Birla Institute of Technology & Science, Pilani
Work-Integrated Learning Programmes Division
First Semester 2019-2020
Comprehensive Examination (Makeup)

Course No.	: PCAM* ZC211
Course Title	: REGRESSION
Nature of Exam	: Closed Book
Weightage	: 40%
Duration	: 3 Hours
Date of Exam	: **09/11/2019 (FN)**

No. of Pages	= 2
No. of Questions	= 6

Q1. I collect a set of data (n = 100 observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression, i.e. $Y = \beta_0 + \beta_1X + \beta_2X^2 + \beta_3X^3 + \mathcal{E}$. [3+3+3+ 3 Marks]

(a) Suppose that the true relationship between X and Y is linear, i.e. $Y = \beta_0 + \beta_1X + \mathcal{E}$. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.
Sol: The cubic regression model is more flexible than the linear regression model. Therefore, we would expect the cubic model to fit better the data, and thus to have lower training RSS.

(b) Answer (a) using test rather than training RSS.
Sol: If the true relationship between X and Y is linear, a cubic regression model is excessively flexible, and we would expect the method to fit test data poorly. Therefore, we would expect the cubic model to have a higher test RSS.

(c) Suppose that the true relationship between X and Y is not linear, but we don't know how far it is from linear. Consider the training RSS for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.
Sol: Same answer as (a).

(d) Answer (c) using test rather than training RSS.
Sol: In this case, we do not know the right amount of flexibility to fit the true underlying model. So there is not enough information to tell which model would give the lower test RSS.

Q2. Suppose there is a uni-variate regression problem and you are given 'N' training examples. The regression model obtained after minimizing sum of squares of error is y = 5.22 + 0.021 x. As the w1 is estimated to be 0.021, which seems relatively smaller value, can we conclude upfront that y is actually not dependent on x. Support your answer with an appropriate reasoning? [6 Marks]

Sol:

We are given 'N' training examples and regression model obtained by minimizing the sum of squares of error is

$$y = 5.22 + 0.021\,x.$$

We can't conclude that the dependent variable, y is not dependent on x just because, $\hat{w}_1$, 0.021, is ~~relative~~ smaller value.

Let $(x_1, t_1), (x_2, t_2), \ldots, (x_N, t_N)$ be 'N' training examples.

$$RSS = \sum_{n=1}^{N} \left(t_n - (5.22 + 0.021 x_n) \right)^2$$

$$SE(\hat{w}_1) = \frac{\sqrt{RSS/(N-2)}}{\sum_{n=1}^{N} (x_n - \bar{x})}$$

$$H_0 : w_1 = 0$$

whereas $H_a : w_1 \neq 0$

In practice, we compute t-statistic given by $t = \frac{\hat{w}_1 - 0}{SE(\hat{w}_1)}$

If there is no relationship between X and Y, then we expect that t-statistic to follow t-distribution with N-2 degrees of freedom.

Consequently, it is a simple matter to compute the probability of observing any number equal to $|t|$ or larger in absolute value, assuming $w_1 = 0$. We call this probability the p-value. Roughly speaking, we interpret the p-value as follows: a small p-value indicates that it is unlikely to observe such a substantial association between the predictor and the response due to chance, in the absence of any real association between predictor and the response.

Hence, if we see a small p-value, then we can infer that there is an association between the predictor and the response.

We reject the null hypothesis - that is, we declare relationship to exist between X and Y, if the p-value small enough (typically if p-value is less than 0.05).

