CS 4774 Homework #1 Writeup

Linear Algebra Review

(1.1) $x_1 = -1$; $x_2 = 0$; $x_3 = 1$; (solved using Gaussian Elimination, but I won't show the steps here).

(1.2) $$\begin{bmatrix} 2 & 2 & 3 \\ 1 & -1 & 0 \\ -1 & 2 & 1 \end{bmatrix} \begin{bmatrix} x1 \\ x2 \\ x3 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix}$$

(1.3) Rank of A is 3, since the row echelon form has no rows of zeroes

(1.4) $$A^{-1} = \begin{bmatrix} 1 & -4 & -3 \\ 1 & -5 & -3 \\ -1 & 6 & 4 \end{bmatrix}; \qquad \det(A) = -1$$

(1.5) $$A^{-1}Ax = A^{-1}b \Rightarrow x = A^{-1}b \Rightarrow x = \begin{bmatrix} 1 & -4 & -3 \\ 1 & -5 & -3 \\ -1 & 6 & 4 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$$

(1.6) $$[-1 \quad 0 \quad 1] . \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix} = 1$$

(1.7) $L_1 = 4;\ L_2 = 6; L_\infty = 2$

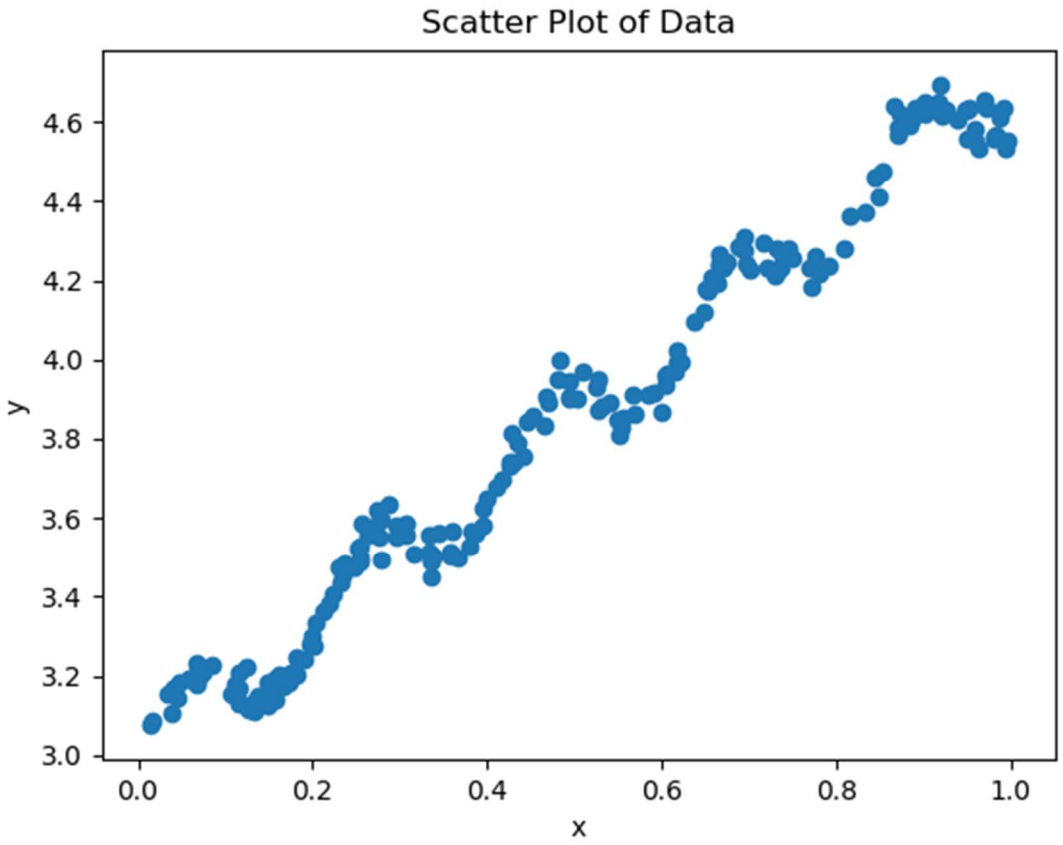
(1.8) $$\begin{bmatrix} 2 & 2 & 3 \\ 1 & -1 & 0 \\ -1 & 2 & 1 \\ -1 & 2 & 1 \end{bmatrix} \begin{bmatrix} x1 \\ x2 \\ x3 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \\ 2 \\ 1 \end{bmatrix}$$

