DA5400- FML -Assignment 1

- Preethi MM21B051

Question 1

We add a column vector of ones to the X matrix to account for bias, while computing W vector using least squares.

The analytical solution for W_ML is given by:

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

Upon calculating the MSE on the test data set we get it to be '66.256'.

Question 2

We set the learning rate (step size) to be 0.001, and max number if iterations to be 1000. We see that W_t converges within 10 iterations for a tolerance of 1e-9.

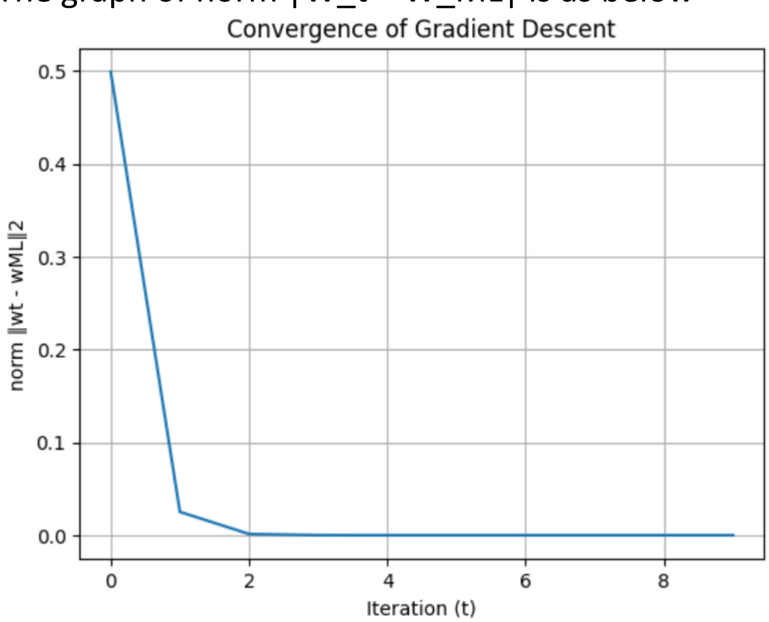
The gradient is given by:

$$\nabla = X^T(Xw - y) + \lambda w$$

And weights are update in each iteration as:

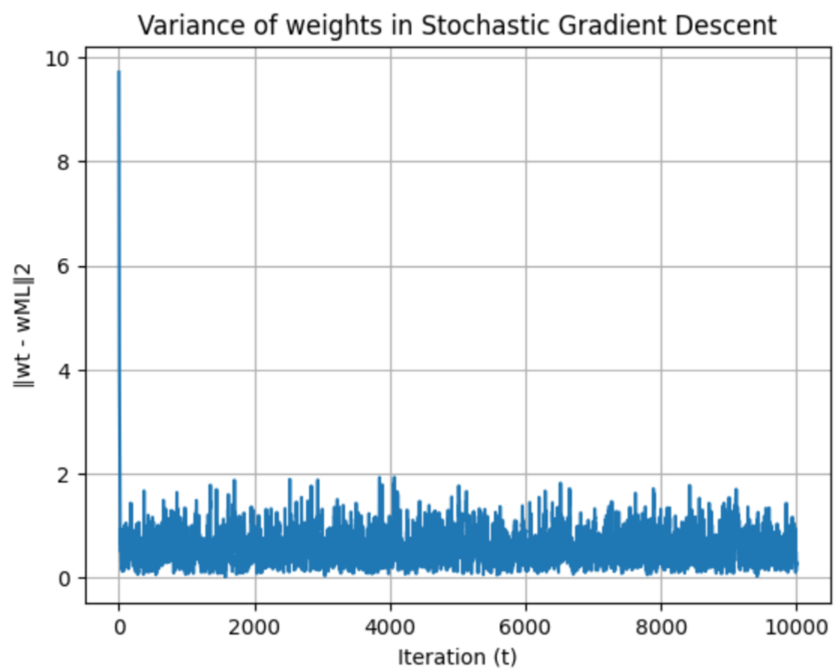
$$w_{t+1} = w_t - \eta \nabla.$$

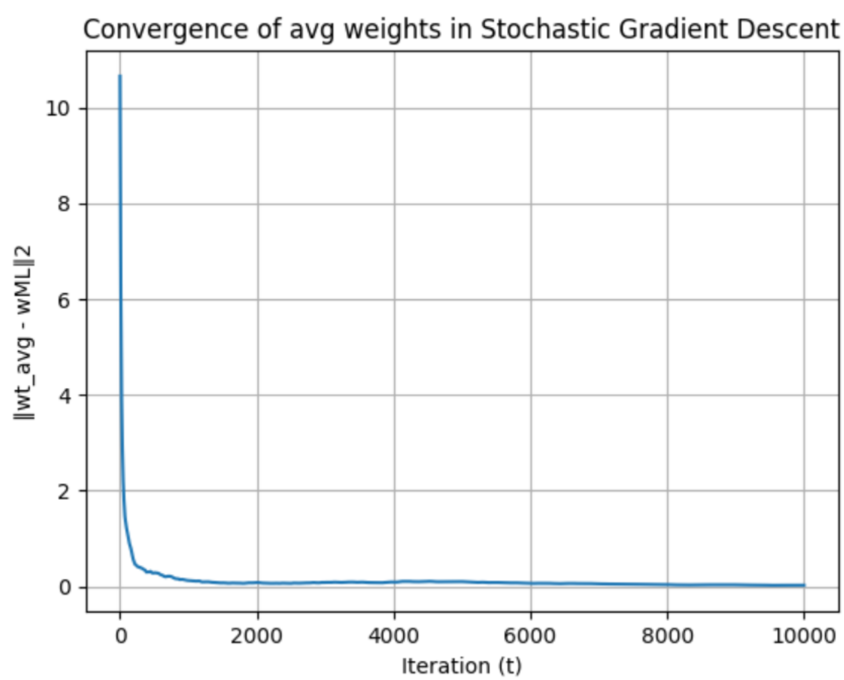
The graph of norm |W_t – W_ML| is as below



Question 3

We repeat the gradient descent process, but instead of using the entire data set each time we randomly choose 100 points and use them to compute the W_t in each iteration. The W_t itself doesn't converge due to the variance in the dataset, however the avg(W_t) converges to W_ML as can be seen in the graphs below:





Question 4

We split the data set into 80% training and 20% as validation set. We take various λ values along the scale and find weights using ridge regression.

$$\nabla = X^T(Xw - y) + \lambda w$$
$$w_{t+1} = w_t - \eta(X^T(Xw_t - y) + \lambda w_t)$$

We find the λ value with gives the lowest MSE on the validation dataset and find its corresponding weights.

Minimum validation error: 160.75 on the validation data set

Corresponding λ value: 0.0031

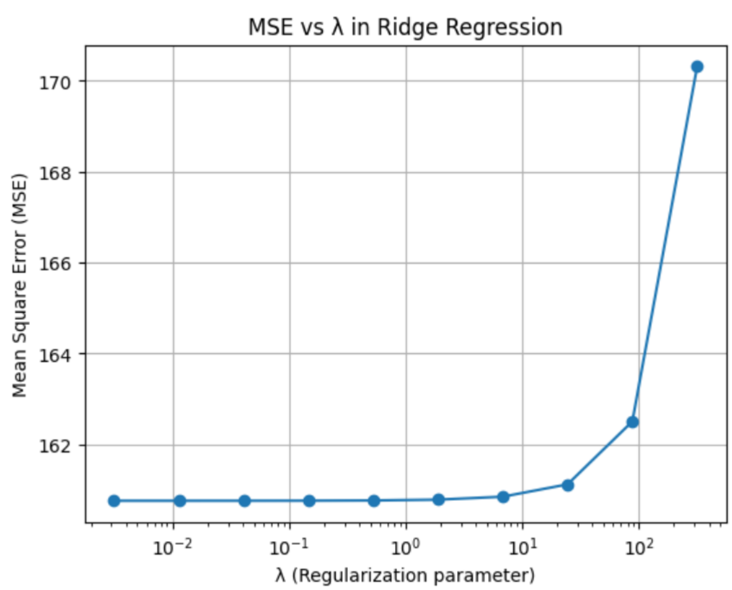
We get our best λ value to be very close to 0, hence when we perform ridge regression for the best lambda it is very close to the gradient descent itself. The gradient and update condition for ridge regression with λ =0, then becomes, which is that of gradient descent.

$$\nabla = X^T(Xw - y) + \lambda w$$
$$w_{t+1} = w_t - \eta\nabla.$$

Therefore, the weights and MSE on the test data set are also similar as below:

W_R for the best λ is [9.89745678 1.76772023 3.52383375]

The minimum MSE value for Ridge Regression is 66.25. The MSE with linear regression is 66.25.



Question 5

The quadratic kernel is chosen for this dataset because the feature-target relationship appears to be quadratic. The kernel is defined as:

$$K(x_1, x_2) = (x_1 \cdot x_2 + 1)^2$$

In kernel regression, the solution for the learned parameter (α) is:

$$\alpha = (K + \lambda I)^{-1} y$$

where (K) is the kernel matrix, (λ) is a regularization parameter, and (y) is the target. Predictions are made as:

$$\hat{y}(x_{\text{test}}) = \sum_{i=1}^{N} \alpha_i K(x_{\text{test}}, x_i)$$

The quadratic kernel is a better fit because it captures the nonlinear, second-order relationship in the data, leading to significantly lower mean squared error (MSE) compared to standard least squares regression. The addition of (λ) ensures the kernel matrix is invertible, addressing numerical instability issues.