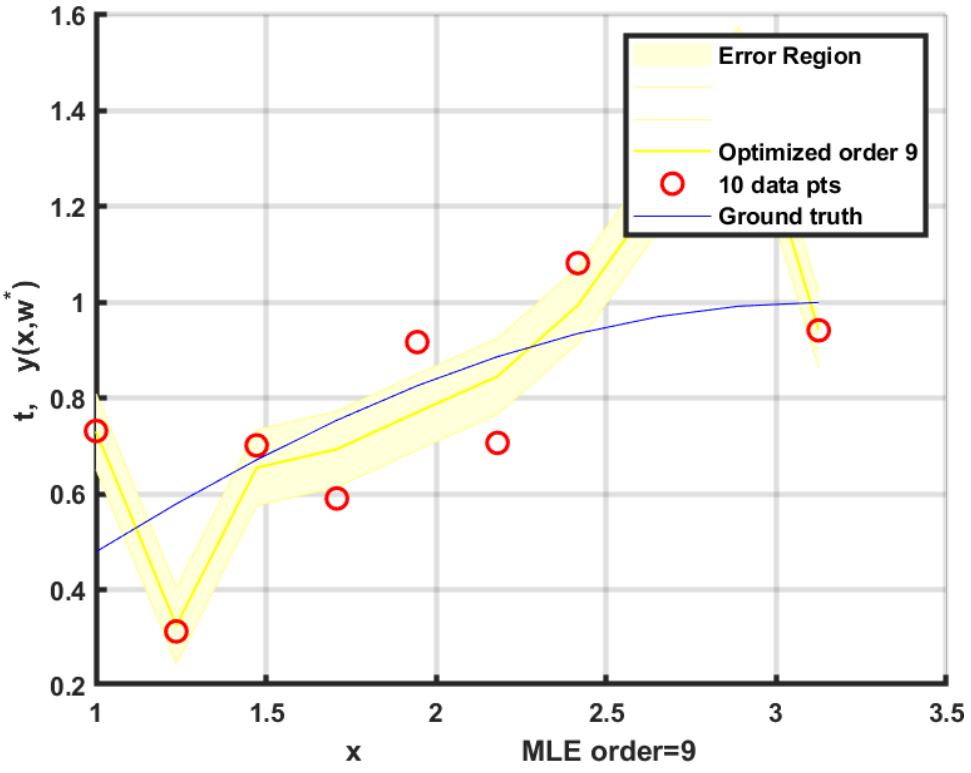

Figure 5: Data: 10pts, M = 9

Comparison between Error minimization and MLE, both looks exactly same.

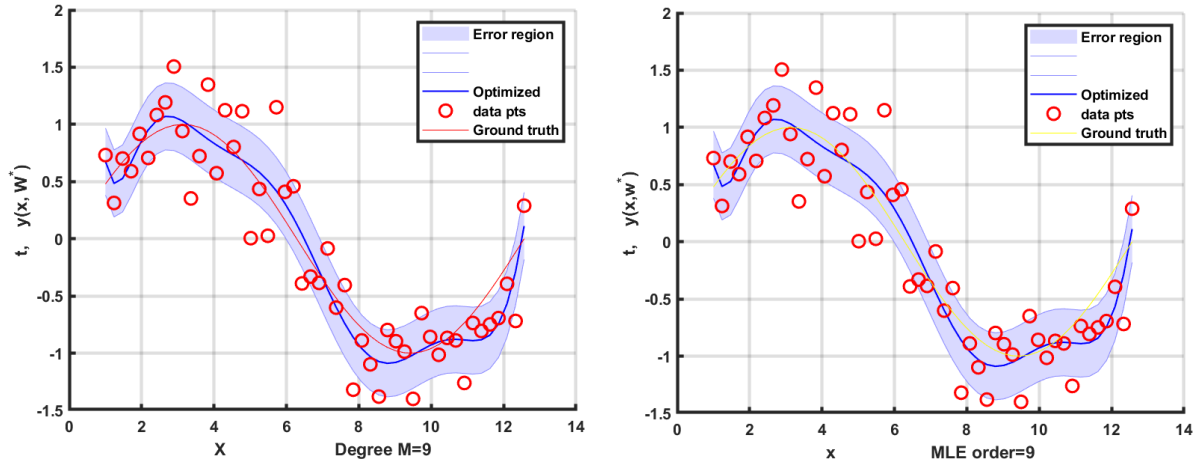

Figure 6: Comparison between EM and MLE with M = 9 , data 50pts

3.4 Maximum Posterior Estimation

Maximum posterior draws its advantages from prior information, which is absent in maximum likelihood. In Posterior, we dont have to add λ, it comes internally in form of (α and β). Most of observation related to variation w.r.t N and M as same as previous 3 tasks, Hence omitted repetition.

Below figure is with 10 data points and error in training data set is about 0.1088, which little better then what we got from various λ values in task 2 (though comparison is not valid as λ and α/β is not equal, but still it is good improvement compared to moving λ from e^{-25} to 1).