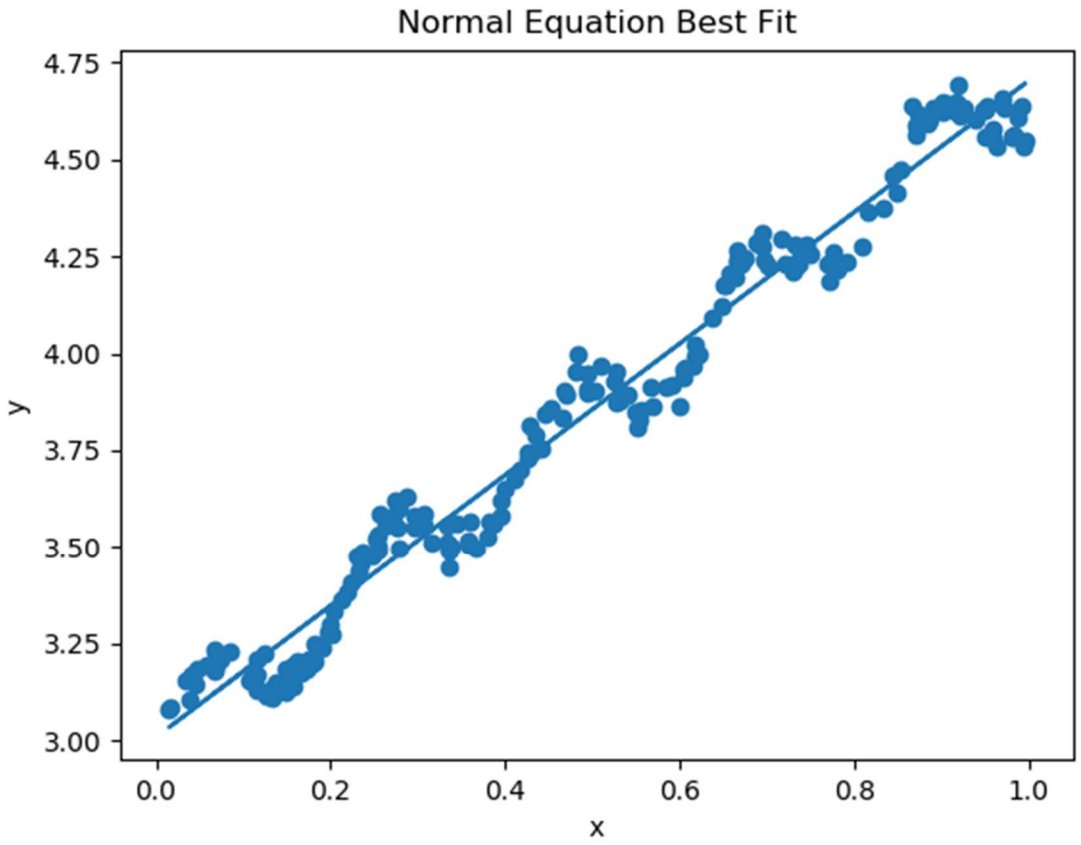
(1.9) Rank of A1 is 3

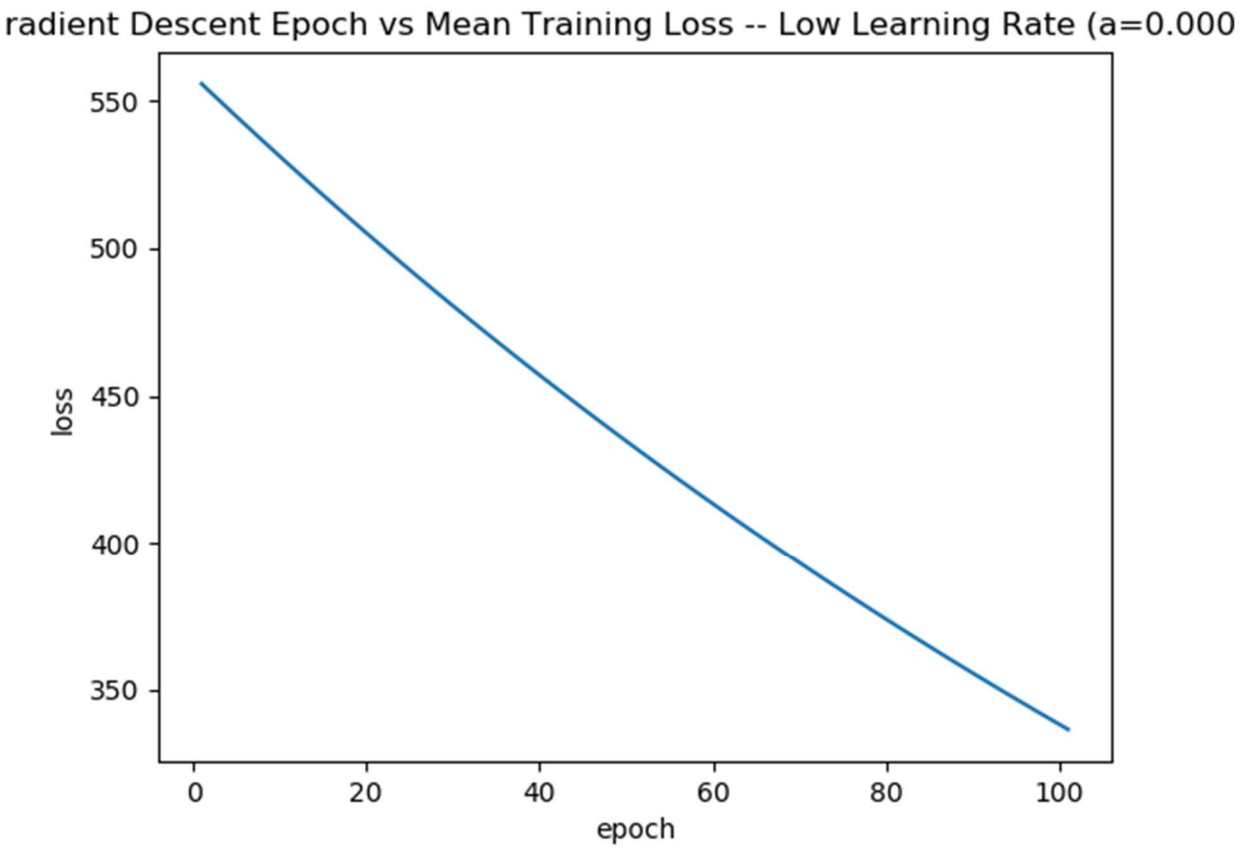
(1.10) No. the system is over-constrained.

Linear Regression Model Fitting

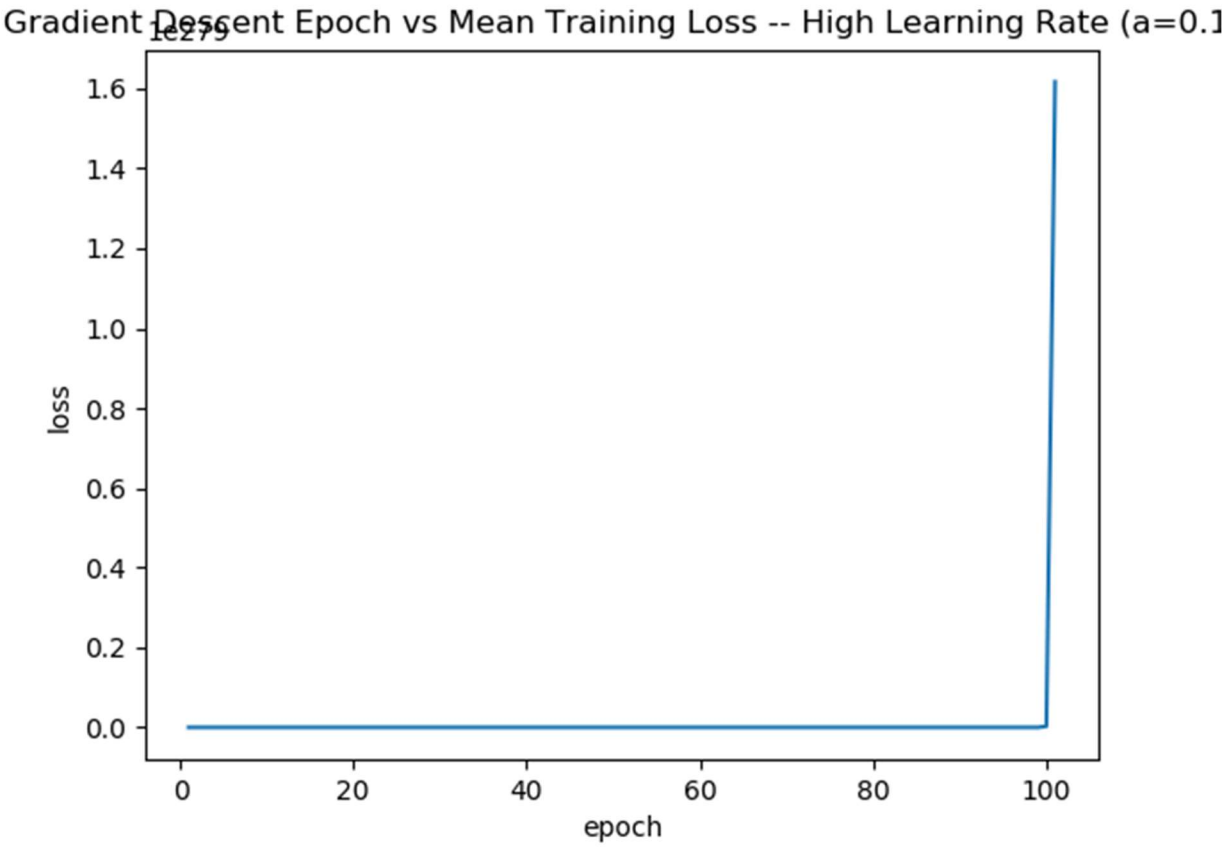
The requested graphs are on the following pages, with written responses after. The learning rate is denoted "a" and found in the title of each graph where requested, and the batch size is denoted "b", similarly found.


