# ELL409 - Assignment 1 Report

Priyal Jain, 2021MT60949

August 2024

## 1 Objective

The aim is to clean the data provided and then build a Linear Regression model to fit on the data, minimizing the least squares loss using algorithms of gradient descent and stochastic gradient descent, implemented from scratch.

## 2 Assumptions

- The convergence criterion is fixed at $\epsilon = 1 \times 10^{-5}$ .
- Total number of epochs is taken to be 10,000 .
- Outliers are removed using interquartile range method with IQR factor = 1.5 .
- Normalization has been done to speed up the convergence.
- Moving average window size is taken to be $k = 10$ .
- Given data is split into training (80%) and validation (20%) sets.
- Decaying learning rate is incorporated in SGD to achieve convergence in a finite number of epochs.

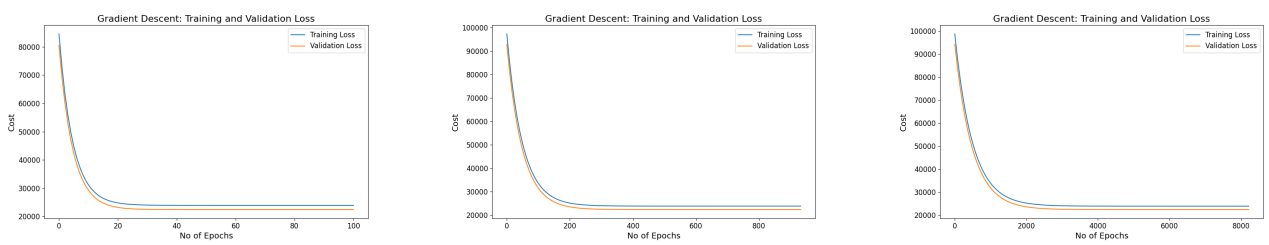
## 3 Gradient Descent

### 3.1 Results

| Learning Rate | Iterations to converge | Time(s) | Final Parameters | Training Loss | Validation Loss |
|---|---|---|---|---|---|
| 0.001 | 8213 | 2.21 | [126.84, 237.55] | 23890.33 | 22395.17 |
| 0.01 | 932 | 0.53 | [126.87, 237.58] | 23890.33 | 22395.25 |
| 0.1 | 100 | 0.04 | [126.88, 237.59] | 23890.33 | 22395.28 |

Table 1: Gradient Descent Results with Different Learning Rates

### 3.2 Training and Validation Loss Plots




(a) Learning Rate: 0.1 (b) Learning Rate: 0.01 (c) Learning Rate: 0.001

Figure 1: Batch size is entire dataset

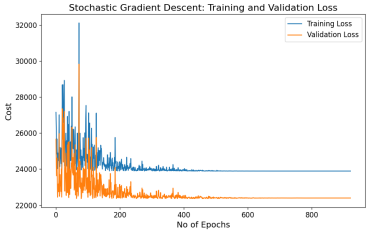
# 4 Stochastic Gradient Descent

## 4.1 Results

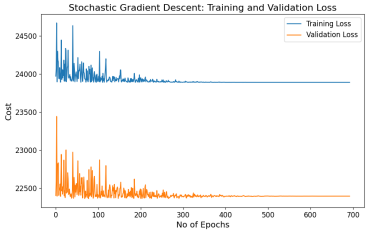
| Learning Rate | Batch Size | Time (s) | Iterations | Final Parameters | Training Loss | Validation Loss |
|---|---|---|---|---|---|---|
| 0.1 | 1 | 42.00 | 921 | [126.90, 237.59] | 23890.33 | 22395.25 |
| 0.1 | 10 | 3.02 | 688 | [126.89, 237.60] | 23890.33 | 22395.28 |
| 0.1 | 100 | 0.41 | 462 | [126.88, 237.60] | 23890.33 | 22395.31 |
| 0.1 | 1000 | 0.13 | 269 | [126.89, 237.60] | 23890.33 | 22395.30 |
| 0.01 | 1 | 25.13 | 692 | [126.88, 237.60] | 23890.33 | 22395.29 |
| 0.01 | 10 | 2.01 | 466 | [126.88, 237.60] | 23890.33 | 22395.31 |
| 0.01 | 100 | 0.41 | 263 | [126.88, 237.61] | 23890.33 | 22395.32 |
| 0.01 | 1000 | 0.24 | 807 | [126.58, 237.23] | 23890.58 | 22394.60 |
| 0.001 | 1 | 180.03 | 4592 | [126.89, 237.60] | 23890.33 | 22395.31 |
| 0.001 | 10 | 9.77 | 2166 | [126.89, 237.59] | 23890.33 | 22395.26 |
| 0.001 | 100 | 3.037 | 5723 | [126.57, 237.22] | 23890.59 | 22394.60 |
| 0.001 | 1000 | 0.17 | 223 | [126.58, 237.23] | 23890.56 | 22394.61 |