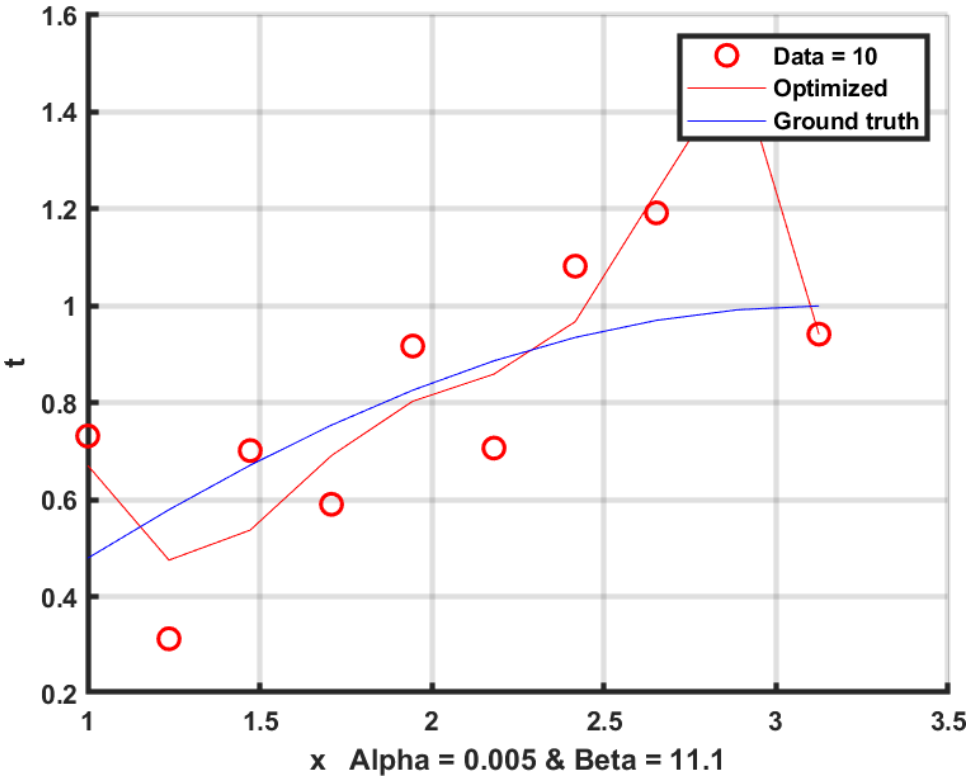

Figure 7: Data: 10pts, $\alpha = 0.005$, $\beta = 11.1$,M = 9

Varying α value. In below graph we can see that on increasing α variance is also getting increased. Note: Difference in image is not so conspicuous, for instance it can be seen at leftmost 2 data point.

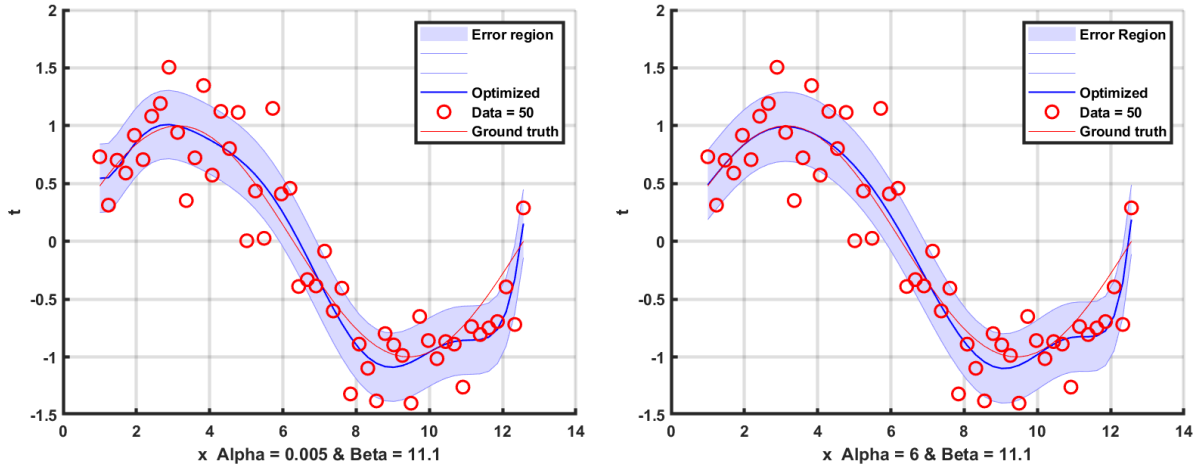

Figure 8: a)$\alpha = 0.005$, b) $\alpha = 6$ with $\beta = 11.1$, data 50pts

4 Conclusion

Increasing number of data points(N) in the dataset will reduce the error training error (Given that degree of polynomial sufficent to handle data distribution). Increasing number degree of polynomial certainly reduce the training error but may increase testing error. To avoid W* values bigger we can intriduce penalty, which is called regularization. We need to make balance between N and M such that we can avoid testing error. Another way of solving this curve fitting is using probability and decision theory. From the observation, it can be concluded that Maximum posterior estimation performs better then linear regression with regularization, but its performance depends on prior. Prior can depend on subject knowledge, hence it may take different time to converge.

Improvement Since most of the problem statement are open ended. 85% of the time was spent in writing project report, which actually turned starting fun into pain. It would have been great if it was just programming assignment.

5 References

References

[1] Bishop, Christopher M. Pattern Recognition and Machine Learning.