CS 335 & CS 337

Adarsh Raj - 190050004

Assignment-1

August 2021

Question 1

Theory - 1.1

Expression for $\nabla mse(w, b)$: Now, we have:

$$mse(w, b) = \frac{1}{N} \sum_{i=1}^{N} \big((wx_i + b) - y\big)^2$$

Also,

$$\nabla mse(w, b) = \begin{pmatrix} \frac{\partial mse(w, b)}{\partial w} & \frac{\partial mse(w, b)}{\partial b} \end{pmatrix}^T$$

Now,

$$\frac{\partial mse(w, b)}{\partial w} = \frac{1}{N} \sum_{i=1}^{N} 2x_i\big((wx_i + b) - y\big)^2$$

$$\frac{\partial mse(w, b)}{\partial w} = \frac{2}{N} \sum_{i=1}^{N} x_i((wx_i + b) - y\big)^2$$

And,

$$\frac{\partial mse(w, b)}{\partial b} = \frac{1}{N} \sum_{i=1}^{N} 2\big((wx_i + b) - y\big)^2$$

$$\frac{\partial mse(w, b)}{\partial b} = \frac{2}{N} \sum_{i=1}^{N} \big((wx_i + b) - y\big)^2$$

Therefore, we have:

$$\nabla mse(w, b) = \begin{pmatrix} \frac{2}{N} \sum_{i=1}^{N} x_i((wx_i + b) - y\big)^2 & \frac{2}{N} \sum_{i=1}^{N} \big((wx_i + b) - y\big)^2 \end{pmatrix}^T$$

Lab - 1.2

(a)

Refer to code file `assignment_1.ipynb`.

(b)

Refer to code file `assignment_1.ipynb`.

(c)

- **Single Variable Gradient Descent:**

  Max iterations used = 10000
  Learning rate used = 0.001

```
Validation loss is 1.9641491793128651
Training Loss loss is 1.8365229381774275
```

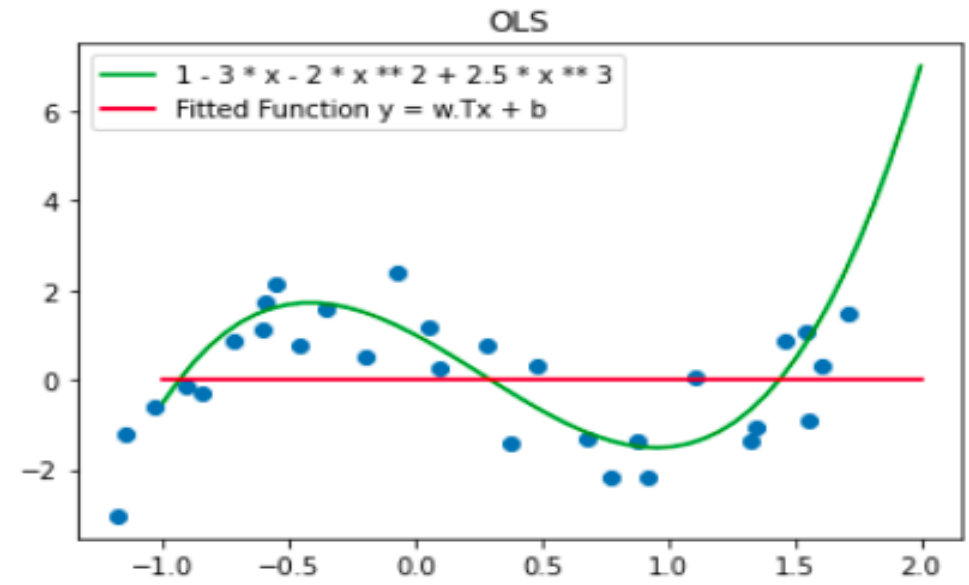

Figure 1: Single Variable Gradient Descent Fitted Function

- **Single Variable Closed Form:**

  Closed form can be expressed as:

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

  Where, X is appended by a column of 1's. W will be a (2x1) column vector, and $w = W[0]$ and $b = W[1]$.

```
------------------------------------------------------------
------
Validation loss is 1.960101492433602
Training Loss loss is 1.8227058521869255
```

