Answers to problems for Week 10

Ans 1

$$w^{(\tau+1)} = w^{(\tau)} - \eta \nabla E_n$$

Sum of the squares error function is given by

$$E_n(w) = \frac{1}{2}\{t_n - w^T \phi(x_n)\}^2$$

Taking gradient we have

$$\nabla E_n(w) = -(t_n - w^T \phi(x_n)\phi(x_n)$$

and substituting in the above equation we get

$$w^{(\tau+1)} = w^{(\tau)} - \eta(-(t_n - w^{(\tau)T}\phi_n))\phi_n)$$

$$w^{(\tau+1)} = w^{(\tau)} + \eta(t_n - w^{(\tau)T}\phi_n)\phi_n)$$

Ans 2

For Bayesian regression Bishop uses a prior

$$p(w) = N(w|m_0, S_0)$$

and a likelihood function

$$p(t|w) = \prod_{n=1}^{N} N(t_n|w^T\phi(x_n), \beta^{-1})$$

He then shows that posterior distribution is given by

$$p(w|t) = N(w|m_N, S_N)$$

$$m_N = S_N(S_0^{-1}m_0 + \beta\phi^T t)$$

$$S_N^{-1} = S_0^{-1} + \beta\phi^T\phi$$

If $S_0 = \alpha^{-1}I$ then

$$m_N = S_N(\alpha m_0 + \beta\phi^T t)$$

$$S_N^{-1} = \alpha I + \beta\phi^T\phi$$

As $\alpha \to 0 S_N^{-1} \to \beta\phi^T\phi$

$$S_N \to \beta^{-1}(\phi^T\phi)^{-1}$$

$$m_N \to \beta^{-1}(\phi^T\phi)^{-1}\beta\phi^T t$$

$$m_N = (\phi^T\phi)^{-1}\phi^T t$$

Table 1: Error Values

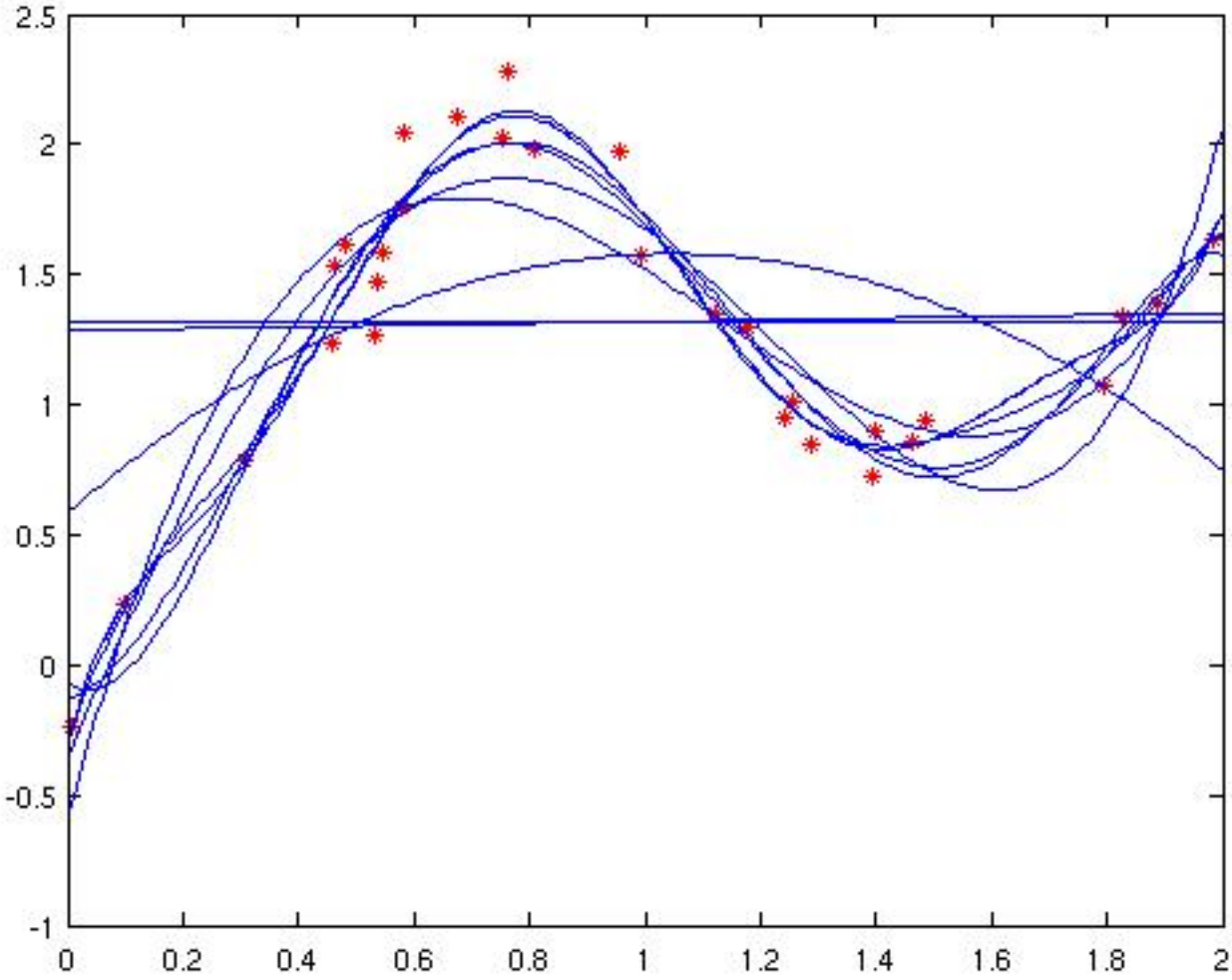
Polynomial degree	Squared loss as given by Equation 3.12
0	4.6633
1	4.6587
2	3.6222
3	0.8091
4	0.6104
5	0.3523
6	0.3389
7	0.2426
8	0.2426
9	0.2365