Scatter Plot of Data
4.6
4.4
4.2
4.0
3.8
3.6
3.4
3.2
3.0
y
0.0
0.2
0.4
0.6
0.8
1.0
x

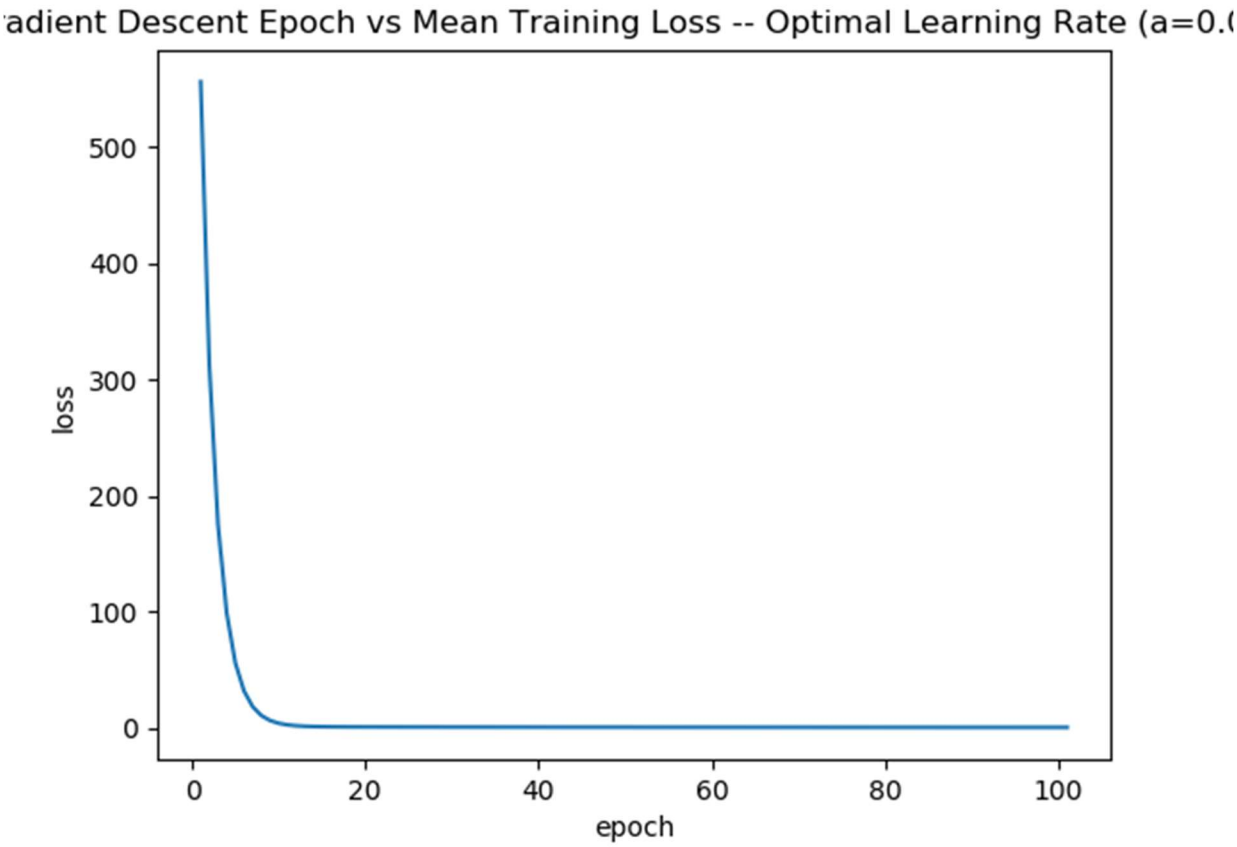

Normal Equation Best Fit
4.75
4.50
4.25
4.00
3.75
3.50
3.25
3.00
y
0.0
0.2
0.4
0.6
0.8
1.0
x


radient Descent Epoch vs Mean Training Loss -- Low Learning Rate (a=0.000
550
500
450
400
350
loss
0
20
40
60
80
100
epoch