Q3. Define the expectation and standard deviation of a random variable. Find the expectation of the following probability distribution. [6 Marks]

X	1	2	3	4	5	6
P(X=x)	1/6	1/3	1/8	1/8	1/8	1/8

Sol:

3.

x	1	2	3	4	5	6
P(x=x)	1/6	1/3	1/8	1/8	1/8	1/8

$$E(x) = \sum_{x} x\, p(x=x)$$

$$= 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{3}\right) + 3\left(\frac{1}{8}\right) + 4\left(\frac{1}{8}\right) + 5\left(\frac{1}{8}\right) + 6\left(\frac{1}{8}\right)$$

$$= \frac{1}{6} + \frac{4}{6} + \frac{1}{8}\left(3+4+5+6\right)$$

$$= \frac{5}{6} + \frac{1}{8}\left(18\right) = \frac{5}{6} + \frac{9}{4} = \frac{10+27}{12}$$

$$= \frac{37}{12}$$

$$\mathrm{var}(x) = \left(1-\frac{37}{12}\right)^2 \frac{1}{6} + \left(2-\frac{37}{12}\right) \frac{1}{3}$$

$$+ \left(3-\frac{37}{12}\right)^2 \frac{1}{8} + \left(\frac{4}{1}-\frac{37}{12}\right)^2 \frac{1}{8}$$

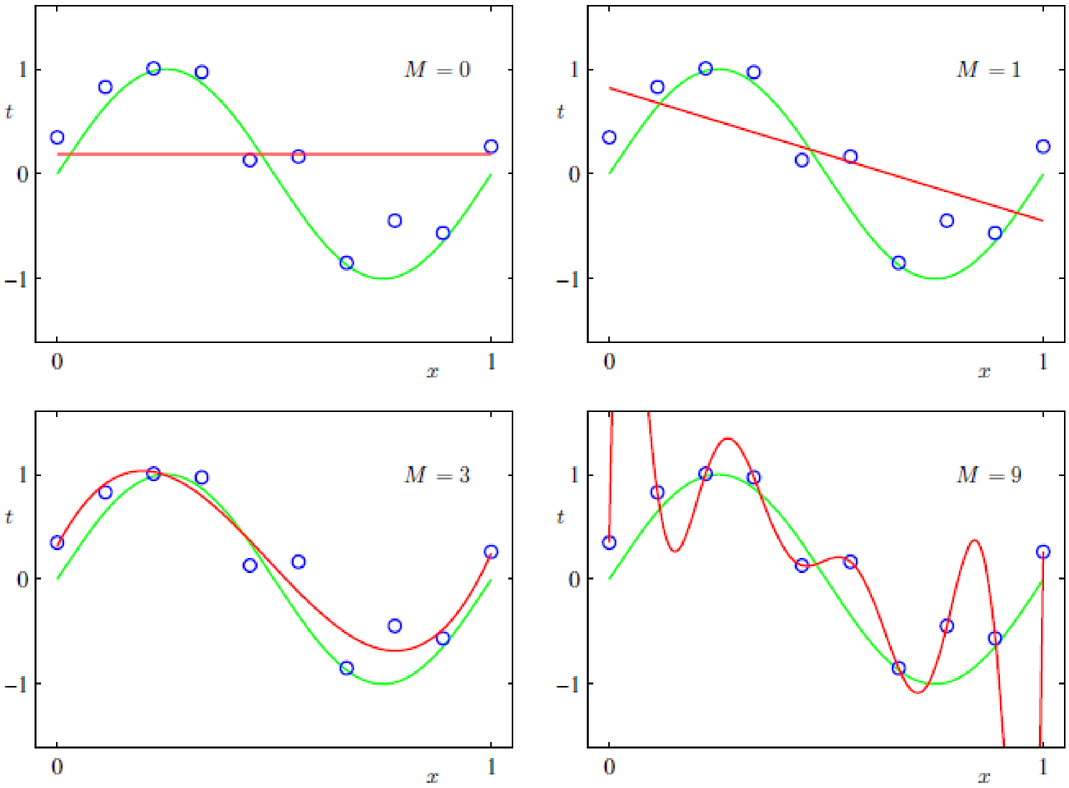
$$+ \left(5-\frac{37}{12}\right)^2 \frac{1}{8} + \left(6-\frac{37}{12}\right)^2 \frac{1}{8}$$

$=$

Q4. What is overfitting? Explain the impact of overfitting on supervised learning with an example. With all mathematical rigor, discuss two methods by which overfitting can be combatted?

