CS5691 Pattern Recognition and Machine Learning
Assignment 2
Submitted By: Ganesh S (ME20B070)

1 Question 1

Common notations used :
$X \in \mathbb{R}^{n \times d}$ is the matrix where each row is a data point, n is the number of data points, d is the number of features, $y \in \mathbb{R}^n$ is the vector of labels, $w \in \mathbb{R}^d$ is the vector of weights.

1.1 Closed form solution

The least squares problem is formulated as an estimation problem in **w** and solved using maximum likelihood estimation.

$$w_{ML} = (X^T X)^{-1} X^T y$$

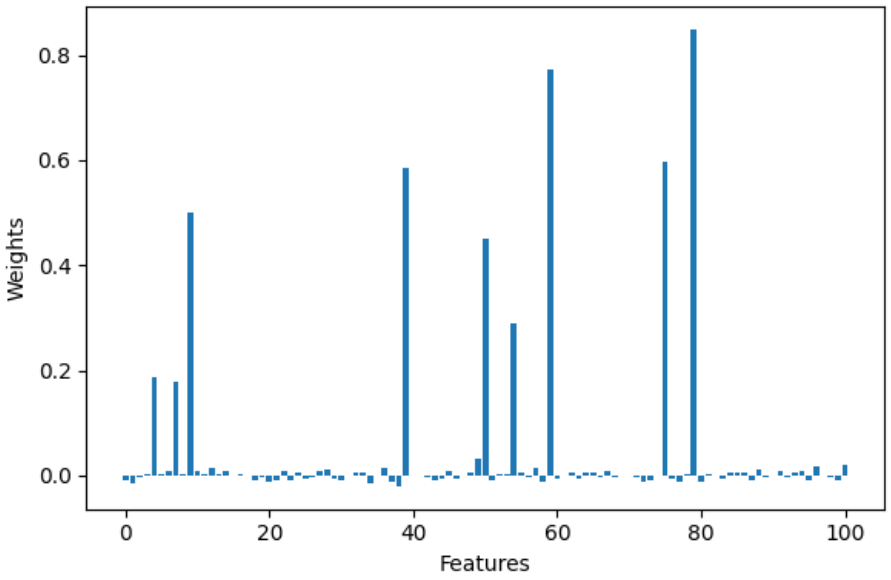


Figure 1: Weights vs Features (denoted by its index starting from 0)

The matrix X was appended with a column of one's to account for biases. But as you can see, the weight corresponding to the bias (index = 100) is very small. Thus we can conclude that the effect of bias in this dataset is minimal. The error in train dataset was **396.85** and test dataset was **185.38**

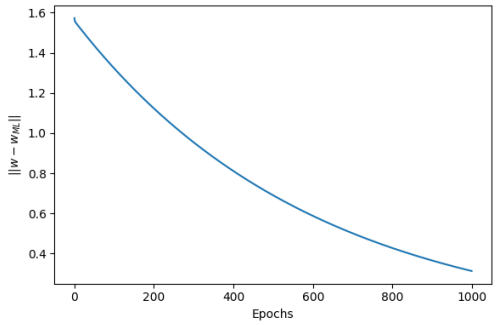
1.2 Gradient descent

All the weights are initialized as 0 and then, gradient descent is performed.

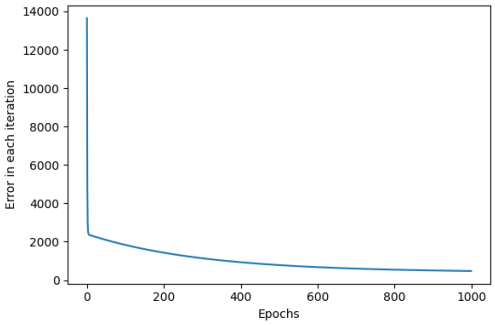
$$w^{t+1} = w^t - \frac{2\alpha}{n}((X^T X)w^t - X^T y)$$

Where α is the learning rate. For this dataset the learning rate was chosen to be 0.01. After doing 1000 iterations the error in the train dataset was **472.51** and in test dataset was **155.55**.