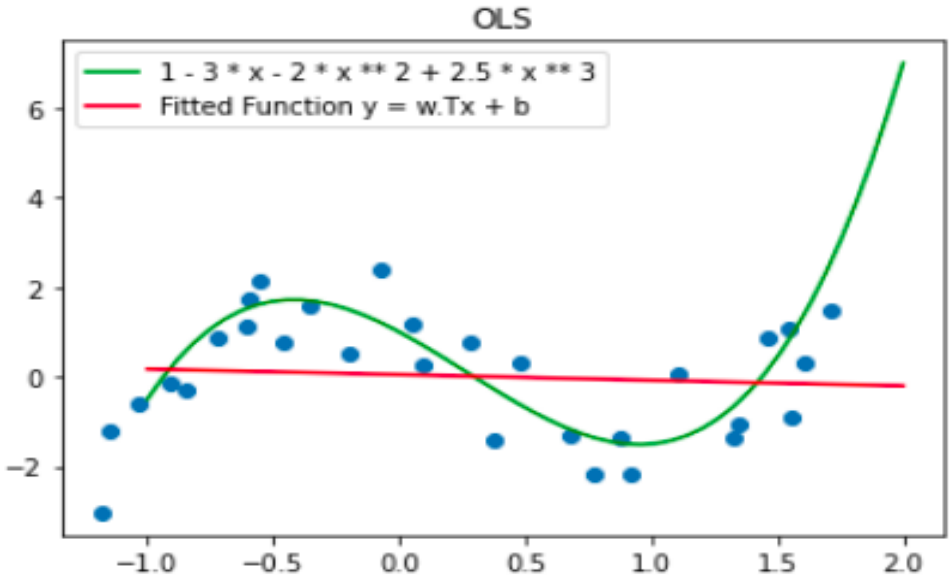

Figure 2: Single Variable Closed Form Fitted Function

(d)

It is **not** possible to obtain a solution using `singlevar_grad()` such that its training loss is strictly less than that of the solution obtained by `singlevar_closedform()`.

This is because the closed form gives the optimal solution if it exists and that is the global minima. Also, the gradient descent converges towards local minima depending on the learning rate and number of iterations. Since for single variable linear regression, we only have one minima which is represented by the closed form, the gradiant descent will always tend towards the global minima.

Therfore, as gradient descent always tends towards the closed form solution, the training loss for it can never be smaller than that of the closed form.

Question 2

Theory - 2.1

(a) Given, the number of samples as N and data matrix $X(N \times d)$.

For response variable y, we have the model:

$$\hat{y} = w_1x_1 + w_2x_2 + ... + w_dx_d + \beta$$

Where, β is the bias term.

To represent the above in matrix form, we can append a column of only 1's to matrix X, for the bias. Now, the predicted output $\hat{Y}$ can be expressed as:

$$\hat{Y} = XW$$

(b) Given, the minimum squared error loss function as:

$$mse = \frac{1}{N}\sum_{i=1}^{N}(\hat{y} - y)^2$$

Equivalent formulation of above in matrix form is as follows:

$$mse = \frac{1}{N}||\hat{Y} - Y||^2 = \frac{1}{N}||XW - Y||^2$$

Now,

$$\frac{\partial mse}{\partial W} = \frac{\partial}{\partial W}\left(\frac{1}{N}||XW - Y||^2\right)$$

Using the fact that, $\frac{\partial||Ax+b||}{\partial x} = 2A^T(Ax + b)$, where A is a matrix and x and b are column vectors, we have:

$$\frac{\partial mse}{\partial W} = \left(\frac{2}{N}X^T(XW - Y)\right)$$

(c) Give, for ridge regression, loss function is expressed as:

$$mse = \frac{1}{N}\sum_{i=1}^{N}(\hat{y} - y)^2 + \lambda||W||^2$$

In matrix form, we have:

$$mse = \frac{1}{N}||\hat{Y} - Y||^2 + \lambda||W||^2 = \frac{1}{N}||XW - Y||^2 + \lambda||W||^2$$