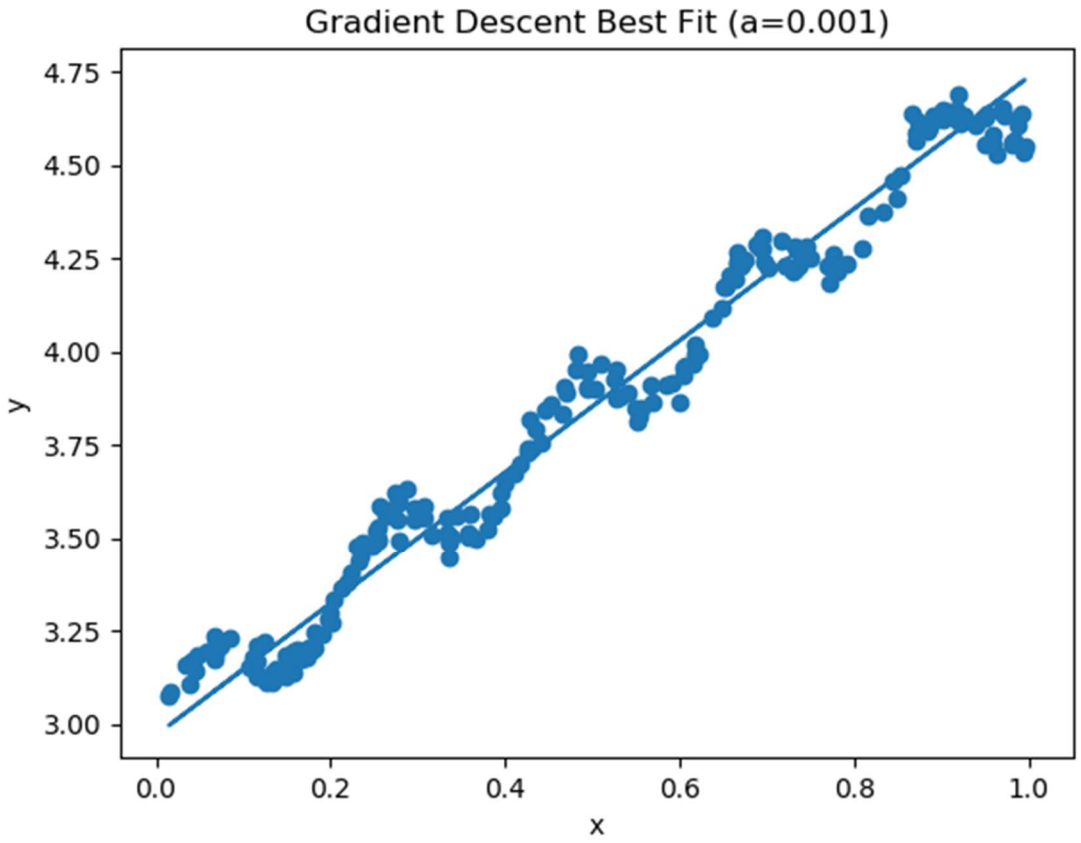
a=0.00001

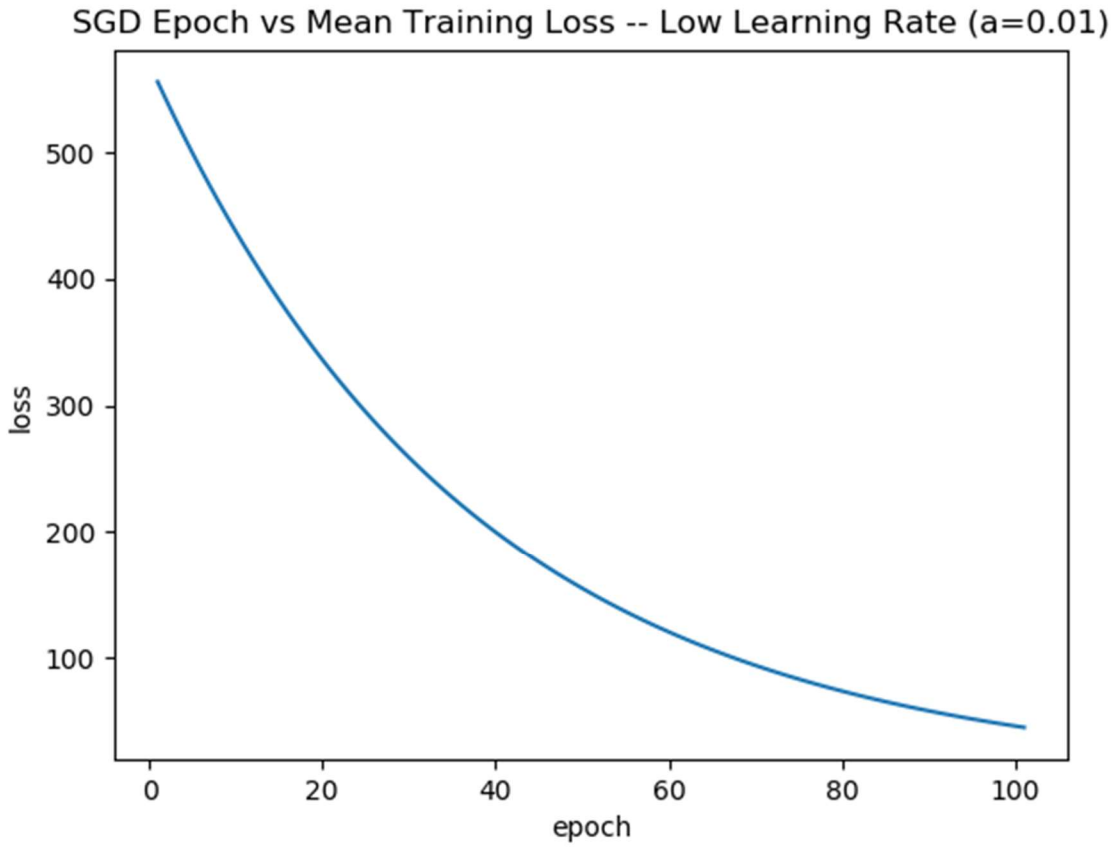

Gradient Descent Epoch vs Mean Training Loss -- High Learning Rate (a=0.1
1e279
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
loss
0
20
40
60
80
100
epoch

