Curve Fitting

Tejas Mahale

January 2018

Abstract :

Regression is taking group of random variables and trying to find out mathematical relationship between them. Simple linear regression is analysis of relationship between a dependent and an independent variable and formulate the linear equation between both the variables[1]. Main idea is getting proper curve by evaluating relationship between dependent and independent variables so in further we can predict future values of dependent variables by providing independent variables. For this purpose, we need to estimate parameters of linear equation. There are multiple ways for parameter estimation for curve fitting. In this project we are concentrating on least square error minimization with and without regularization, probabilistic models of normal distribution using maximum likelihood (MLE) and maximum posterior(MAP) estimation. All these models gives similar results with some improved versions in terms of either accuracy or curve fitting[2].

Content :

1.Introduction:

1.1 Least square error minimization:

If we define linear model as:

$y(x, w) = w_0 + w_1 x^1 + w_2 x^2 + \ldots\ldots + w_M x^M$

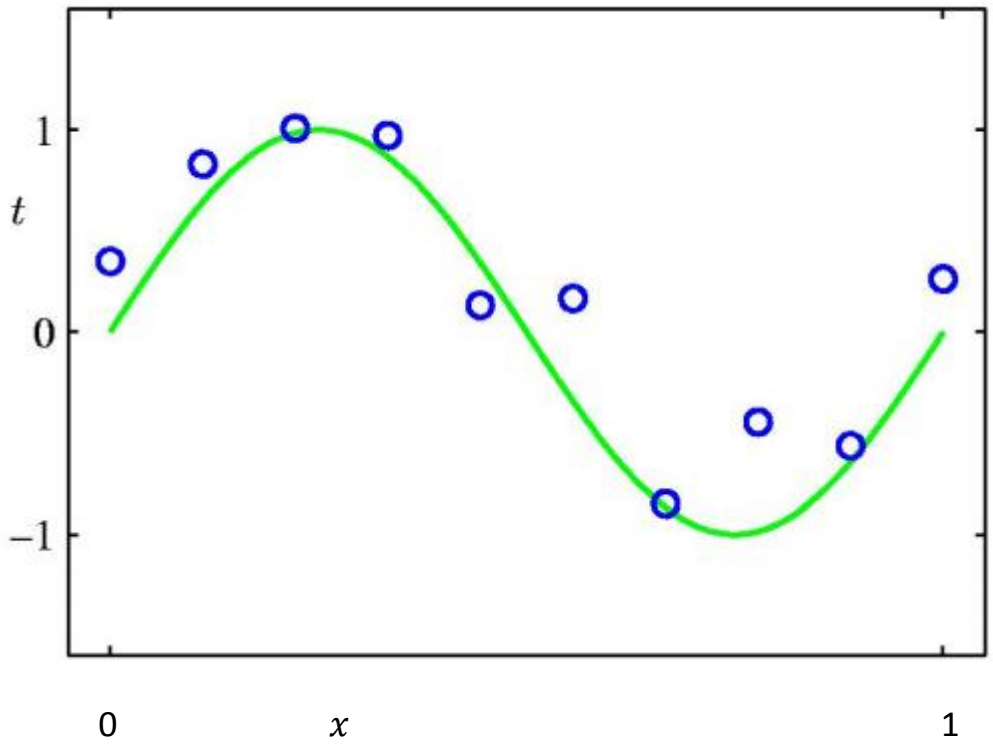


fig 1.1.1 Plotting of N = 10 point

here M= Order of polynomial

So we already have result of points in t observations for each point on x axis. Here we tries to find $y(x, w)$ for corresponding x point. Ultimately we are learning W matrix from given observed points.

Once we get $y(x, w)$, we use least square error using our original observations t as follow:

$$E(w) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, w) - t_n\}^2 \qquad Eq\ 1.1.1$$

Least square error is solution of over-determined equations[3]. We try to minimize $E(w)$, so we are ultimately trying to get a curve for which summation projection of every point on the curve is minimum.

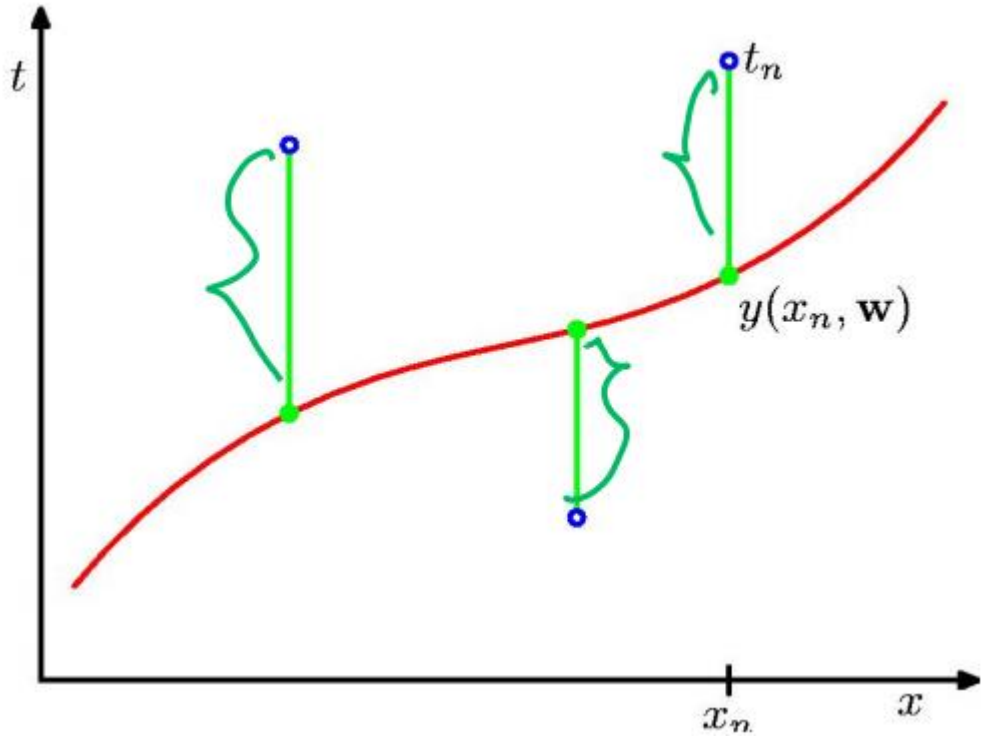


figure 1.1.2 Sum of square error (from class notes)

We will try to find W^* matrix from square error equation. We will get W^* from X and T matrices as :

$$\boldsymbol{W^* = (X^T X)^{-1} X^T T} \qquad Eq\ 1.1.2$$

We will prove this in section 2.2.1 Mathematics later in this project report

1.2 Linear curve fitting using regularization:

In this section we are using same sum square error equation along with regularization parameter (λ) to get rid of over fitting.

Over fitting : Curve fitting performance of training examples is excellent while performance of testing examples becomes worse as compare to training examples.

In linear regression set-up, when we have high degree of polynomial (M), curve tries to fit each and every data point and tries to pass through each of testing data points for minimum square error, leads to inappropriate shape causing over-fitting.

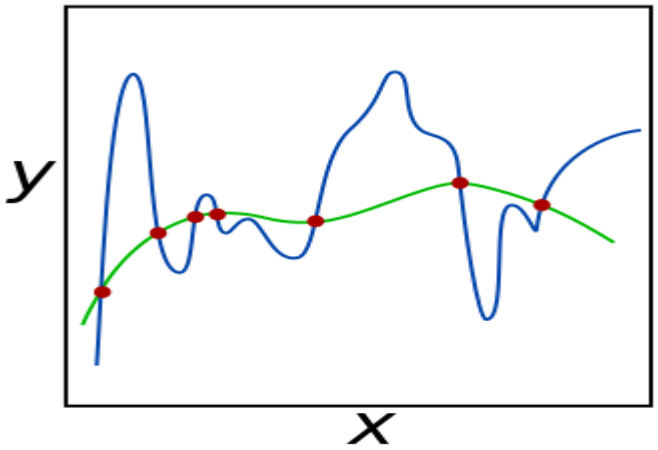


fig 1.2.1 Regularization (From Wikipedia)

Green curve is represents regularization model while blue curve is without regularization. Both curves passes through same data points, green curve (regularized) is more smoother.

Regularization can be introduced to square error function by adding penalty term to error function in order to discourage the coefficients of W to reach large values while minimizing the error function[2].

The simplest such penalty term takes the form of a sum of squares of all of the coefficients, leading to a modified error function as :

$$E(w) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, w) - t_n\} + \frac{\lambda}{2}||W||^2 \qquad Eq\ 1.2.1$$

In this case , W^* would be :

$$\boldsymbol{W^* = \ (X^T X \ + \ \frac{\lambda}{2} \ I)^{-1} X^T \ T}$$

I = Identity matrix, λ = Regularization parameter

we will prove this result in 2.2 Regularization mathematics section later in this project report.

There is alternative way for curve fitting using probability theory.