Now, from vector calculus, we have $\frac{\partial||Ax+b||}{\partial x} = 2A^T(Ax+b)$ and $\frac{\partial||W||^2}{\partial W} = 2W$, hence the expression for $\frac{\partial mse}{\partial W}$ is:

$$\frac{\partial mse}{\partial W} = \frac{2}{N}X^T(XW - Y) + 2\lambda W$$

(d) For data matrix X (appended by a column of ones for bias term), the closed form expression for OLS is:

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

The above equation will not have any solution if the matrix being inverted is not invertible, i.e, $(X^T X)$ is non invertible. Therefore it can be deduced that X is not fully column rank (all columns are not linearly independent) and its determinant will be zero.

As gradient descent is a first-order iterative algorithm for finding a local minimum of a differentiable function and it takes steps of small size in the opposite direction of gradient, it always converges to a local minimum. In this case too, it will converge to a local minimum depending on the parameters including number of epochs and learning rate.

Lab - 2.2

(a) Refer to code file `assignment_1.ipynb`.

(b)
- **Multi Variable Gradient Descent:**
  Max iterations used = 20000 and Learning rate used = 0.001

```
Validation loss if 0.7809731952659338
Training Loss loss if 0.5079821693811677
```

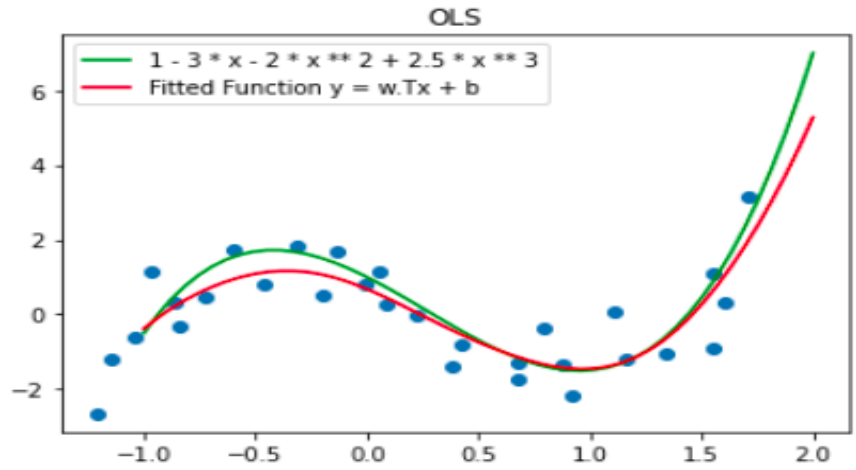

Figure 3: Multi Variable Gradient Descent Fitted Function

- **Multi Variable Closed Form:**

```
------
Validation loss if 0.7789309011457897
Training Loss loss if 0.47300430018197204
```

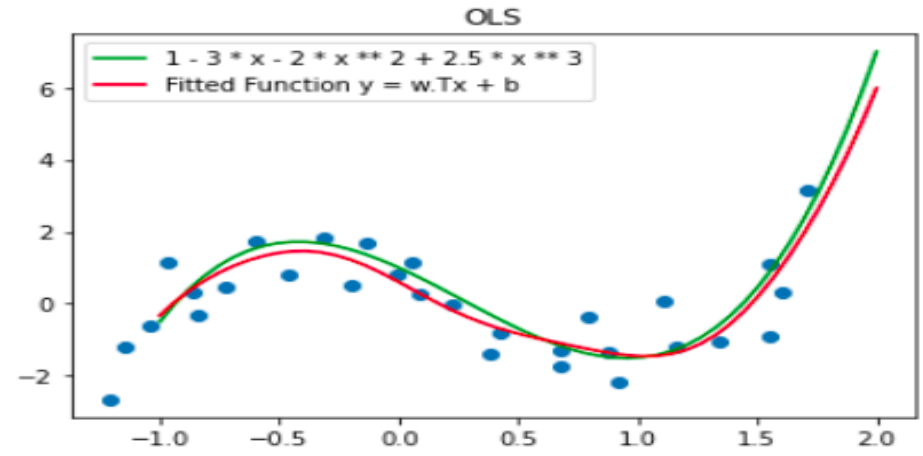

Figure 4: Multi Variable Closed Form Fitted Function

(c) • **Ridge Regression Gradient Descent:**
Max iterations used = 20000 and Learning rate used = 0.001
lambda(λ) = 0.01

```
Validation loss if 0.9034410514220805
Training Loss loss if 0.6327770244023692
```