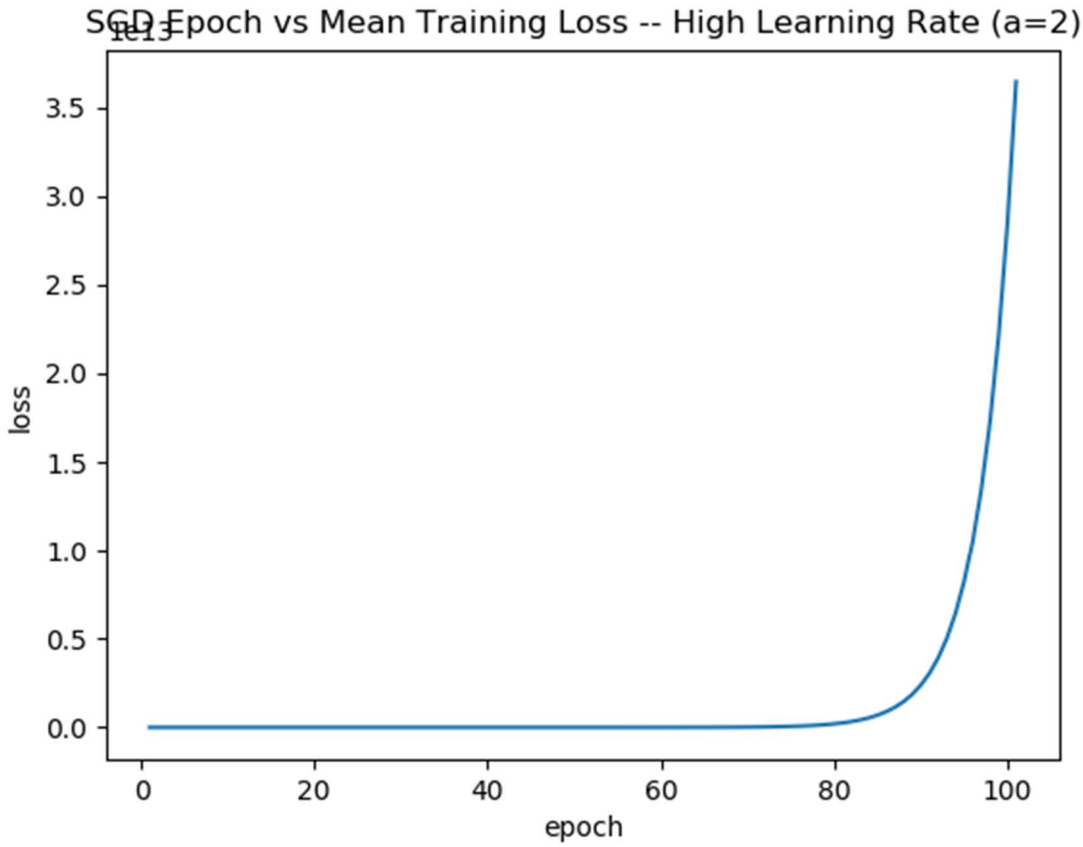
a=0.1

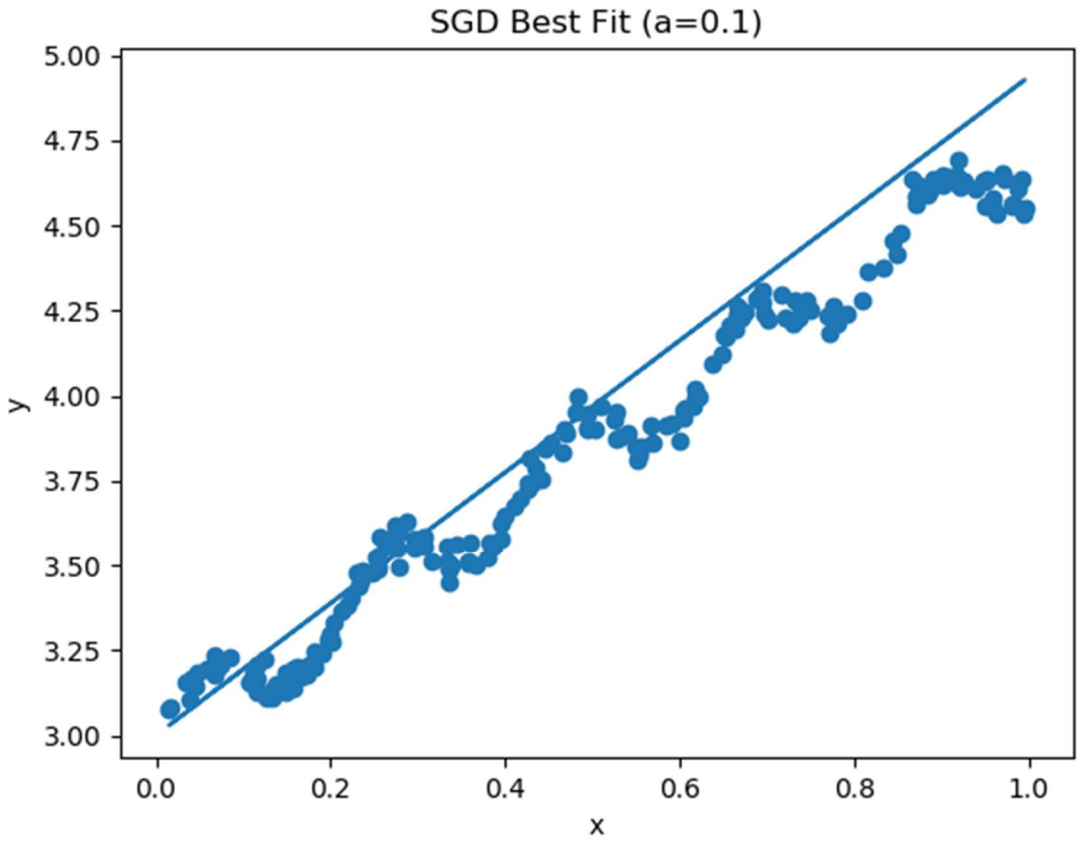

adient Descent Epoch vs Mean Training Loss -- Optimal Learning Rate (a=0.(
500
400
300
200
100
0
loss
0
20
40
60
80
100
epoch