Bayesian probabilities :

With context to maximum likelihood estimation and maximum prior estimation we would like to review Bayesian probability.

$$p(Y\,|X) = \frac{p(X\,|Y)\,p(Y)}{p(X)} \;\; ; \;\; p(x) = \sum_{Y} p(X\,|Y)\,p(Y)$$

$$Posterior \;\propto Likelihood \text{ x } \text{Prior}$$

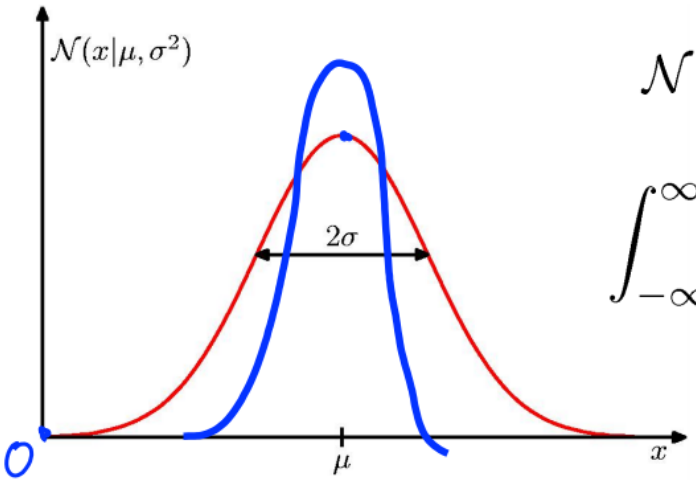
In our case the effect of observed data $D = \{t_1, t_2\; t_3 \ldots\ldots\ldots t_n\}$ can be expressed through conditional probability $p(D|w)$ [2] using *p(D):*

$$p(w|D) = \frac{p(D|w)\;\; P(w)}{p(D)} \qquad\qquad Eq\ 1.3.1$$

Gaussian Distribution:

In our case, we are going to use identically independent Gaussian(Normal) distribution model.

$$\mathcal{N}\left(x|\mu,\sigma^2\right) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$



$$\mathcal{N}(x|\mu,\sigma^2) > 0$$

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right)\,\mathrm{d}x = 1$$

here μ = mean and σ^2 = variance of Gaussian distribution.

1.3 Maximum Likelihood Estimation:

In maximum likelihood estimation, we will try to find parameters that the position of curve for given points. In other way, maximizing the likelihood function is minimizing error function[2].

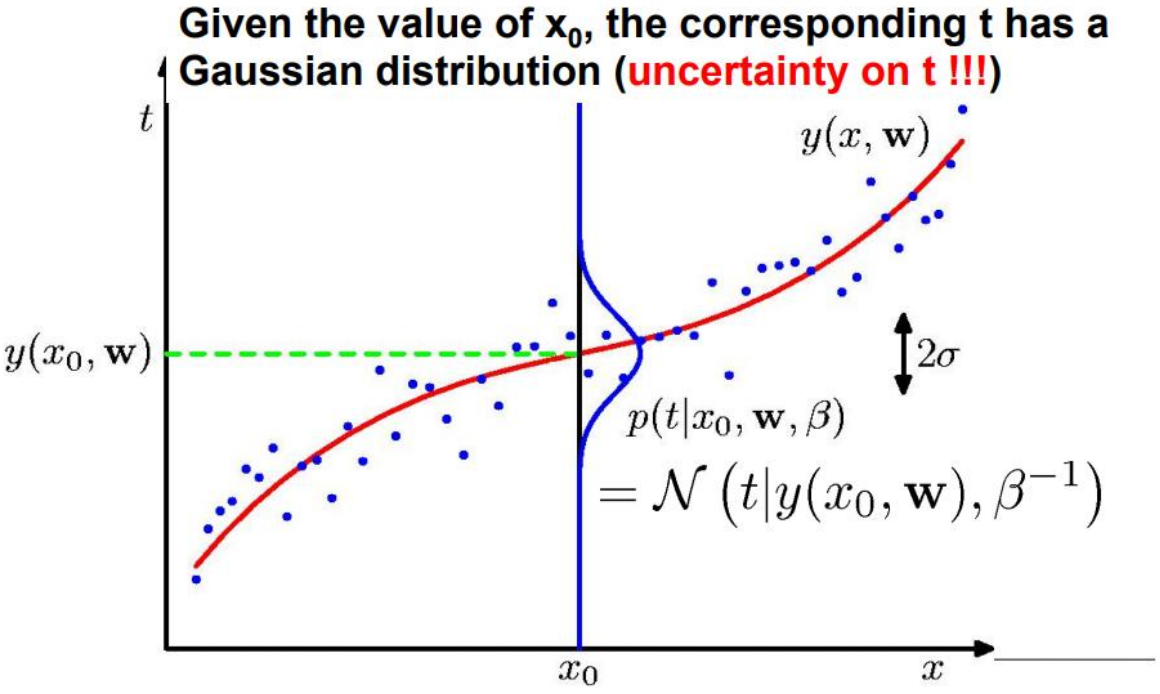
for MLE, $p(W|X,t,\beta\ ) \ \propto p(t|X,W,\beta)$

The parameters of a Gaussian distribution are the mean (□) and variance (σ^2)[2]. Given observations, $x_1, x_2, x_3 \ldots\ldots x_N$ the likelihood of those observations for a certain μ and σ 2 (assuming that the observations came from a Gaussian distribution)[4] is

$$p(x_1, x_2, x_3 \ldots\ldots x_N \mid □, \sigma^2) = \prod_{n=1}^{N} N\ (x_N \mid □, \sigma^2) \qquad Eq\ 1.3.1$$

where

$$N(x \mid □, \sigma^2) = \frac{1}{\sqrt{2\ \pi\ \sigma^2}} \exp\{\frac{-1}{2\sigma^2}\ \ (x-\mu)^2\} \qquad Eq\ 1.3.2$$



here $\sigma^2 = \beta^{-1}$

Our goal is choosing the value of W such that probability of observed data is maximize ie. we will find values of W and β by maximising the probability.

For that purpose we will use negative log likelihood function which ease our computation which makes function monotonically deceasing[2].

$$\ln p\ (X|\ ?, \sigma^2) = \frac{-1}{\sqrt{2\ \pi\ \sigma^2}} \sum_{n=1}^{N} (\ x_n - \ \mu)^2 - \frac{N \ln \sigma^2}{2} - \frac{N \ln(2\pi)}{2} \quad Eq\ 1.3.3$$

we have derived parameter estimation equations in section 2.3 MLE mathematics and we get following results:

$$\mu_{ML} = \frac{1}{N} \sum_{n=1}^{N} x_n \ ;\ \sigma^2{}_{ML} = \frac{1}{N} \sum_{n=1}^{N} (\ x_n - \ \mu_{ML})^2$$

We will use above result for our model by taking negative log likelihood on $y(x_{n,}\ W)$:

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

$$\ln p(t|X, W, \beta) = \frac{-\beta}{2} \sum_{n=1}^{N} \{\ y(x_n, W) - \ t_n\}^2 \ + \frac{N}{2} \ln \beta - \frac{N}{2} \ln 2\pi$$

we will get $\boldsymbol{\frac{1}{\beta_{ML}} = \frac{1}{N} \sum_{n=1}^{N} \{\ y(x_n, W_{ML}) - \ t_n\}^2}$

we have derived parameter estimation equations in section 2.3 MLE mathematics.

1.4 Maximum Posterior estimation

This is similar to what we observed in maximum likelihood estimation except here we introduce prior distribution over W with zero mean and precision α [2].

$$p(W\,|\alpha) = N(W\,|0, \alpha^{-1}\,I) = \frac{\alpha}{2\pi}^{\frac{M+1}{2}} \exp\left(\frac{-\alpha}{2}\,W^T\,W\right) \quad Eq\ 1.4.1$$

For MAP,

$$p(W|X, t, \alpha, \beta\,) \quad \propto \quad p(t|X, W, \beta)\,p(W|\alpha)$$

we already have

$$\ln p(t|X, W, \beta) = \frac{-\beta}{2} \sum_{n=1}^{N} \{\,y(x_n, W) - \,t_n\}^2 \,+ \frac{N}{2}\ln\beta - \frac{N}{2}\ln 2\pi$$

so we can find that the maximum of the posterior is given by the minimum of

$$\beta\,E(W) = \frac{\beta}{2} \sum_{n=1}^{N} \{\,y(x_n, W) - \,t_n\}^2 \,+ \frac{\alpha}{2}\,W^T W \qquad Eq\ 1.4.2$$

Maximum posterior is used over maximum likelihood because it avoids over fitting with regularization.

We can see that maximizing the posterior distribution is equivalent to minimizing the regularized sum-of-squares error function with a regularization parameter given by $\lambda = \frac{\alpha}{\beta}$.