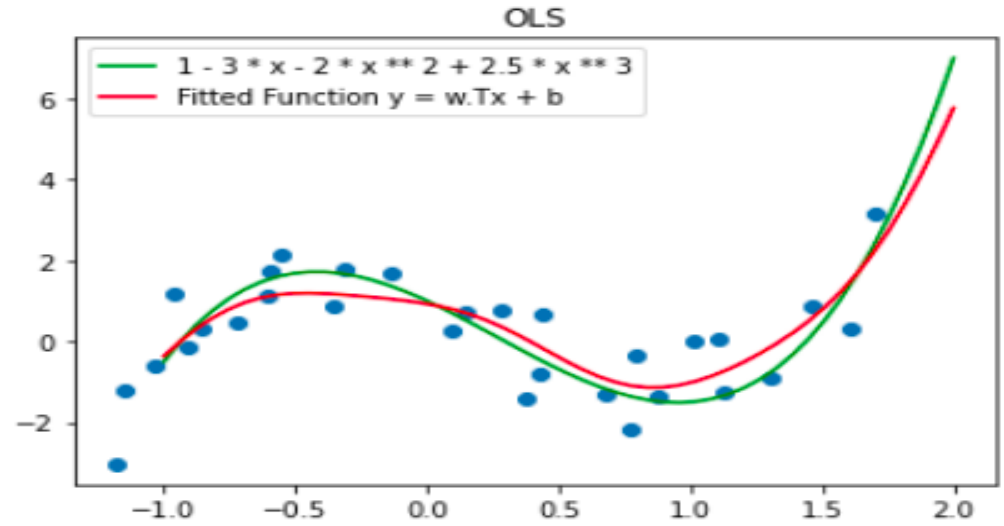

Figure 5: Ridge Regression Gradient Descent Fitted Function

• **Ridge Regression Closed Form:**

```
Validation loss if 0.9743196638683792
Training Loss loss if 0.6489435234832814
```

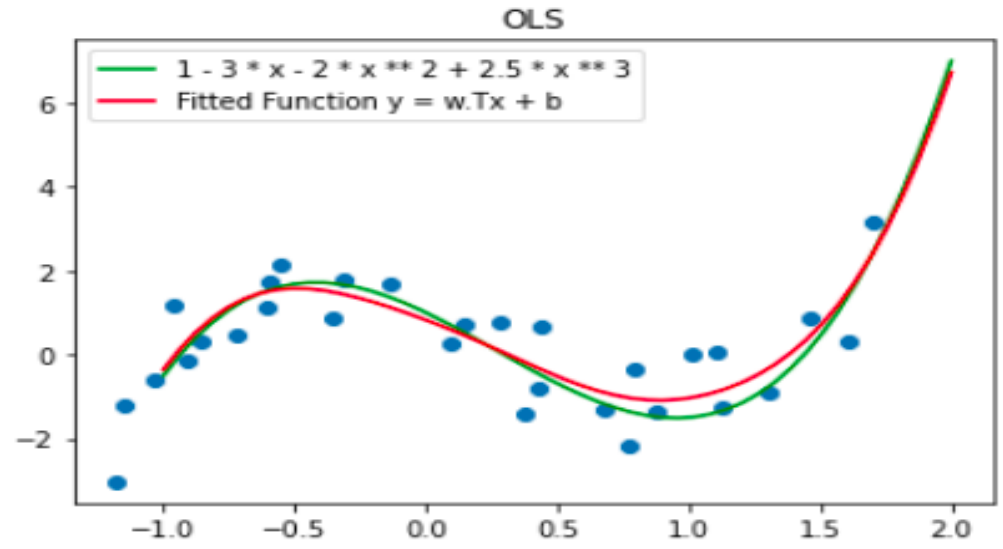

Figure 6: Ridge Regression Closed Form Fitted Function

Theory - 3.1

(a)