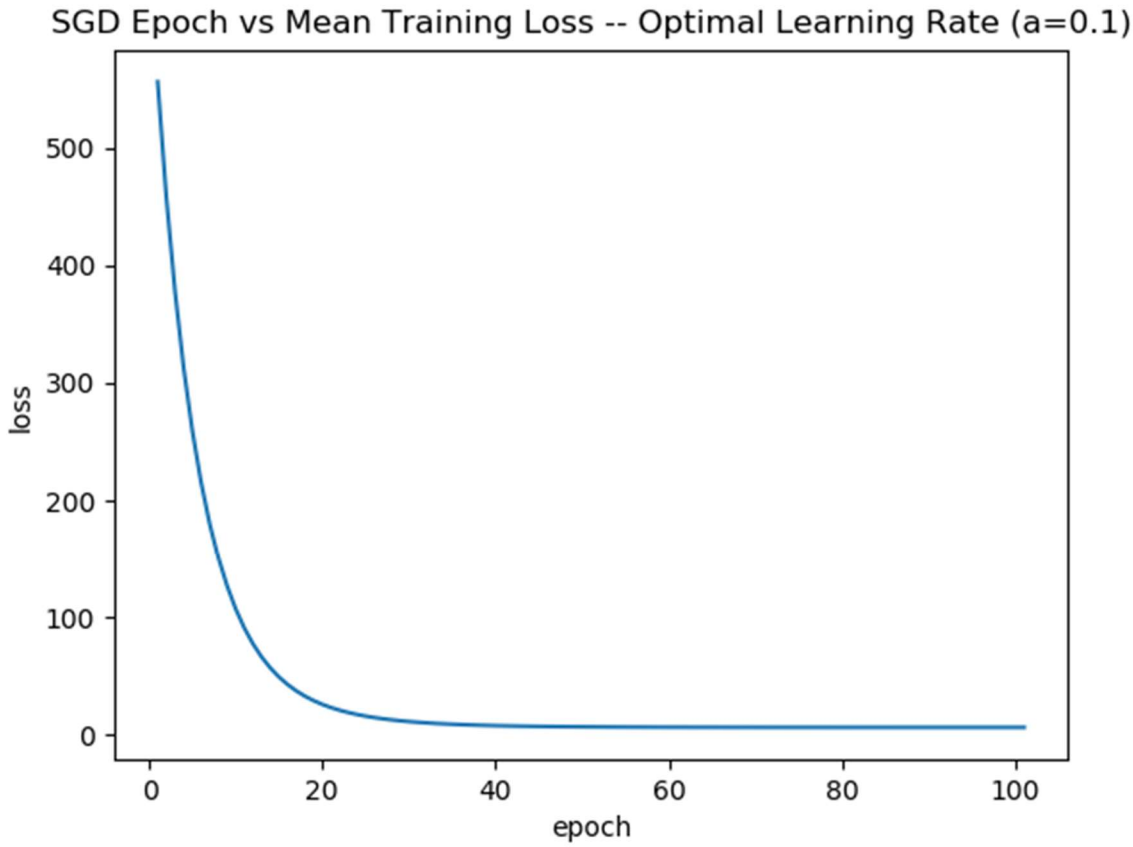
a=0.001

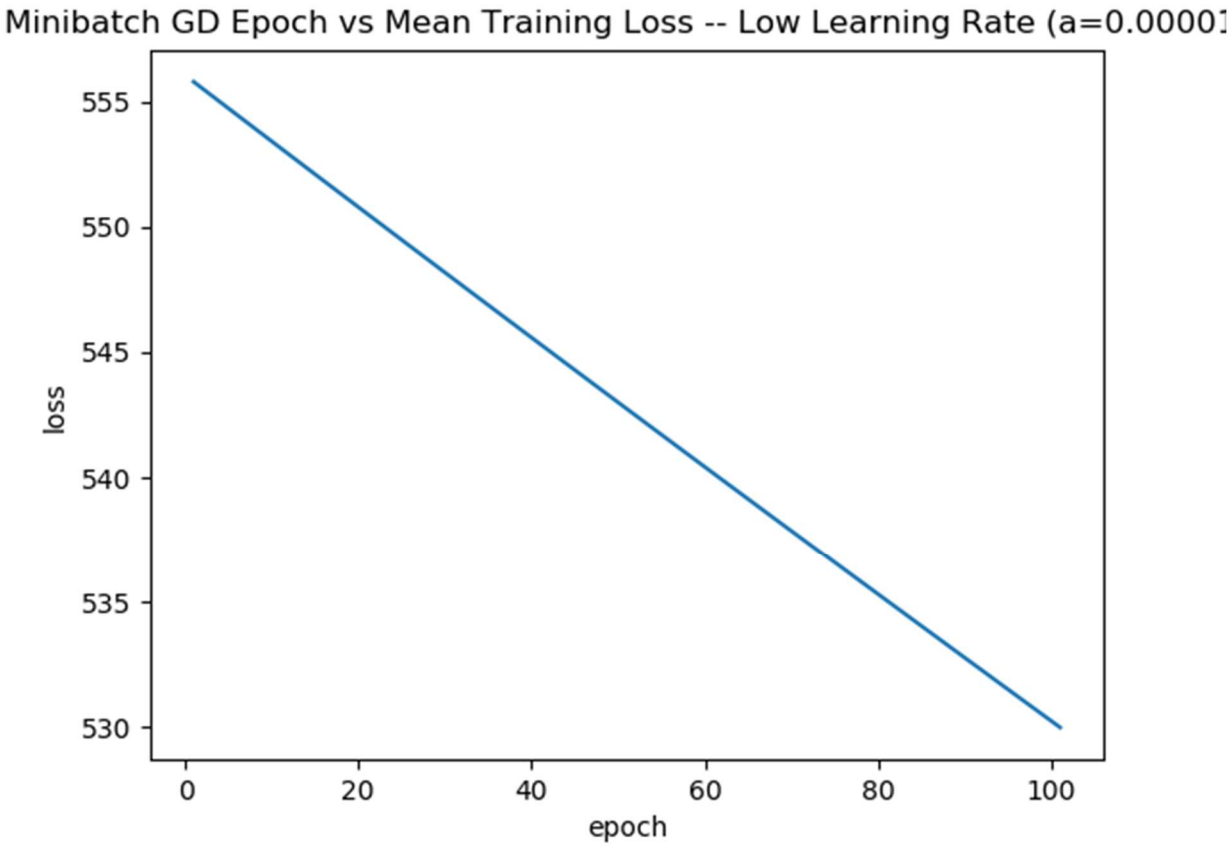

Gradient Descent Best Fit (a=0.001)
4.75
4.50
4.25
4.00
3.75
3.50
3.25
3.00
y
0.0
0.2
0.4
0.6
0.8
1.0
x

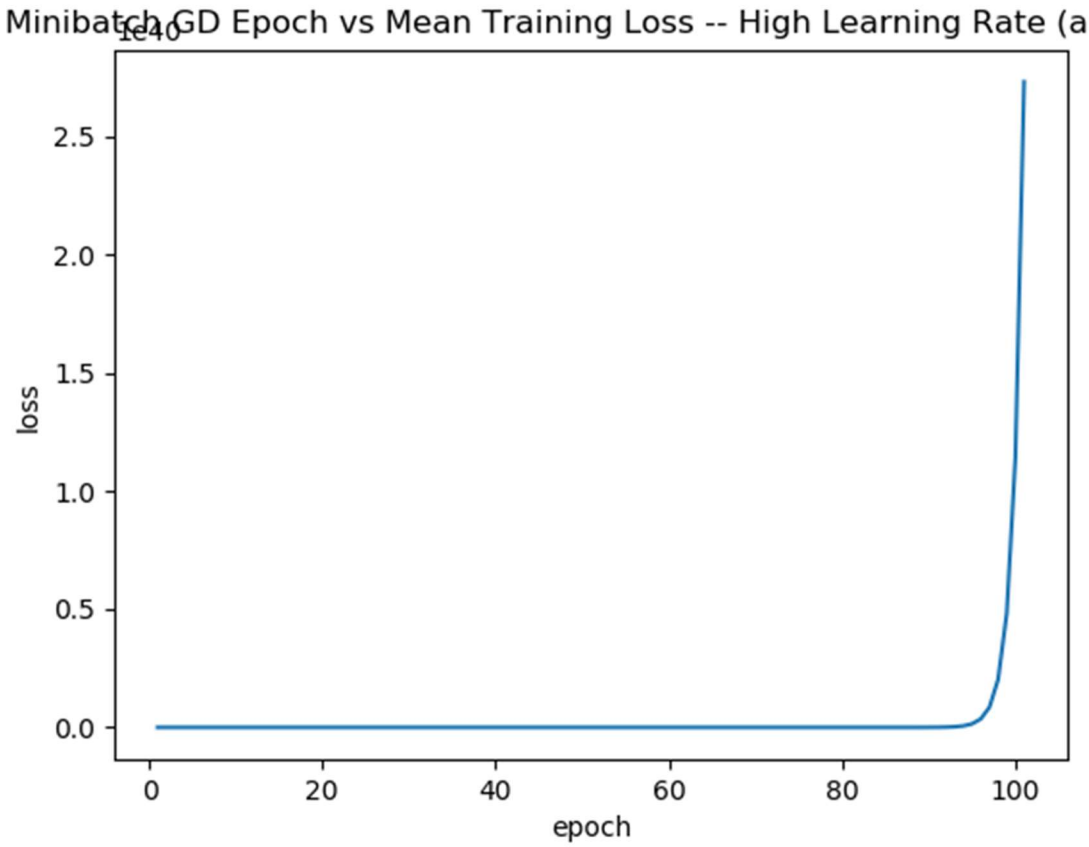

SGD Epoch vs Mean Training Loss -- Low Learning Rate (a=0.01)
loss
epoch
500
400
300
200
100
0
20
40
60
80
100

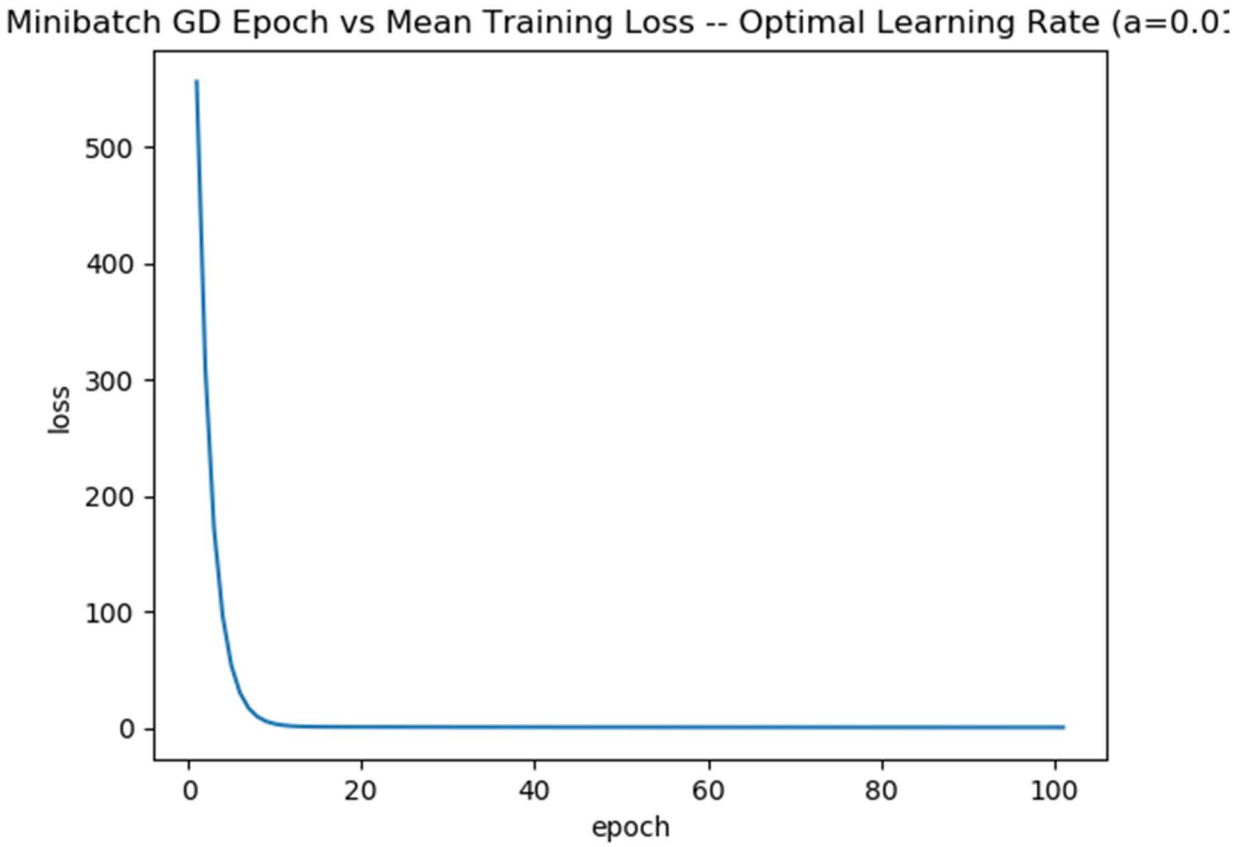

SGD Epoch vs Mean Training Loss -- High Learning Rate (a=2)
1e13
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
loss
0
20
40
60
80
100
epoch