Section 2 : Approach

2.1 Data used:

In all four cases, we have used N=50 point data. Data is generated using *generateData.m* function. It is basically sin(x) data points in addition of random noise value.

For case 2 regularization, to show Erms error on training and testing sets, we divided 50 points data in training set (N=40) and testing set (N =10).

x = 50 points in between 1 to 4π generated using linspace

y= 50 points generated using function sin(x)

t = 50 points generated by y + random noise

2.2 Mathematics :

Throughout the project we followed simple step, finding error function and minimizing the error function with respect to parameter which we wants to estimate.

2.2.1 Mathematics for Least Square error computation:

Error function for square error can be given as :

$$E(w) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, w) - t_n\}^2$$

We need to minimize this equation, ie. we need to find W^* for which E(W^*) is minimum.

$y(x, w) = w_0 + w_1 x^1 + w_2 x^2 + \ldots\ldots + w_M x^M \qquad here\ M = order$

Lets convert polynomials into matrix form:

$$X = \begin{bmatrix} x_1^0 & x_1^1 & \ldots\ldots & x_1^M \\ x_2^0 & x_2^1 & \ldots\ldots & x_2^M \\ & & \ldots\ldots\ldots & \\ x_N^0 & x_N^1 & \ldots\ldots & x_N^M \end{bmatrix}$$

$W = [w_0 \ w_1 \ w_2 \ \ldots\ldots\ldots\ldots\ldots\ldots \ w_M]$

$T = [t_0 \ t_1 \ t_2 \ \ldots\ldots\ldots\ldots\ldots\ldots \ t_N]$

So error function can be represented as:

$E(W) = \frac{1}{2} \ (XW - T)^T \ (XW - T)$ ----------> Eq 2.2.1.1

Taking transpose inside

$E(W) = \frac{1}{2} \ \{ (W^T X^T - T^T)(XW - T)\}$

inner multiplication,

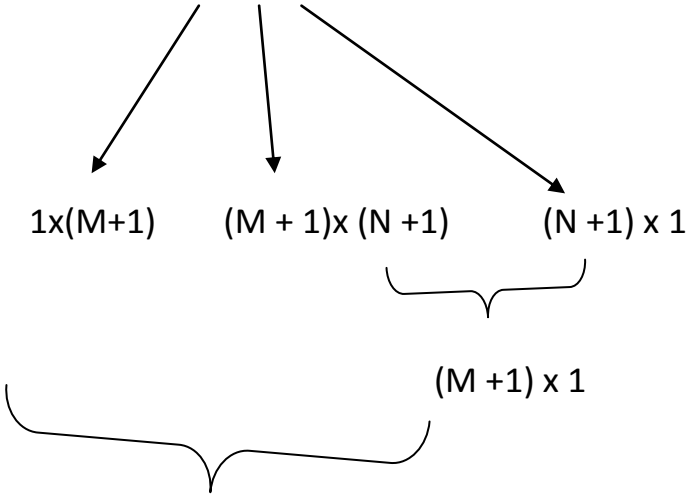
$$E(W) = \frac{1}{2}\{ W^T X^T XW - W^T X^T T - T^T XW + T^T T\} \quad \text{--------} > \text{Eq 2.2.1.2}$$

Differentiating w. r. t W

$$\frac{\partial E(W)}{\partial w} = \frac{1}{2}\left\{\frac{\partial (W^T X^T XW)}{\partial w} - \frac{\partial (W^T X^T T)}{\partial w} - \frac{\partial (T^T XW)}{\partial w} + \frac{\partial (T^T T)}{\partial w}\right\}$$

------- > Eq 2.2.1.3

Consider $W^T X^T T$



(1 x 1) => Scalar

$$\frac{\partial (W^T X^T T)}{\partial w} = X^T T \quad and \quad \frac{\partial (T^T XW)}{\partial w} = \frac{\partial (W^T X^T T)^T}{\partial w} = X^T T$$

----------Eq 2.2.1.4

Substituting Eq 2.2.1.4 in Eq 2.2.1.3 equating this equation to zero to get minimum :

$$\frac{\partial E(W)}{\partial w} = \frac{1}{2}\{ 2X^T XW^* - X^T T - X^T T + 0\} = 0$$

$$2X^T XW^* - X^T T - X^T T = 0$$

$$2X^T XW^* = 2\, X^T T$$

$X^T X W^* = X^T T$

Multiplying $(X^T X)^{-1}$ on both sides,

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

$$\boldsymbol{W^* = (X^T X)^{-1} X^T T} \qquad \textbf{Eq 2.2.1.5}$$

Now minimized squared error function become :

$$\boldsymbol{E(w^*) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, w^*) - t_n\}^2}$$

where $Y(X, W^*) = XW^*$

2.2.2 : Least square error with Regularization

Here we introduce regularization parameter (λ) in square error equation. This particular choice of regularizer is known in the machine learning literature as *weight decay* because in sequential learning algorithms, it encourages weight values to decay towards zero, unless supported by the data[2].

Here we are using L2 regularization.

So square error can be given as :

$$E(w) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, w) - t_n\}^2 + \frac{\lambda}{2}\mathrm{w}^2$$

converting above equation in matrix form:

$$E(W) = \frac{1}{2}(XW - T)^T (XW - T) + \frac{\lambda}{2}\mathrm{W}^T W$$ ----------> Eq 2.2.2.1

Taking transpose inside

$$E(W) = \frac{1}{2}\{(W^T X^T - T^T)(XW - T)\} + \frac{\lambda}{2}\mathrm{W}^T W$$

inner multiplication,

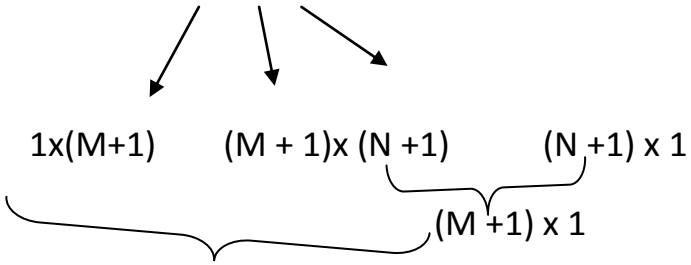
$$E(W) = \frac{1}{2}\{W^T X^T XW - W^T X^T T - T^T XW + T^T T\} + \frac{\lambda}{2}\mathrm{W}^T W$$

Differentiating w. r. t W

$$\frac{\partial E(W)}{\partial w} = \frac{1}{2}\left\{\frac{\partial (W^T X^T XW)}{\partial w} - \frac{\partial (W^T X^T T)}{\partial w} - \frac{\partial (T^T XW)}{\partial w} + \frac{\partial (T^T T)}{\partial w} + \lambda \frac{\partial (\mathrm{W}^T W)}{\partial w}\right\}$$

--------> Eq 2.2.1.2

Consider $W^T X^T T$



(1 x 1) => Scalar

$$\frac{\partial\,(W^T X^T T)}{\partial w} = X^T T \quad and \quad \frac{\partial (T^T XW)}{\partial w} = \frac{\partial (W^T X^T T)^T}{\partial w} \; = \; X^T T$$

----------Eq 2.2.2.3

Substituting Eq 2.2.2.3 in Eq 2.2.2.2 equating this equation to zero to get minimum :

$$\frac{\partial\, E(W)}{\partial w} = \; \frac{1}{2}\{ \; 2X^T XW^* \; - \; X^T T \; - X^T T + 0 \; + 2\lambda W^*\} = \; 0$$

$$2(X^T X + \lambda I)W^* \; - \; X^T T \; - X^T T = 0$$

$$2(X^T X + \lambda I)W^* = 2\, X^T T$$

$$(X^T X + \lambda I)W^* = \; X^T T$$

Multiplying $(X^T X + \; \lambda I)^{-1}$ on both sides,

$$(X^T X + \; \lambda I)^{-1}(X^T X + \lambda I)\, W^* = \; (X^T X + \; \lambda I)^{-1} X^T T$$

$$\boldsymbol{W^* = \; (X^T X + \; \lambda I)^{-1} X^T T} \qquad \textbf{-----> Eq 2.2.1.4}$$

Now minimized squared error function became :

$$\boldsymbol{E\left(w^*_{reg}\right) = \frac{1}{2} \sum_{n=1}^{N} \{ y\left(x_n, w^*_{reg}\right) - \; t_n \}^2 \; + \frac{\lambda}{2}\, w^{*2}_{reg}}$$

where $Y\left(X, W^*_{reg}\right) = \; XW^*_{reg}$

2.2.3 Maximum Likelihood estimation:

for MLE, $p(W|X,t,\beta\ ) \ \propto p(t|X,W,\beta)$

$p(t|X,W,\beta)$ is likelihood estimation of dependent parameter t w.r.t independent parameter X. We will try to find estimate of W and β.

Here we are using normal iid distribution.