Given, prior distribution of w is a Gaussian with mean ($\mu = \mu_o$) and variance ($\sigma^2 = 1$). Hence, the prior distribution $p(w)$ (pdf of a gaussian distribution) can be expressed as:

$$p(w) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(w-\mu)^2}{2\sigma^2}\right)$$

$$\implies p(w) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(w-\mu_o)^2}{2}\right)$$

(b)

Assumption: Data likelihood is Gaussian, i.e, $p(y|x;w) = \mathcal{N}(wx, 1)$.

(c)

Given that the dataset $\mathcal{D}$ is independent and identically distributed, we have:

$$p(\mathcal{D}|w) = \prod_{i=1}^{N} p((x_i, y_i)|w) = \prod_{i=1}^{N} p(y_i|x_i; w)$$

Also,

$$\prod_{i=1}^{N} p(y_i|x_i; w) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(y_i - wx_i)^2}{2}\right)$$

$$\implies p(\mathcal{D}|w) = \frac{1}{(\sqrt{2\pi})^N} \exp\left(\sum_{i=1}^{N}\left(\frac{-(y_i - wx_i)^2}{2}\right)\right)$$

(d)

By Bayes' theorem, we have:

$$p(w|\mathcal{D}) = \frac{p(\mathcal{D}|w)p(w)}{p(\mathcal{D})}$$

Also,

$$p(\mathcal{D}) = \int_{-\infty}^{\infty} p(\mathcal{D}|w)p(w)dw$$

$$\implies p(w|\mathcal{D}) = \frac{p(\mathcal{D}|w)p(w)}{\int_{-\infty}^{\infty} p(\mathcal{D}|w)p(w)dw}$$

$$\implies p(w|\mathcal{D}) = \frac{\frac{1}{(\sqrt{2\pi})^N}\exp\left(\sum_{i=1}^{N}\left(\frac{-(y_i-wx_i)^2}{2}\right)\right)\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-(w-\mu_o)^2}{2}\right)}{\int_{-\infty}^{\infty}\frac{1}{(\sqrt{2\pi})^N}\exp\left(\sum_{i=1}^{N}\left(\frac{-(y_i-wx_i)^2}{2}\right)\right)\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-(w-\mu_o)^2}{2}\right)dw}$$

On cancelling out the constants in numerator and denominator, and combining the exponential terms we have:

$$p(w|\mathcal{D}) = \frac{\exp\left(-\frac{1}{2}\big(\sum_{i=1}^{N}(-(y_i-wx_i)^2)+(-(w-\mu_o)^2)\big)\right)}{\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2}\big(\sum_{i=1}^{N}(-(y_i-wx_i)^2)+(-(w-\mu_o)^2)\big)\right)dw}$$

(e)

Using the conjugate prior assumption to calculate $p(w)$ analytically, and hence ignoring the denominator term.

(f)

On simplifying the numerator, we have: (Equation 1)

$$p(w|\mathcal{D}) \propto \exp\left(-\frac{1}{2}\left[w^2\big(\sum_i xi^2+1\big)-2w\big(\sum_i y_ix_i+\mu_o\big)+\sum_i y_i^2+\mu_o^2\right]\right)$$

(g)

Posterior is also gaussian by virtue of conjugate prior. Hence, (Equation 2)

$$p(w|\mathcal{D}) = \mathcal{N}(\mu_N,\sigma_N^2) \propto \exp\big(-\frac{1}{2\sigma_N^2}(w-\mu_N)^2\big)$$

(h)

Formulating equation 1 above as follows:

$$p(w|\mathcal{D}) \propto \exp\left(-\frac{1}{2}\big(\sum_i xi^2+1\big)\left[w^2-2w\frac{\big(\sum_i y_ix_i+\mu_o\big)}{\big(\sum_i xi^2+1\big)}+C\right]\right)$$