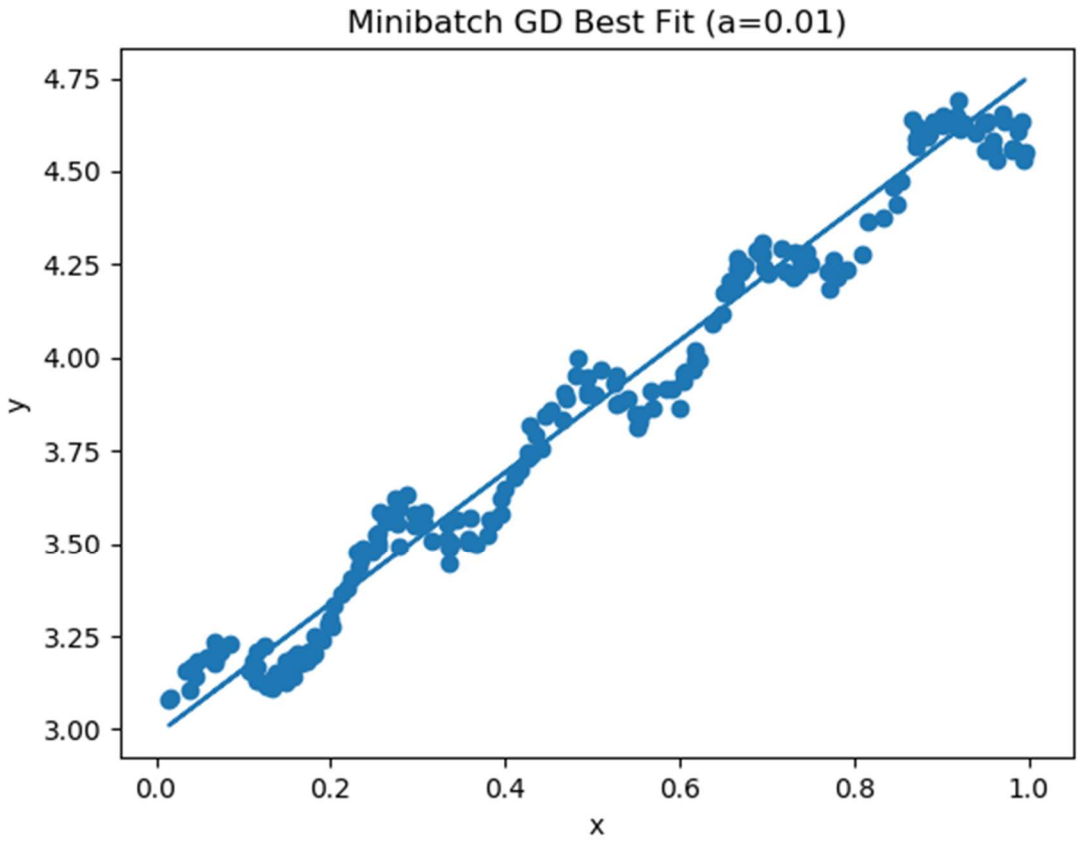

SGD Best Fit (a=0.1)
5.00
4.75
4.50
4.25
4.00
3.75
3.50
3.25
3.00
y
0.0
0.2
0.4
0.6
0.8
1.0
x

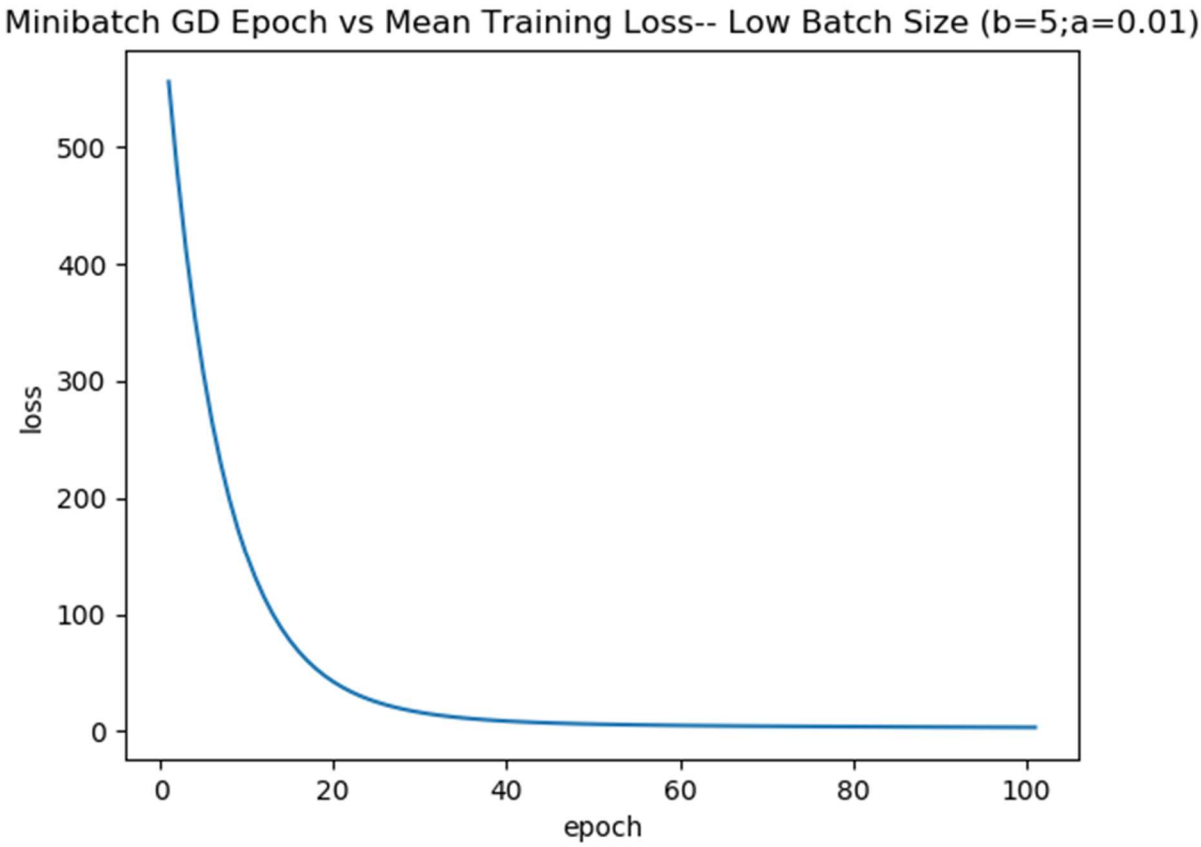

SGD Epoch vs Mean Training Loss -- Optimal Learning Rate (a=0.1)
loss
epoch
0
100
200
300
400
500
0
20
40
60
80
100