From the above graph we observe that the difference between the weights in each iteration and the weights obtained by maximum likelihood estimation reduces after each iteration. During gradient descent, the parameter moves to the minima of the function that is being optimized. Similarily the weights after each iteration w^t move towards the minima of the least squares function i.e. w_{ML}



(a) $||w^t - w_{ML}||$



(b) $||Xw^t - y||^2$

Figure 2: Variation of $||w^t - w_{ML}||$ and error in each iteration

Now, the step-size was taken to be variable. We take the step-size η is given as a square-summable function :

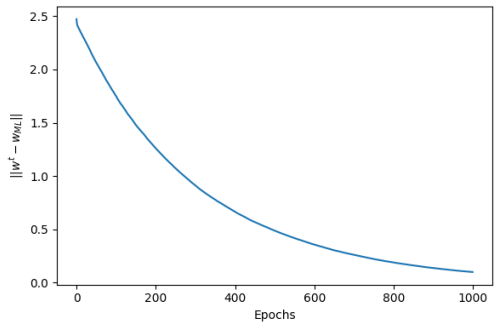
$$\eta^t = \frac{\alpha}{t}$$
$$w^{t+1} = w^t - \frac{2\alpha}{nt}((X^T X)w^t - X^T y)$$

The error peaked at the start and then dipped down quickly. This performed worse in the train dataset and gave an error of **569.49**. But it performed marginally better in the test dataset by giving an error **142.72**.

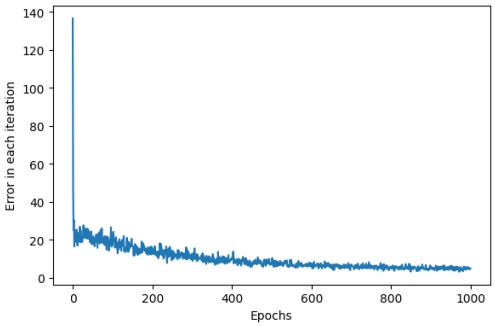
1.3 Stochastic gradient descent

Stochastic gradient using a batch-size of 100 was performed. At each iteration 100 points are randomly sampled from the train-data set $(\tilde{X})$

$$w^{t+1} = w^t - \frac{2\alpha}{n}((\tilde{X}^T \tilde{X})w^t - \tilde{X}^T y)$$



(a) $||w^t - w_{ML}||$



(b) $||Xw^t - y||^2$

Figure 3: Variation of $||w^t - w_{ML}||$ and error in each iteration

We observe that the error flucuates during each iteration but follows a decreasing trend. This is due to the random nature of the stochastic gradient descent algorithm. The error in train dataset is **477.40** and the error in the test dataset is **155.78** similar to that in gradient descent algorithm.
But this algorithm is way quicker since computation is performed on $\tilde{X} \in R^{100\times101}$ instead of $X \in R^{10000\times101}$

2 Question 2

2.1 Ridge Regression

The objective function for ridge regression is:

$$f(w) = ||Xw - y||^2 + \lambda||w||^2$$

Gradient descent is performed as:

$$w^{t+1} = w^t - \frac{2\alpha}{n}\left(\left(X^TX + \lambda I\right)w^t - X^Ty\right)$$

Where λ is a hyper parameter and I is the identity matrix $\in R^{d\times d}$

2.2 Cross Validation

5-fold cross validation was done to obtain the optimal λ. 100 linearly separated samples of λ was generated between 0.1 and 10. The data set is split into 5 folds and in each iteration, one of the fold is taken for validation and other 4 folds are taken for training. This is done for all the folds and average error in validation set is obtained. The error in the validation set is compared for different λ and compared to find the optimum λ.