Where, C is a constant term not dependent on w.
Now, from equation 2, we have:

$$p(w|\mathcal{D}) \propto \exp\big(-\frac{1}{2\sigma_N^2}(w^2-2w\mu_N+\mu_N^2)\big)$$

Comparing the above equations for μ_N and σ_N, we have:

$$-\frac{1}{2\sigma_N^2} = -\frac{1}{2}\Big(\sum_i xi^2 + 1\Big)$$

$$\implies \sigma_N^2 = \frac{1}{\big(\sum_i xi^2 + 1\big)}$$

Similarly,

$$\frac{\mu_N}{\sigma_N^2} = \sum_i x_i y_i + \mu_o$$

$$\mu_N = \frac{\sum_i x_i y_i + \mu_o}{\sum_i xi^2 + 1}$$

Therefore we have,

$$p(w|\mathcal{D}) = \mathcal{N}\left(\frac{\sum_i x_i y_i + \mu_o}{\sum_i xi^2 + 1}, \frac{1}{\sum_i xi^2 + 1}\right)$$

(i)

As $N \to \infty$, we have infinite number of data samples, hence, the summation of all data points, i.e $\sum_i x_i$, will tends towards infinity.

Therefore we have:

$$\lim_{N\to\infty} \sigma_N^2 = \lim_{N\to\infty} \frac{1}{\sum_i xi^2 + 1} = 0$$

Also, as N increases, both $\sum_i x_i$ and $\sum_i x_i y_i$ becomes infinitely large. hence we have:

$$\lim_{N\to\infty} \mu_N = \lim_{N\to\infty} \frac{\sum_i x_i y_i + \mu_o}{\sum_i xi^2 + 1} = \frac{\sum_i x_i y_i}{\sum_i xi^2}$$

(j)

As $N \to \infty$, intuitively we can observe the following:

As we have a large number of data samples, the prior information about w becomes irrelavant and hence μ_o disappears from the expression μ_N. Also, the expression for μ_N tends towards MLE estimate of w, without a prior distribution.

For variance, due to large number of data samples, the Maximum A Posteriori Estimate becomes more and more significant. In thr limit $N \to \infty$, we are absolutely confident about w, hence the variance of posterior distribution tends to zero.

MLE Estimate - 3.2

(a)

MLE estimate is defined as, $w^* = argmax_w p(\mathcal{D}|w)$. Now, let the log likelihood distribution of w be represented as $LL(w)$. Now, we have:

$$LL(w) = \log(p(\mathcal{D}|w)) = -\log(\sqrt{2\pi}) - \frac{1}{2}\sum_{i=1}^{N}(wx_i - y)^2$$

To find the maximum likelihood estimate, we need the derivative of $p(\mathcal{D}|w)$ to be zero or, the derivative of the log likelihood to be zero, i.e $\frac{\partial LL(w)}{\partial w} = 0$.Hence, $w^* = argmax_w LL(w)$. So, we have:

$$LL^{'}(w) = \frac{\partial LL(w)}{\partial w} = 0$$

$$\implies 0 - \sum_i x_i(wx_i - y_i) = 0$$

$$\implies w^* \sum_i x_i^2 = \sum_i x_i y_i$$

$$w^* = \frac{\sum_i x_i y_i}{\sum_i x_i^2}$$

The above w^* is the MLE estimate for w.

(b)

When $N \to \infty$, the Bayesian Estimate, i.e MAP estimate(μ_N) approaches towards the MLE estimate. This is because, as the data samples increases our belief of w, that it belongs to some prior distribution decreases and hence the prior information becomes irrelevant, i.e the value of μ_o decreases compared to the summation terms. Hence at limit infinity, the term vanishes from the equation of MAP estimate.

Since, MLE is one of the estimators which achieves the lease MSE among different estimators, the MAP estimate converges to the MLE at $N \to \infty$.