[6 Marks]

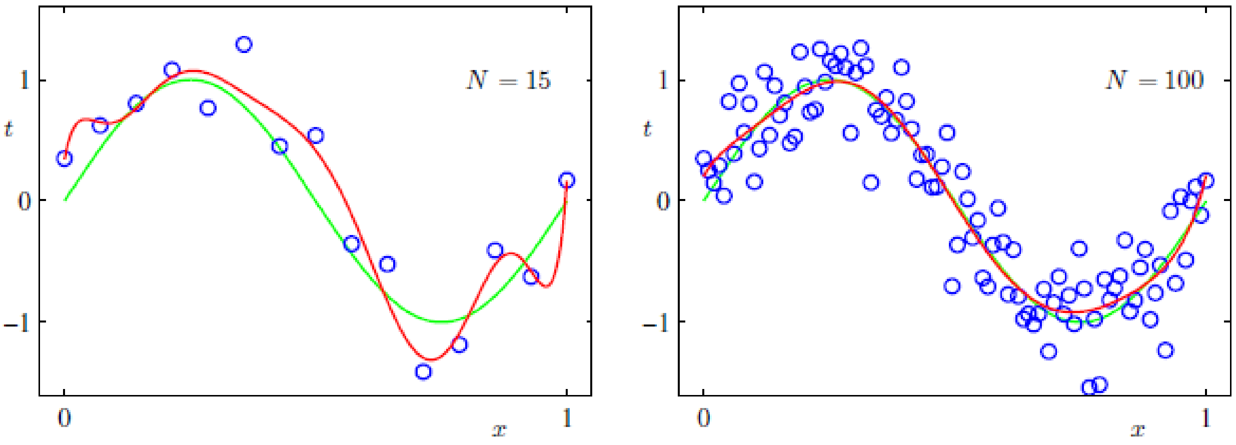
Sol:



We notice that the constant (M = 0) and first order (M = 1) polynomials give rather poor fits to the data and consequently rather poor representations of the function sin(2πx). The third order (M = 3) polynomial seems to give the best fit to the function sin(2πx) of the examples shown in the above. When we go to a much higher order polynomial (M = 9), we obtain an excellent fit to the training data. In fact, the polynomial passes exactly through each data point and $E(w^*) = 0$. However, the fitted curve oscillates wildly and gives a very poor representation of the function sin(2πx). This latter behavior is known as over-fitting.

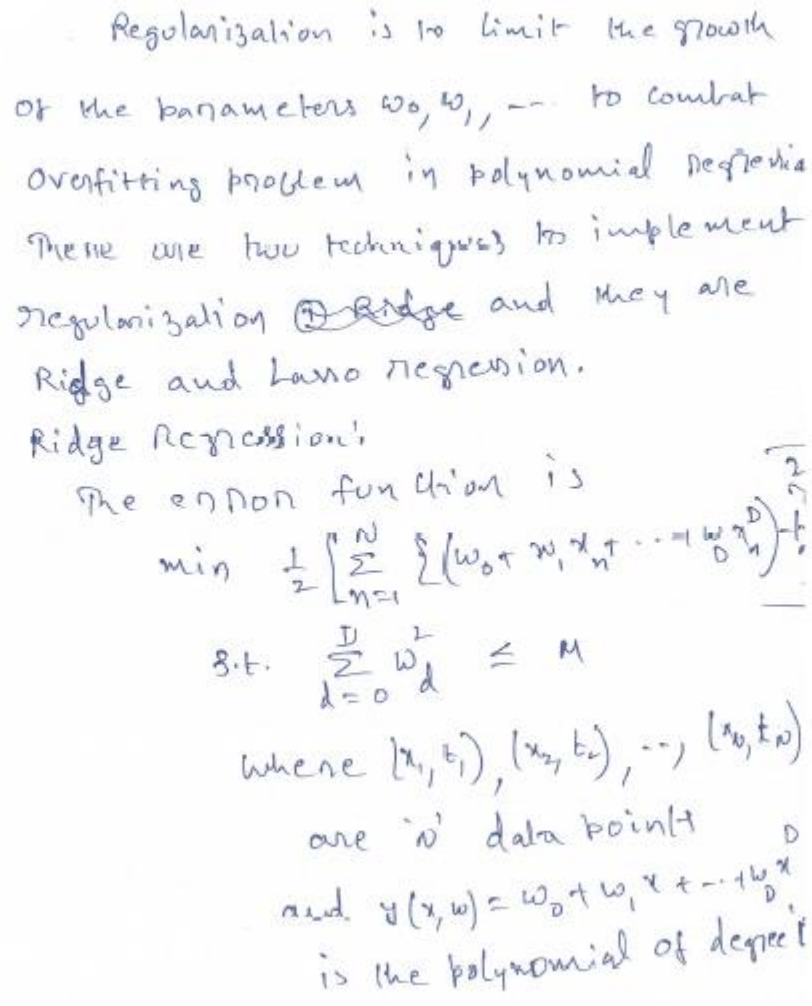
The over-fitting problem can be combatted by the following two methods:

1. Increase the size of training data set:



Plots of the solutions obtained by minimizing the sum-of-squares error function using the M = 9 polynomial for N = 15 data points (left plot) and N = 100 data points (right plot). We see that increasing the size of the data set reduces the over-fitting problem.

2. Regularization to combat overfitting:



Regularization is to limit the growth of the parameters $w_0, w_1, \ldots$ to combat overfitting problem in polynomial regression. There are two techniques to implement regularization ~~① Ridge~~ and they are Ridge and Lasso regression.

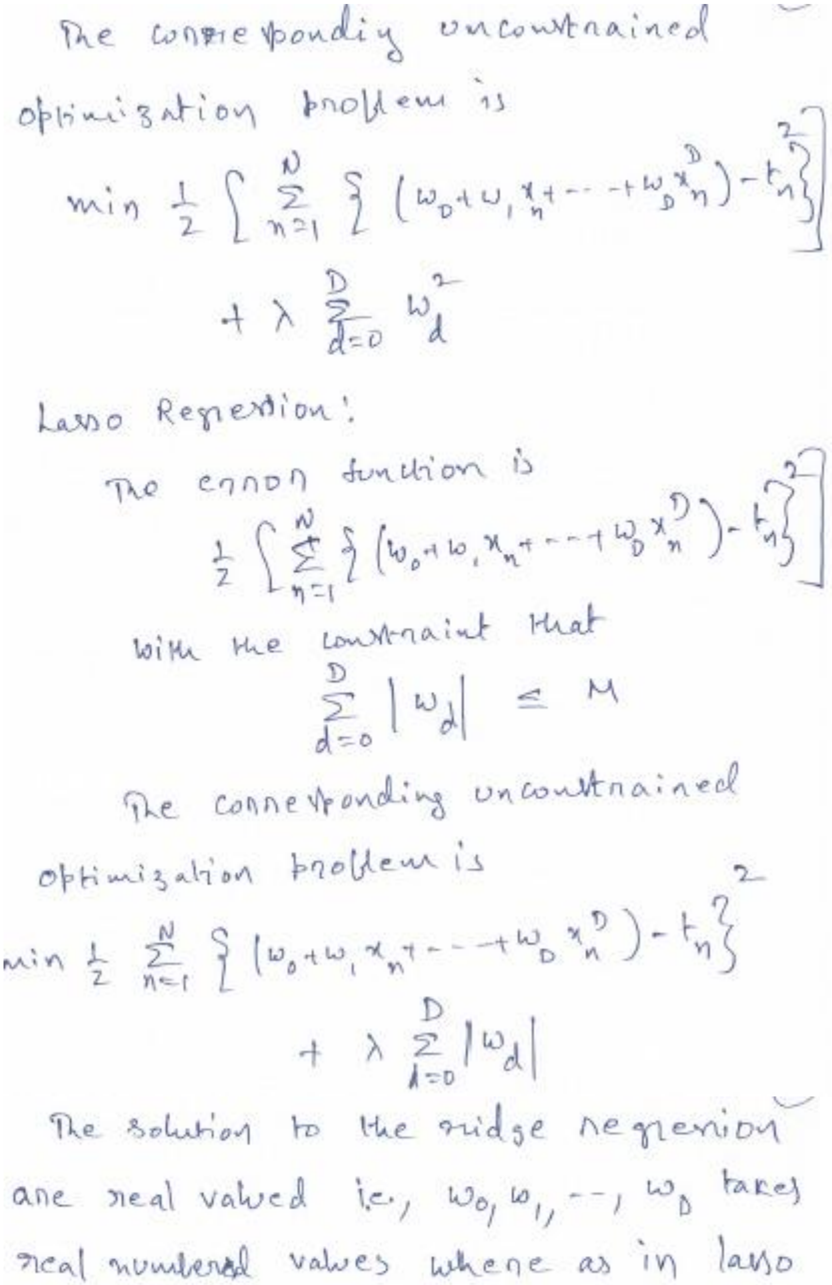
Ridge Regression:

The error function is

$$\min \frac{1}{2}\left[\sum_{n=1}^{N}\left\{(w_0 + w_1 x_n + \cdots + w_D x_n^D) - t_n\right\}^2\right]$$

$$\text{s.t.} \sum_{d=0}^{D} w_d^2 \leq M$$

where $(x_1, t_1), (x_2, t_2), \ldots, (x_N, t_N)$ are 'N' data points and $y(x, w) = w_0 + w_1 x + \cdots + w_D x^D$ is the polynomial of degree



The corresponding unconstrained optimization problem is

$$\min \frac{1}{2}\left[\sum_{n=1}^{N}\left\{\left(w_0+w_1x_n+\cdots+w_Dx_n^D\right)-t_n\right\}^2\right] + \lambda\sum_{d=0}^{D} w_d^2$$

Lasso Regression:

The error function is

$$\frac{1}{2}\left[\sum_{n=1}^{N}\left\{\left(w_0+w_1x_n+\cdots+w_Dx_n^D\right)-t_n\right\}^2\right]$$

with the constraint that

$$\sum_{d=0}^{D}|w_d| \leq M$$

The corresponding unconstrained optimization problem is

$$\min \frac{1}{2}\sum_{n=1}^{N}\left\{\left(w_0+w_1x_n+\cdots+w_Dx_n^D\right)-t_n\right\}^2 + \lambda\sum_{d=0}^{D}|w_d|$$

The solution to the ridge regression are real valued i.e., $w_0, w_1, \ldots, w_D$ takes real numbered values where as in lasso regression they are integer valued.

Q5. Do you think regression model built using gradient descent algorithm, stochastic gradient descent algorithm and batch gradient descent algorithm would be the same? Support you reasoning. [6 Marks]

Sol: The regression models built using gradient descent algorithm, stochastic gradient descent algorithm and batch gradient descent algorithm might be different. The initial seed for these algorithms need not be the same and hence all three models will be distinct. Though initial seed is the same for all these algorithms, weights will be updated differently at the end of the first iteration and hence weight updates will be different with each iteration leading to distinct final weights and models.

Q6. Can the learning rate (in gradient descent algorithms), eta, be any random value? What are the consequences of choosing random value for eta? [4 Marks]
Sol: If eta is taken to be a relatively larger value (say 1000) then gradient descent algorithm might not converge as change in the dependent variable is high with each iteration. If eta is taken to be too small value (say 0.0000001) then then change in the dependent variable is considerable small and hence it takes many more iterations to converge. Hence a moderate value is to be chosen for eta.