The weights are given by the closed form solution:

$$w = \left(X^TX + \lambda I\right)^{-1}X^Ty$$

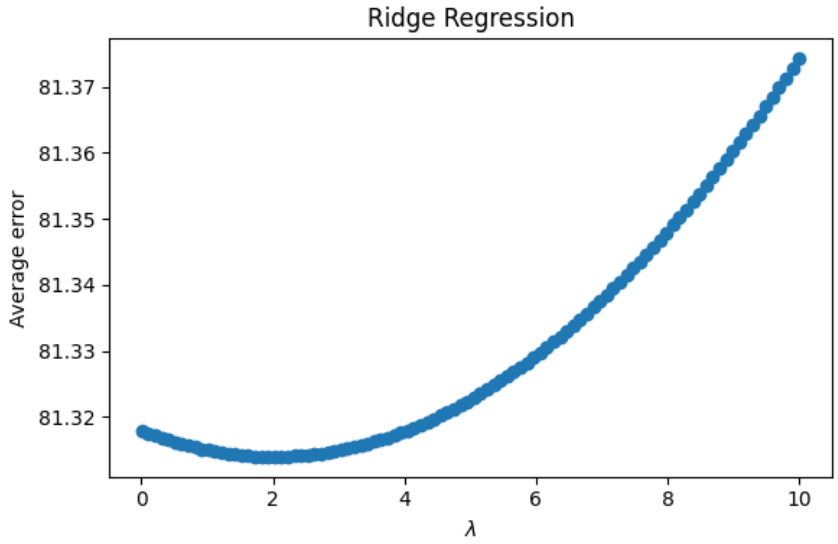
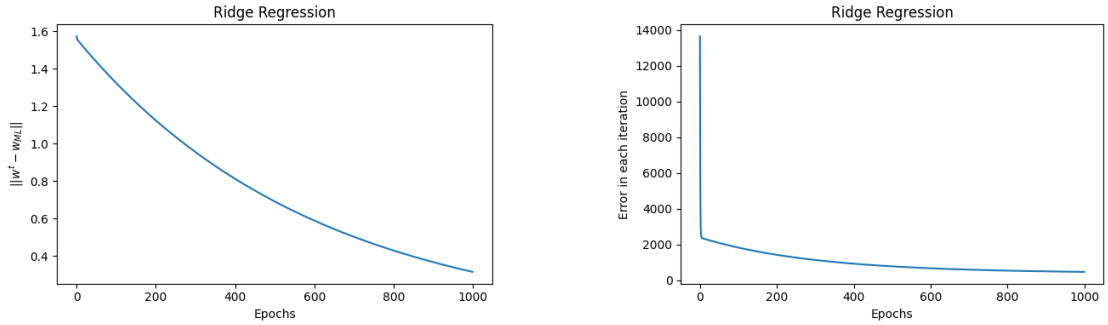


Figure 4: Variation of the average error in the validation set with λ

The optimum λ was found to be **2.028**.



(a) $||w^t - w_{ML}||$ (b) $||Xw^t - y||^2$

Figure 5: Variation of $||w^t - w_{ML}||$ and error in each iteration in Ridge regression

The test error for w_R using $\lambda = 2.028$ after 1000 iterations is is **155.39**. The test error for w_{ML} is **185.38** The train error was minimum in w_{ML} because it is the solution to maximum likelihood estimation. This doesn't guarantee that it will perform better in the test dataset. The expected error i.e. the mean squared error,

$$\mathbb{E}[(\hat{w} - w)^2] = Var(\hat{w}) + (\mathbb{E}[\hat{w}] - w)^2$$

By using ridge regression, we reduce the variance of our estimator by increasing its bias, in the hope of reducing the mean squared error. This works out so the error is lower in the test dataset for w_R

3 Conclusions

The train and test errors for various methods is given below:

Method	**Train Error**	**Test error**
Closed Form	398.85	185.38
Gradient Descent	472.51	155.55
Gradient Descent(with $\frac{\alpha}{t}$ as learning rate)	569.49	142.72
Stochastic Gradient Descent	477.40	155.78
Ridge Regression with $\lambda = 2.028$	473.43	155.59

The test error was found to be lowest for gradient descent followed by ridge regression. Since gradient descent is just an optimization method to solve the least squares problem (maximum likelihood estimation), the error due to gradient descent might have been affected by the initial conditions. It is safe to say that on average, ridge regression gives the best results.