Lab - 3.3

(a)

Code of `bayesian_lr()`, refer to the code file `assignment_1.ipynb`.
Plot:

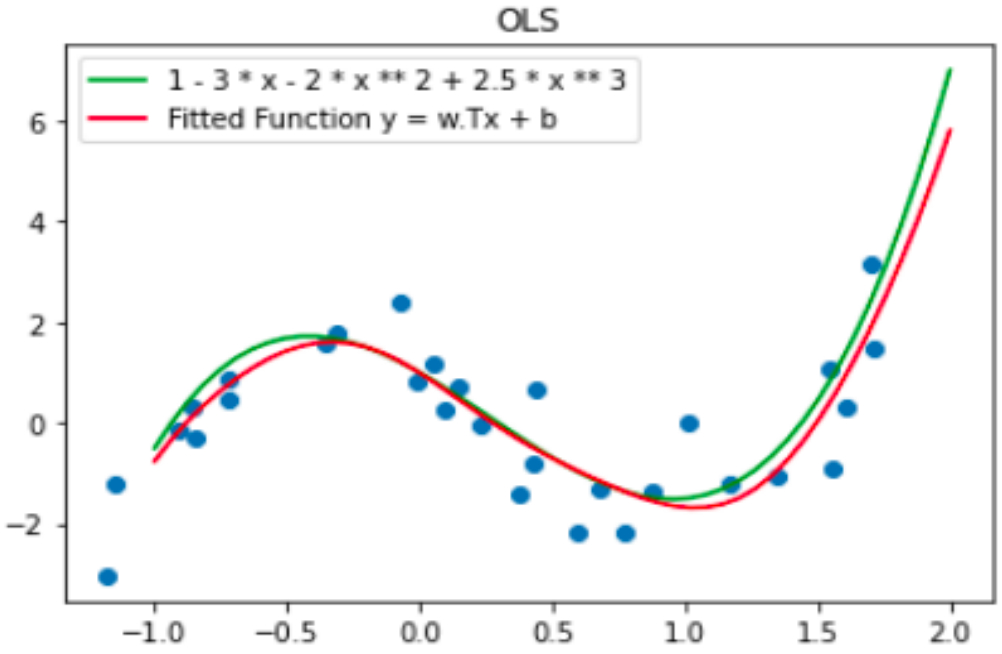

Figure 7: Bayesian Regression Fitted Function

(b)

The parameters of Bayesian posterior estimate for W are:
$\mu_N = \sum_N(\sum_0^{-1}\mu_0 + \frac{1}{\sigma^2}X^Ty)$ and $\sum_N = (\sum_0^{-1} + \frac{1}{\sigma^2}X^TX)^{-1}$

Now, $\sum_o$ is easily invertible as it equals $0.5I$, where I is the identity matrix. Also, for $\sum_N$, it can be proved that the matrix $(\sum_0^{-1} + \frac{1}{\sigma^2}X^TX)$ is always invertible.

Note that the proof of invertibility of matrix $M = (X^TX + \lambda I)$ suffices for above. Now, M is the sum of (symmetric) positive semidefinite matrices, so it is itself positive semidefinite. Moreover, since M is a positive semidefinite matrix, it is invertible if and only if it is positive definite. That is, it suffices to show that $v^TMv > 0$ whenever v is a non-zero vector.

$$v^TMv = v^T(X^TX + \lambda I)v = v^TX^TXv + \lambda v^TAv = \|Xv\|^2 + \lambda\|Av\|^2 > 0$$

Hence, it is concluded that matrix M is always invertible under given conditions. Therefore, $\sum_N$ can always be calculated. Therefore, we will not encounter the problem as of question 2.1 (d).

Conclusion 4

Bayesian regression gives the best estimate and takes much less time to run compared to other regression methods. This is because of the small number of data samples provided, hence matrix calculations are pretty fast while calculating the bayesian posterior distribution. This observation can vary depending on the number of training data samples.

Comparing training and validation loss for single variable, multi variable and ridge regression, multi variable regression gives minimum training loss and validation loss in most of the scenarios (random sampling of data), for about 20000 epochs and learning rate (= 0.001). Epochs were increased to fit data more.

Although, ridge regression performs better on large number of iterations (say, 50000) than multi variable regression. Also, the lambda can be tuned ($\lambda < 0.01$) to increase the ridge regression performance. All the closed form solutions perform better than their gradient descent counterparts.

Plots are attached above pertaining to each of the regression analysis mentioned.