Ans 3

a The plot of the polynomials of various degrees is attached in this pdf. The sum of **half** the squared loss as given by equation 3.12 are given in the table.

b Once we randomly select half the data as the training and the remaining data as testset data, the training set error decreases monotonically as the degree k increases. Comparing it to the previous part there is a slight improvement in the average error. This may be due to the fact as the number of the points and degrees converge the polynomial is better able to fit the points. The test set error first decreases and hits lowest at around degree 3 and then increases with increase in the degree suggesting that there is overfitting beyond degree 3

c Lamda values above zero cause the weights especially of the high powers of x to be much smaller reducing the risk of overfitting. The training set errors are larger but testing set errors are reduced overall.
Matlab program for plotting the polynomials of various degrees and computing the errors is also attached in this document.


2.5
2
1.5
1
0.5
0
-0.5
-1
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2

```
load -ascii a10_data.txt;
[rows,cols] = size(a10_data);
X=a10_data(:,1);
Y=a10_data(:,2);
C1=X.^0;
C2=X.^1;
C3=X.^2;
C4=X.^3;
C5=X.^4;
C6=X.^5;
C7=X.^6;
C8=X.^7;
C9=X.^8;
C10=X.^9;
M=[C1,C2,C3,C4,C5,C6,C7,C8,C9,C10];
Errors=zeros(10,1);
plot(X,Y,'r*');
XT=0:0.01:2;
hold on;
for i=1:10
   Phi=M(:,1:i)
   WML=inv(Phi'*Phi)*Phi'*Y ;
   WMLREV = WML(end:-1:1);
   YT=polyval(WMLREV,XT);
   for j=1:30
   Errors(i,1)=Errors(i,1)+(Y(j)-polyval(WMLREV,X(j)))^2
   end
   Errors(i,1)=Errors(i,1)*0.5
   plot(XT,YT,'b');
   hold on
end
hold off
Errors
```

Ans 4

Equation 3.30 which is the constraint can be written as

$$\sum_{j=1}^{M} |w_j|^q - \eta \leq 0$$

Now adding half of the above equation to the error equation 3.12 using the langrange method we have

$$L = \frac{1}{2}\sum_{n=1}^{N}\{t_n - w^T\phi(x_n)\}^2 + \frac{\lambda}{2}(\sum_{j=1}^{M} |w_j|^q - \eta)$$

Now the above equation and the regularized error function 3.29 have the same form in terms of w. Hence minimizing 3.29 is equivalent to minimizing the 3.12 subject to

constraint 3.30. Minimizing the above equation and using E.9 , E.10 and E.11 we get the value of η as

$$\eta = \sum_{j=1}^{M} |w_j|^q$$

Ans 5

Using $y(x, w) = w^T \phi(x)$, $p(t|x, w, \beta) = N(t|y(x, w), \beta^{-1})$ and $p(w|\mathbf{t}) = N(w|m_N, S_N)$ we can write equation 3.57

$$p(t|\mathbf{t}, \alpha, \beta) = \int p(t|w, \beta) p(w|\mathbf{t}, \alpha, \beta) dw$$

as

$$p(t|\mathbf{t}, \alpha, \beta) = \int N(t|\phi(x)^T w, \beta^{-1}) N(w|m_N, S_N) dw$$

Using equation 2.113 $p(x) = N(x|\mu, \Lambda^{-1})$ and equation 2.114 $p(y|x) = N(y|Ax + b, L^{-1})$ with posterior result 2.115 $p(y) = N(y|A\mu + b, L^{-1} + A\Lambda^{-1}A^T)$ Now our first factor in the integrand is the likelihood and the second factor is the prior. Comparing the results we get the result $p(t|x, \mathbf{t}, \alpha, \beta) = N(t|m_N^T \phi(x), \sigma_N^2(x))$ where the variance is given by

$$\sigma_N^2(x) = \frac{1}{\beta} + \phi(x)^T S_N \phi(x)$$