$N(t_n \mid y(x_n,W),\beta^{-1})= \frac{\sqrt{\beta}}{\sqrt{2\,\pi}} \exp\{ \frac{-\beta}{2} \ \ (y(x_n,W)-t_n)^2\}$.....Eq 2.2.3.1

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

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

Taking log likelihood on both sides:

$$\ln p(t|X,W,\beta) = \frac{-\beta}{2} \sum_{n=1}^{N} (y(x_n,W)-t_n)^2 + \frac{N}{2}\ln(\beta) - \frac{N}{2}\ln(2\pi)$$

.....Eq 2.2.3.2

Differentiating above Eq w .r.t W, we will get

$$\frac{1}{p(t|X,W,\beta)} = \frac{-\beta}{2} \frac{d\ (E(W))}{dW}$$

here $E(w) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n,w) - t_n\}^2$

So to maximize $p(t|X,W,\beta)$ we need to minimise $E(w)$ which we already calculated, we get

$$\boldsymbol{W^* = (X^T X)^{-1} X^T T} \qquad \text{from Eq 2.2.1.5}$$

Now let's estimate β .

from Eq 2.2.3.2

$$\ln p(t|X,W,\beta) = \frac{-\beta}{2}\sum_{n=1}^{N}(y(x_n,W)-t_n)^2 + \frac{N}{2}\ln(\beta) - \frac{N}{2}\ln(2\pi)$$

Differentiating w.r.t β equating to zero maximise $p(t|X,W,\beta)$:

$$\frac{1}{p(t|X,W,\beta)} = \frac{-1}{2}\sum_{n=1}^{N}(y(x_n,W)-t_n)^2 + \frac{N}{2\beta} = 0$$

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

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

These W^* and β^{-1} are used to maximise the $p(t|X,W,\beta)$ while curve fitting.

2.2.4 Maximum Posterior Mathematics:

For MAP,

$$p(W|X,t,\alpha,\beta\ ) \quad \propto \quad p(t|X,W,\beta)\ p(W|\alpha) \quad \text{-------Eq 2.2.4.1}$$

We can now determine **w** by finding the most probable value of **w** given the data, in other words by maximizing the posterior distribution. This technique is called maximum posterior, or simply MAP.

We will try to find estimate of W and β.

Here we are using normal iid distribution

$$p(W\ |\alpha) = N(W\ |0,\alpha^{-1}\ I) = \frac{\alpha}{2\pi}^{\frac{M+1}{2}} \exp\left(\frac{-\alpha}{2}\ W^T\ W\right) \quad from\ \ Eq\ 1.4.1$$

$$\ln p(t|X,W,\beta) = \frac{-\beta}{2}\sum_{n=1}^{N}(\ y(x_n,W) - t_n)^2 + \frac{N}{2}\ln(\beta) - \frac{N}{2}\ln(2\pi)$$

from Eq 2.2.3.2

Taking negative ln on Eq 2.2.4.1 and combining Eq 1.4.1 and Eq 2.2.3.2

Maximum posterior solution :

$$W^*_{map} = \text{argmax}_{\text{W}}\ p(W|X,t,\alpha,\beta\ )$$

$$W^*_{map} = \text{argmax}_{\text{W}}\ \{\ln p(t|X,W,\beta)\ + \ln p(W|\alpha)\ \}$$

$$W^*_{map} = \text{argmax}_{\text{W}}\ \left\{\frac{\beta}{2}\sum_{n=1}^{N}(\ y(x_n,W) - t_n)^2 - \frac{N}{2}\ln(\beta) + \mathcal{N}2\ln 2\pi - \text{M}+12\ \ln\alpha + \text{N}2\ln 2\pi + \alpha 2\ W^T W\right.$$

Let's ignore the constant and take negative sign out

$$W^*_{map} = \text{argmin}_W \{ \frac{\beta}{2} \sum_{n=1}^{N} (y(x_n, W) - t_n)^2 + \frac{\alpha}{2} W^T W \}$$

Let's find minimum value of $\frac{\beta}{2} \sum_{n=1}^{N} (y(x_n, W) - t_n)^2 + \frac{\alpha}{2} W^T W$

$$E(W) = \frac{\beta}{2} \sum_{n=1}^{N} (y(x_n, W) - t_n)^2 + \frac{\alpha}{2} W^T W$$

Differentiating w. r. t W

$$\frac{\partial E(W)}{\partial w} = \frac{\beta}{2} \left\{ \frac{\partial (W^T X^T X W)}{\partial w} - \frac{\partial (W^T X^T T)}{\partial w} - \frac{\partial (T^T X W)}{\partial w} + \frac{\partial (T^T T)}{\partial w} \right\} + \frac{\alpha}{2} \frac{\partial (W^T W)}{\partial w} = 0$$

Dividing by β :

$$2(X^T X + \frac{\alpha}{\beta} I) W^* - X^T T - X^T T = 0$$

$$2(X^T X + \frac{\alpha}{\beta} I) W^* = 2 X^T T$$

$$(X^T X + \frac{\alpha}{\beta} I) W^* = X^T T$$

Multiplying $(X^T X + \frac{\alpha}{\beta} I)^{-1}$ on both sides,

$$(X^T X + \frac{\alpha}{\beta} I)^{-1} (X^T X + \frac{\alpha}{\beta} I) W^* = (X^T X + \frac{\alpha}{\beta} I)^{-1} X^T T$$

$$\boldsymbol{W^*_{MAP} = (X^T X + \frac{\alpha}{\beta} I)^{-1} X^T T}$$

3. Observations :

We used 50 point data set and got following results:

3.1 Least Square error:

a) We can observe that, for order 0 and 1, we get horizontal linear line.
b) As we increased order to 3, we got non linear curve, but it's shape is far away from original sinusoidal curve.
c) With order = 6, we got best possible curve for least square error. It is sinusoidal and close to original sin wave.
d) Order 9 curve tries to fit most of points which leads to over-fitting. Curve becomes oscillatory with this order.
e) E_{rms} observed table with respect to order

Order (N)	E_{rms}
0	0.8443
1	0.8442
3	0.5111
6	0.2350
9	0.2011

Fig 3.1.1

From figure 3.1.1, you can see that as order increases, curve gets more liberty to mould its shape and hence Erms reduces. But with higher order like N=9, overfitting problem arises.

Figures:

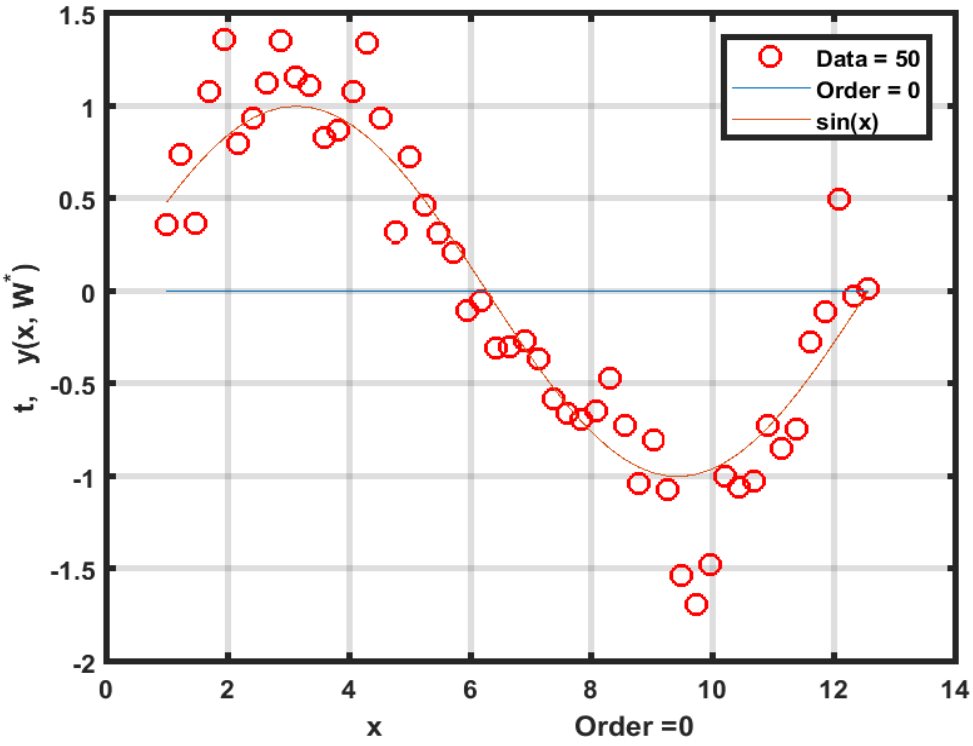


Fig 3.1.2

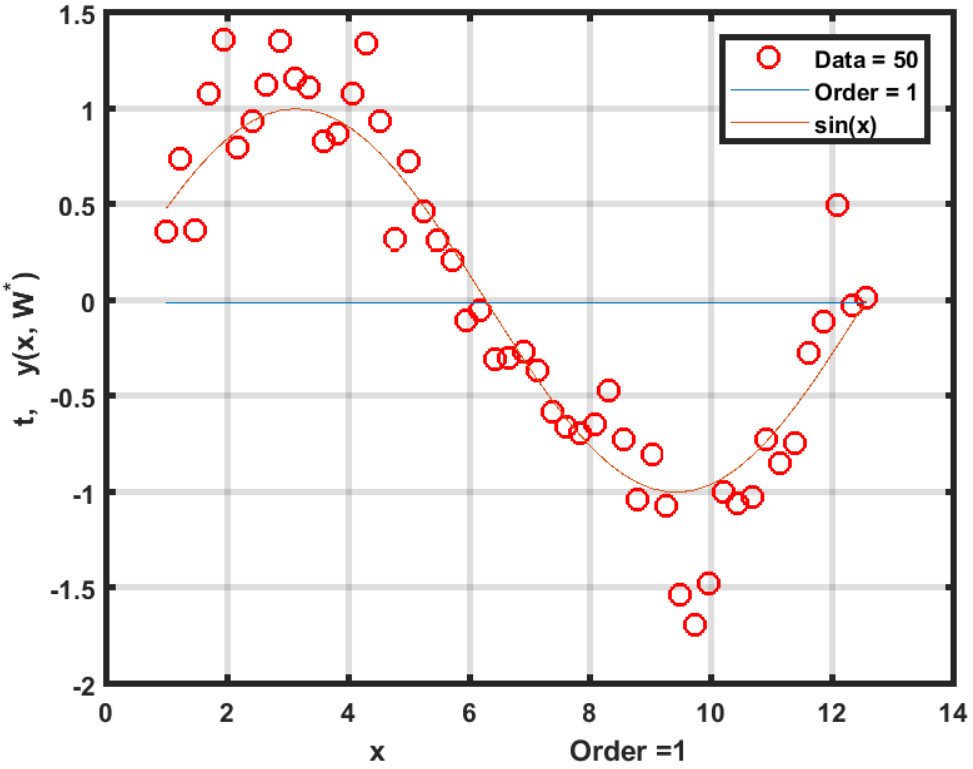


Fig 3.1.3

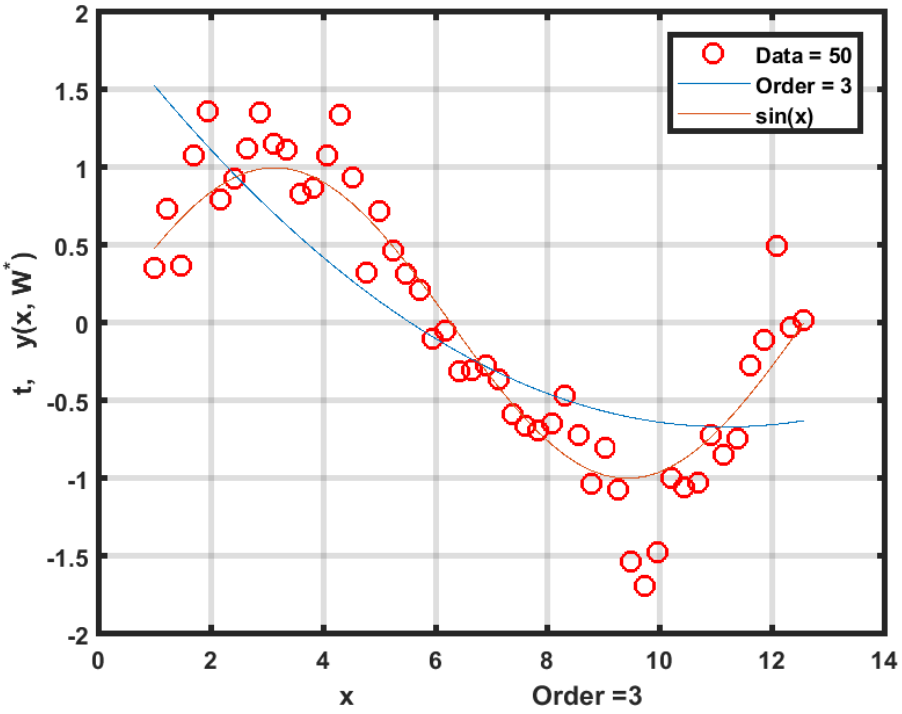


fig 3.1.4

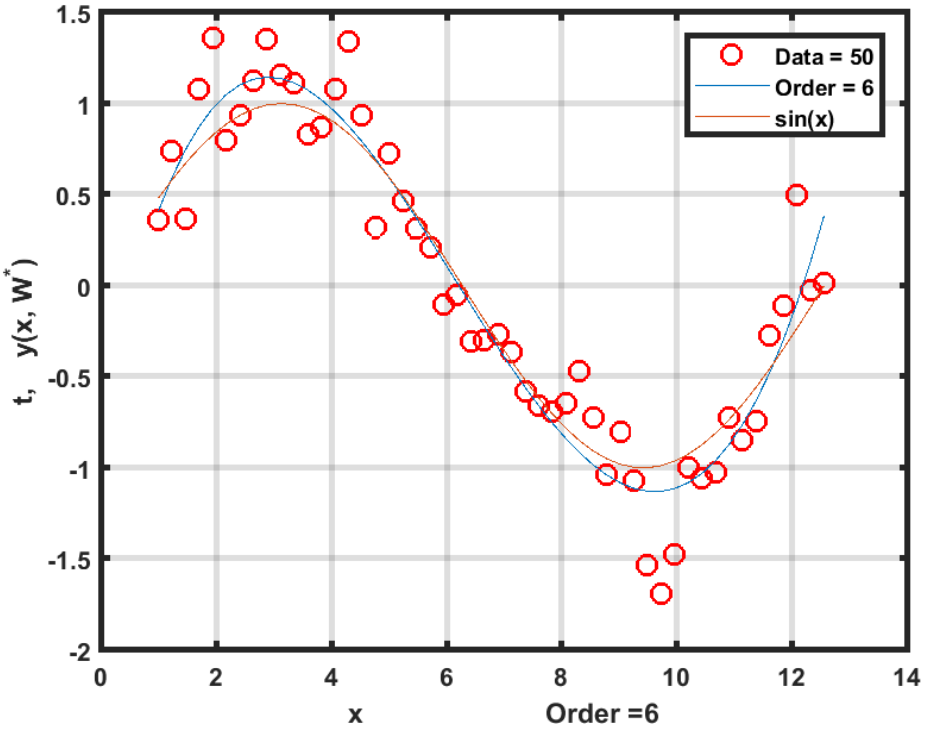


fig 3.1.5

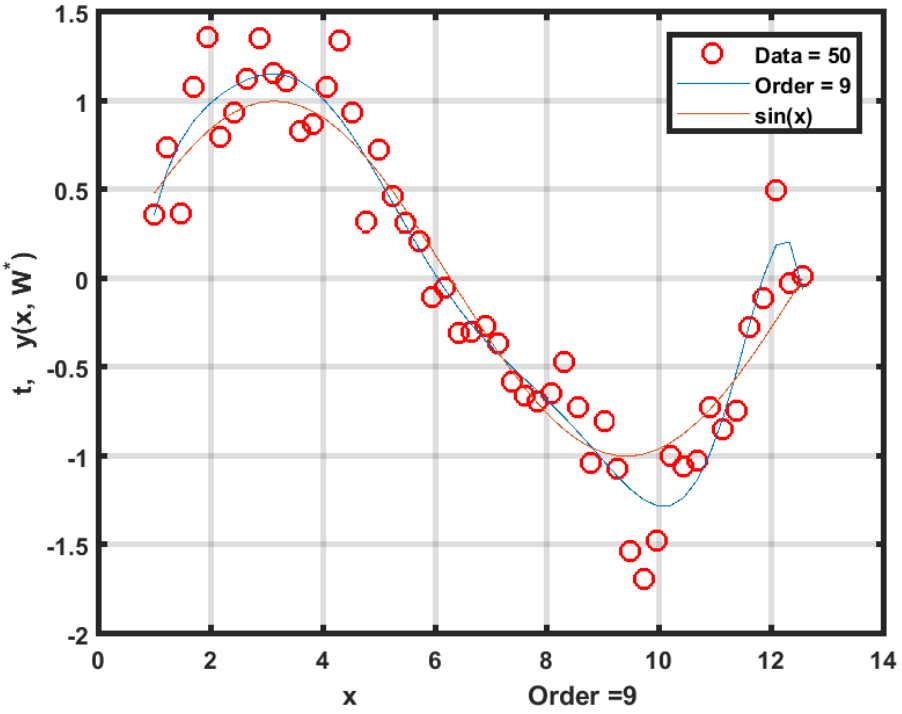


fig 3.1.6

Coefficient (W) :