Table 2: Stochastic Gradient Descent Results for different batch sizes for each learning rate

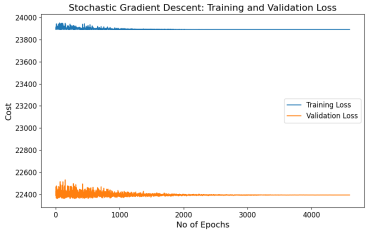
## 4.2 Training and Validation Loss Plots




(a) Learning Rate: 0.1

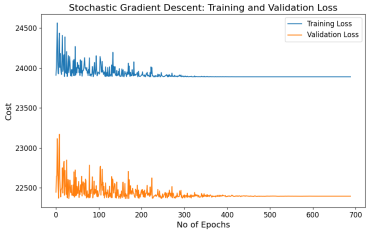



(b) Learning Rate: 0.01

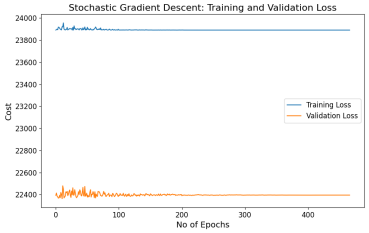



(c) Learning Rate: 0.001

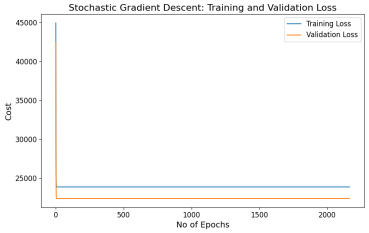
Figure 2: Batch size 1




(a) Learning Rate: 0.1

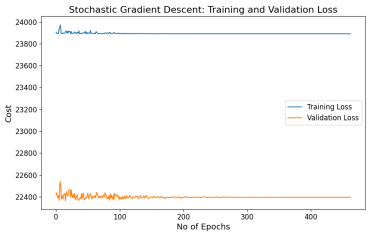



(b) Learning Rate: 0.01

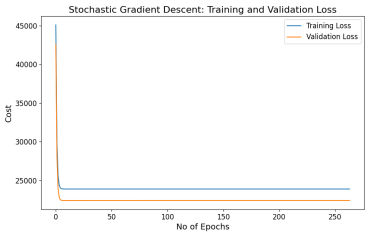



(c) Learning Rate: 0.001

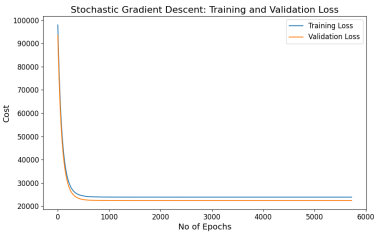
Figure 3: Batch size 10




(a) Learning Rate: 0.1

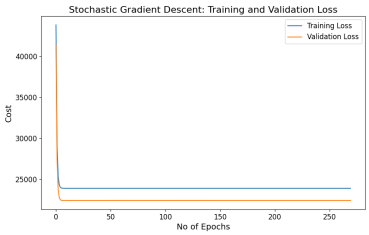



(b) Learning Rate: 0.01

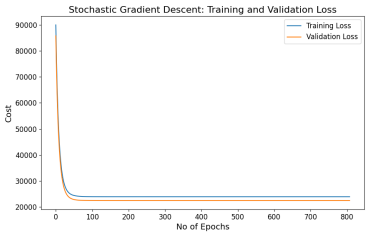



(c) Learning Rate: 0.001

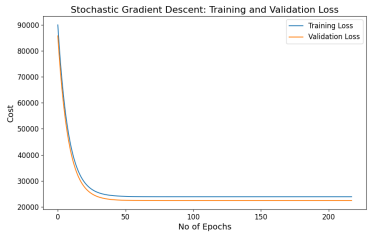
Figure 4: Batch size 100




(a) Learning Rate: 0.1




(b) Learning Rate: 0.01




(c) Learning Rate: 0.001

Figure 5: Batch size 1000

# 5 Observations

- As learning rate is decreased, more iterations are required for convergence.
- As batch size is increased, the time taken to converge reduces.
- For SGD, time for convergence is very high because the cost values oscillate.
- Learning rate decay after every epoch:
  - When $\text{lr} \geq 0.01$, $\text{lr} = \text{lr} \times 0.99$ is used.
  - When $\text{lr} < 0.01$, $\text{lr} = \text{lr} \times 0.999$ is used.
- For outlier detection, Interquartile range method gives better results here as compared to z-score because the data is not normally distributed and has some extreme values.