(W)	N=1	N=3	N=6	N=9
W_1	-0.0118	1.98045	-0.7423	-3.3289
W_2		-0.4733	1.4754	7.6394
W_3		0.02114	-0.349415541983368	-6.0685
W_4			0.0236646901493272	2.7403
W_5			-0.00044	-0.7291
W_6			6.16918e-06	0.11420
W_7				-0.01032
W_8				0.0004976
W_9				-9.8773273e-06

fig 3.1.7

3.2 Results for least square error with Regularization:

Figures:

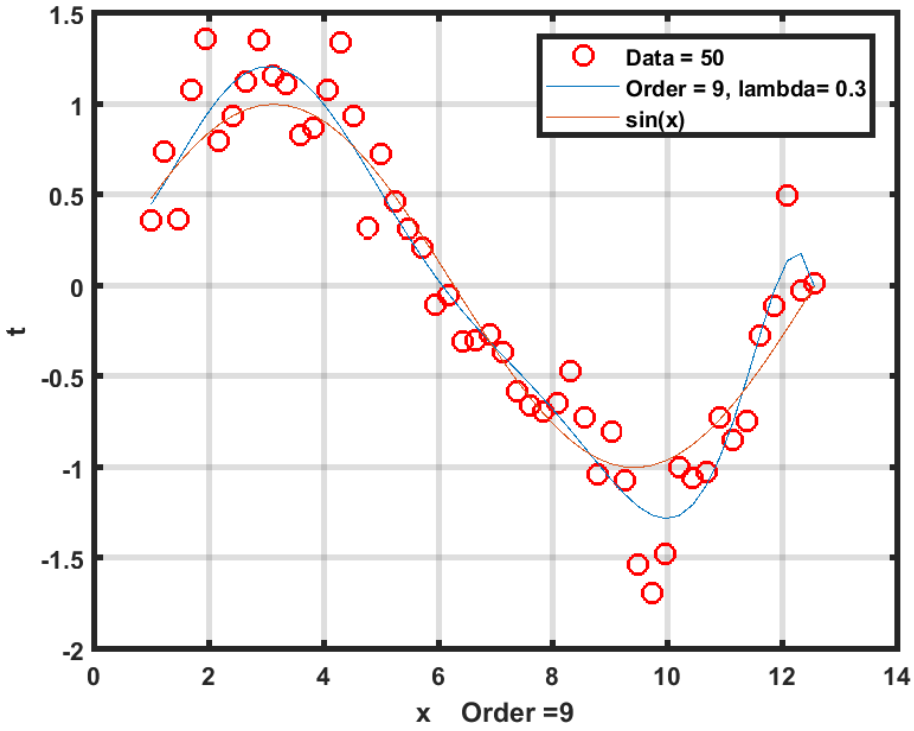


fig 3.2.1

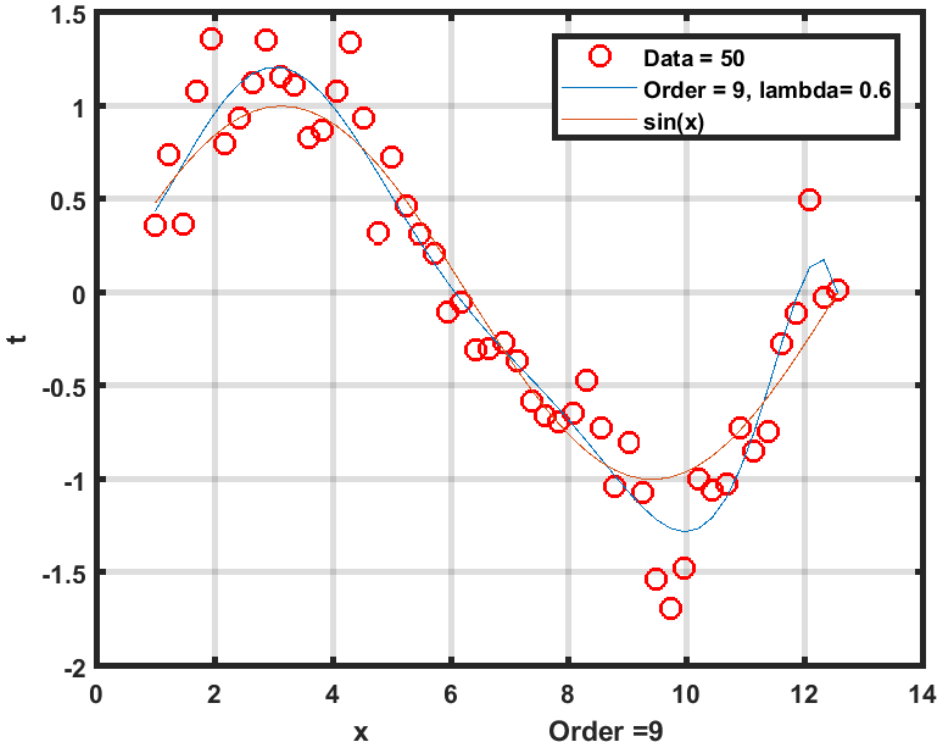


fig 3.2.2

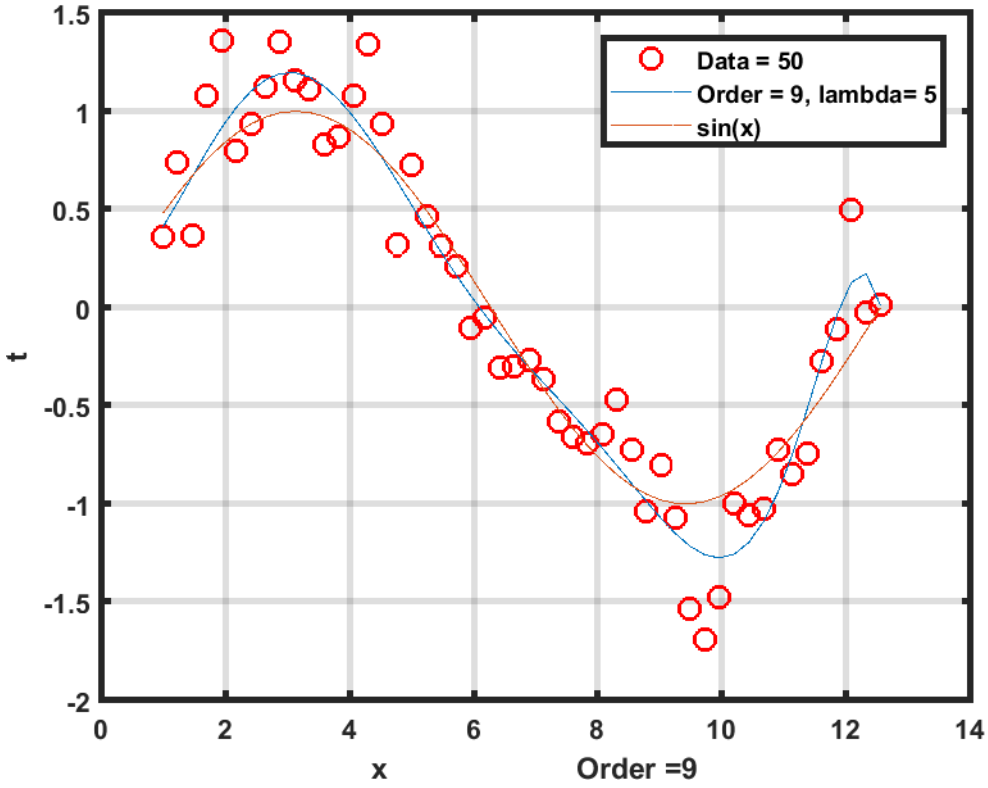


fig 3.2.3

Since for Order = 9, even though it is over-fitting, effect of regularization is minimum as W coefficients are not in huge range. Difference between these values is not erratic.

Table of W matrix with order N = 9

W	λ =0	$\lambda = 0.3$	$\lambda = 0.6$	$\lambda = 5$
W_1	-3.3289	0.1509	0.1264	0.08326
W_2	7.6394	0.1281	0.1258	0.12258
W_3	-6.0685	0.0838	0.1209	0.16138
W_4	2.7403	0.1861	0.1589	0.12685
W_5	-0.7291	-0.1290	-0.1203	-0.108786
W_6	0.11420	0.03110	0.02964	0.027438
W_7	-0.01032	-0.0036	-0.003475	-0.0032480
W_8	0.0004976	0.00020	0.0001985	0.000186532
W_9	-9.8773273e-06	-4.5719459e-06	-4.442639e-06	-4.188181831e-06

fig 3.2.4

From fig 3.2.4 we can see the effect of regularization on weights. As regularization parameter value increase, weights tends to be more distributed.

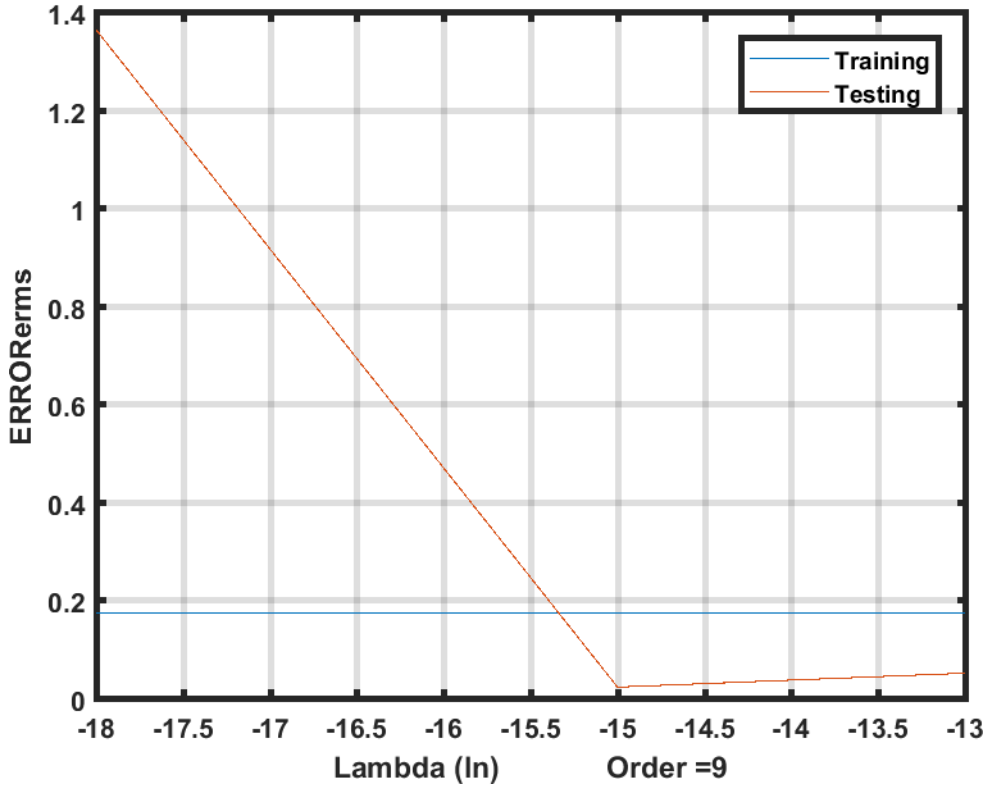


fig .3.2.5

	ln(lambda) = -18	ln(lambda) = -15	ln(lambda) = -13
Training Erms	0.17573723	0.17573906	0.17575118
Testing Erms	1.36579477	0.02496061	0.05336623

fig 3.2.6

3.3 Results for Maximum likelihood estimation:

figures:

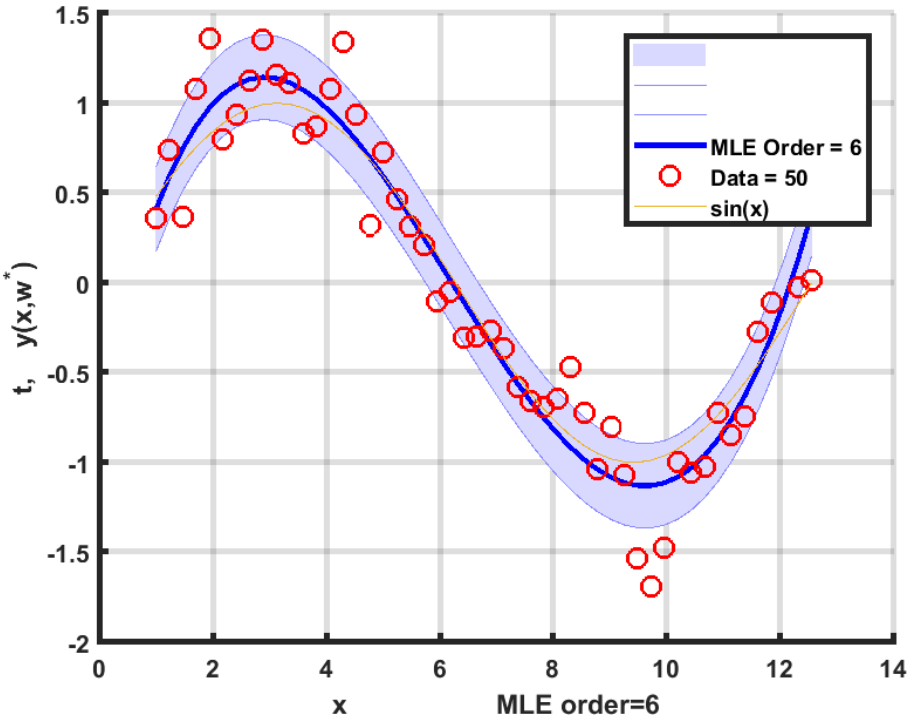


fig 3.3.1

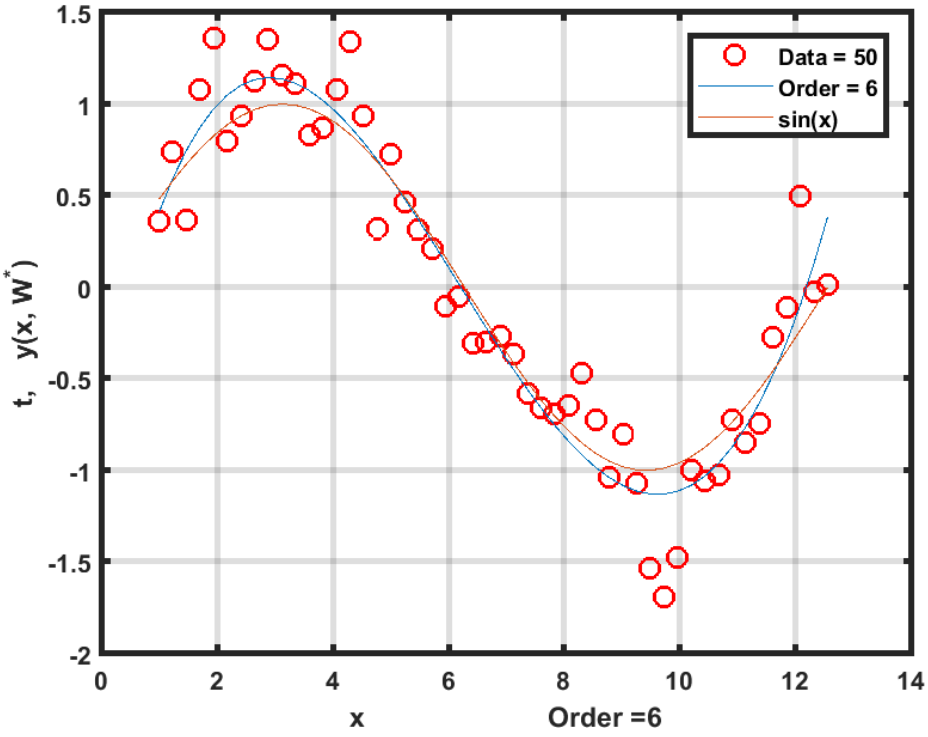


fig 3.1.5

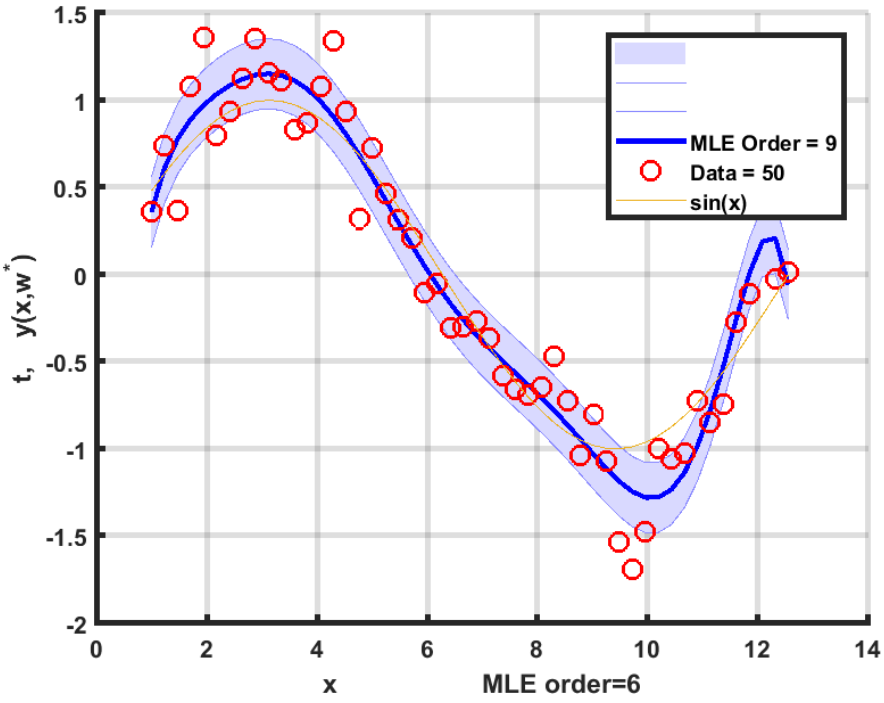


fig 3.3.2

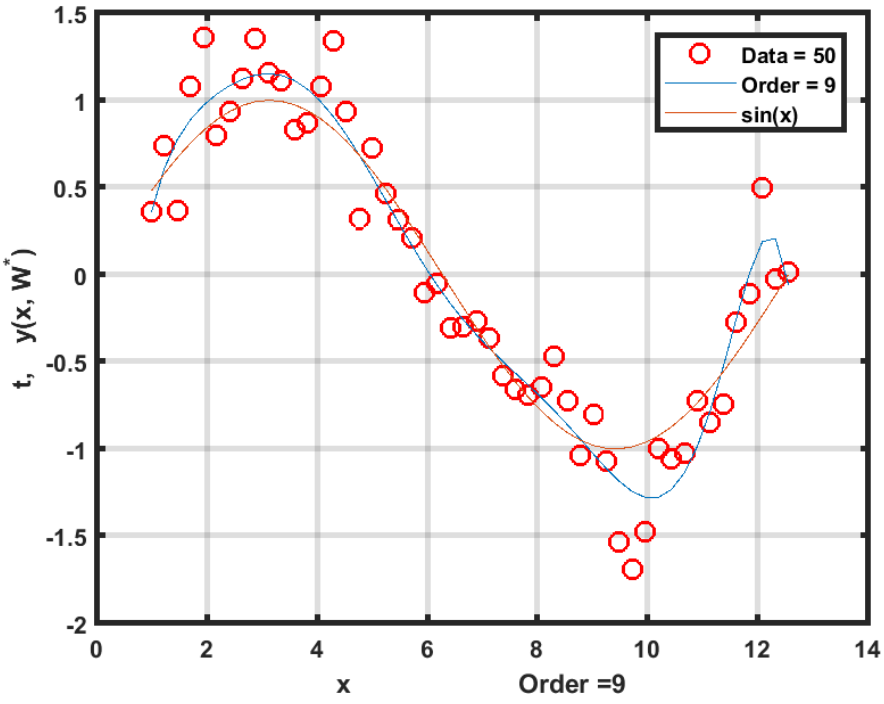


fig 3.1.6

By comparing MLE image 3.3.1 and least square image 3.1.5, we can say that shaded region is covering most of points.

Order	Variance (β^{-1})
6	0.0552
9	0.0404

3.4 Maximum Posterior estimation

figures:

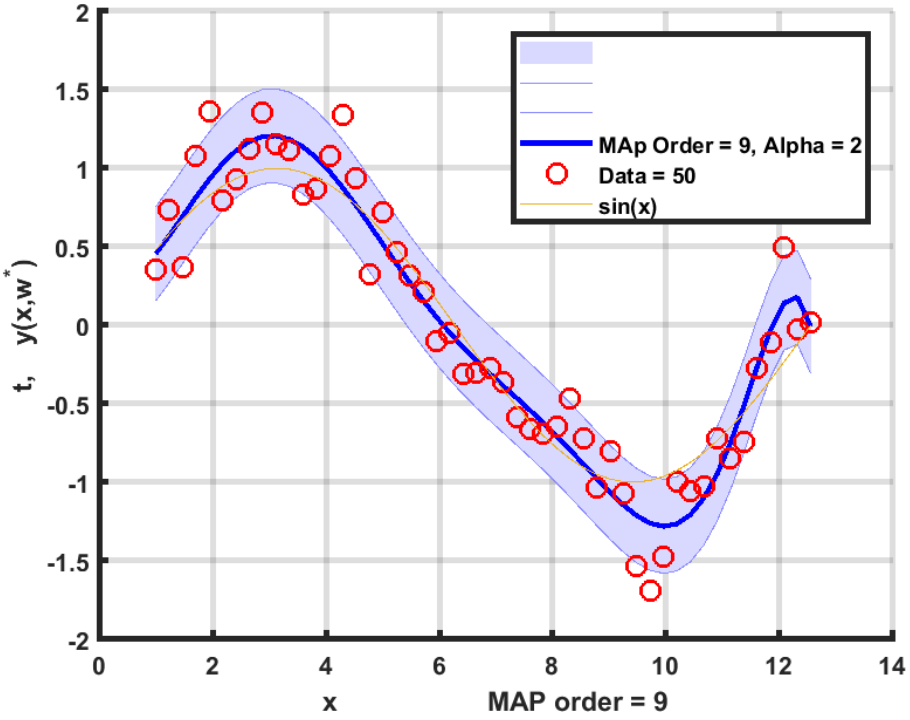


fig 3.4.1

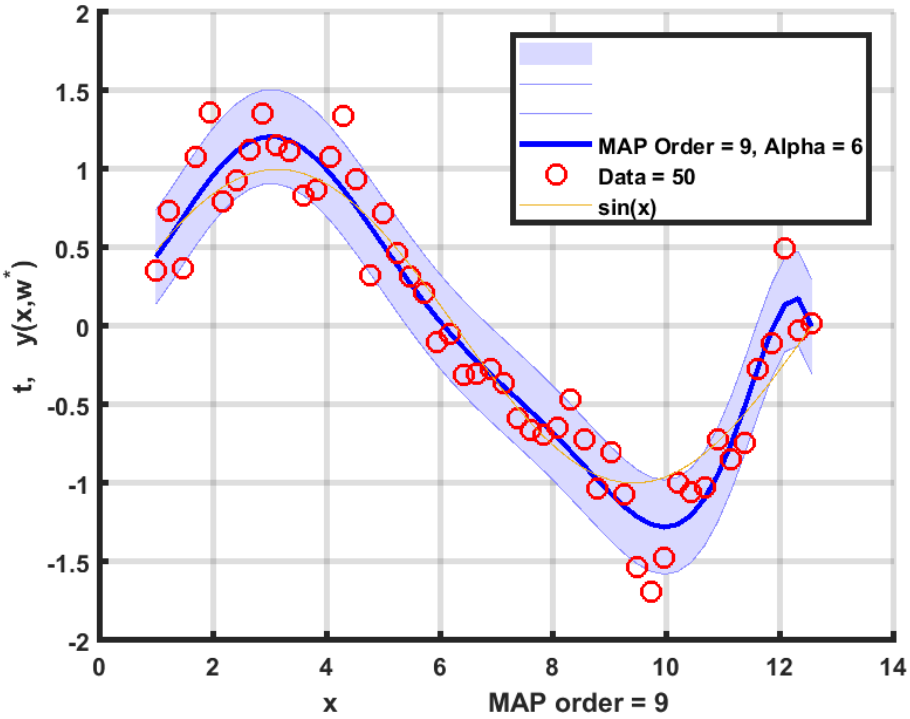


Fig 3.4.2

Table :

In this case our beta is constant Beta = 11.11

W	Alpha = 0	Alpha = 2	Alpha =6
W_1	-3.3289	0.16882	0.129904
W_2	7.6394	0.13646	0.125893
W_3	-6.0685	0.04577	0.116330
W_4	2.7403	0.21240	0.162462
W_5	-0.7291	-0.1371	-0.12151
W_6	0.11420	0.03245	0.029837
W_7	-0.01032	-0.00373	-0.00349
W_8	0.0004976	0.000211	0.000199
W_9	-9.8773273e-06	-4.6883667e-06	-4.45981e-06

Above table shows that maximum posterior estimator is natural regularizer. As we increase Alpha value, weights tend to be more distributed. As we increase alpha value, positive weights decreased and value of negative weights increased.

4. Conclusion:

4.1 Least Square Error method:

This method gave good results for order 6 and 50 points. But if you keep the number of points same and increase the order, sum of square error would be decreased further but this lead to over-fitting problem. Being said that, least square error minimization is useful optimal estimation for unknown parameters. Square error estimations are used in computation of every other model parameter estimations like MLE or MAP.

4.2 Regularization:

Improper weight balance on square error minimization lead to over-fitting. By adding penalty term in square error equation limited the weight coefficients to reach large values. In our case, regularization was more effective in order 9 where over-fitting problem occurred.

4.3 Maximum Likelihood estimation:

This alternative way for curve fitting gave similar results as compare to square error by using maximum likelihood solution under normal distribution noise assumption. To avoid over fitting, we can marginalise over parameters using Bayesian settings[2].

4.4 Maximum Posterior estimation :

By adding prior along with likelihood, posterior estimations can be maximised. Maximum posterior estimation provides regularization and more consistency as compare to MLE. If prior distribution is constant, there is no difference in MLE and MAP.

5. References

1] Gülden Kaya Uyanık, Neşe Güler, A Study on Multiple Linear Regression Analysis, Procedia - Social and Behavioral Sciences, Volume 106, 2013,

2] Bishop: Pattern Recognition and Machine Learning

[3] Stanford notes: https://see.stanford.edu/materials/lsoeldsee263/05-ls.pdf

[4] Wei Ho, Michael Ye, Princeton course COS424 : Interacting with data