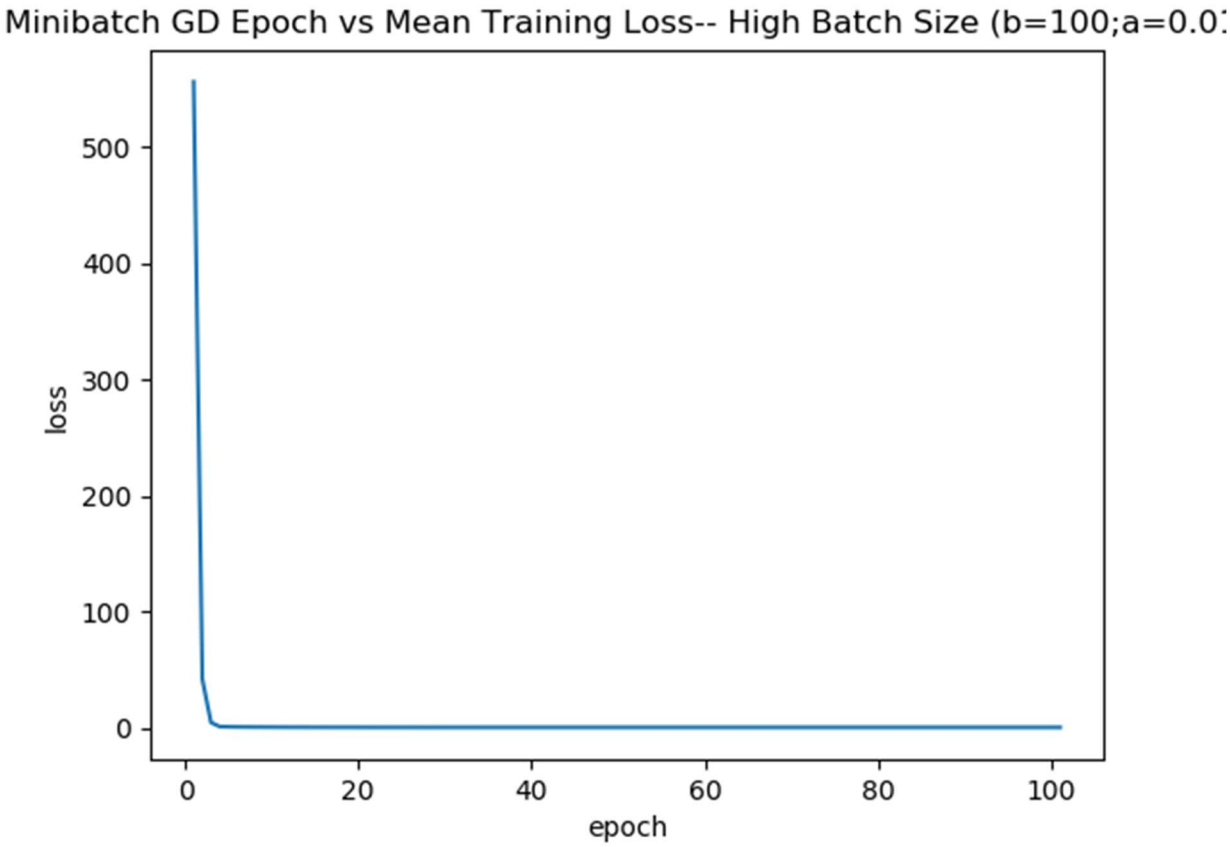

Minibatch GD Epoch vs Mean Training Loss -- Low Learning Rate (a=0.0000
555
550
545
540
535
530
loss
0
20
40
60
80
100
epoch

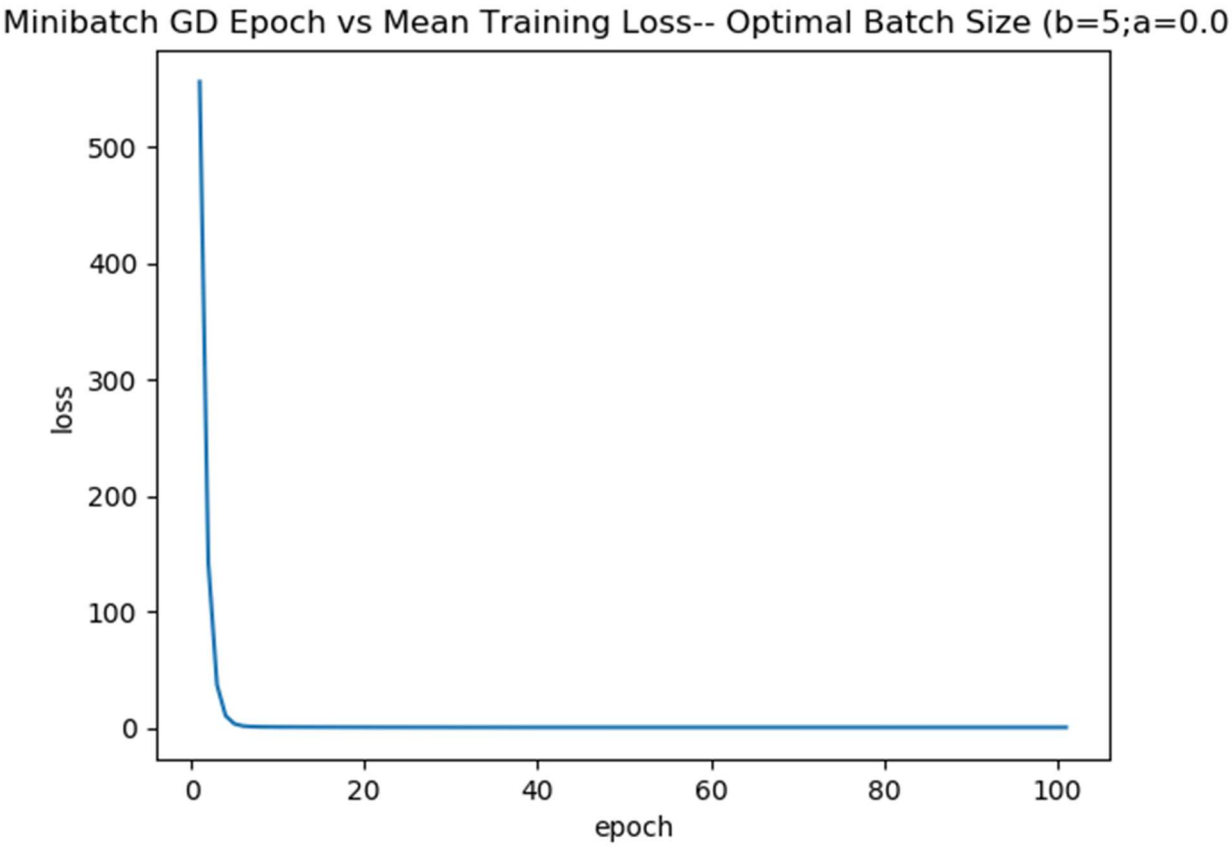

Minibatch GD Epoch vs Mean Training Loss -- High Learning Rate (a=0.1)
1e40
2.5
2.0
1.5
1.0
0.5
0.0
loss
0
20
40
60
80
100
epoch


Minibatch GD Epoch vs Mean Training Loss -- Optimal Learning Rate (a=0.0
500
400
300
200
100
0
loss
0
20
40
60
80
100
epoch

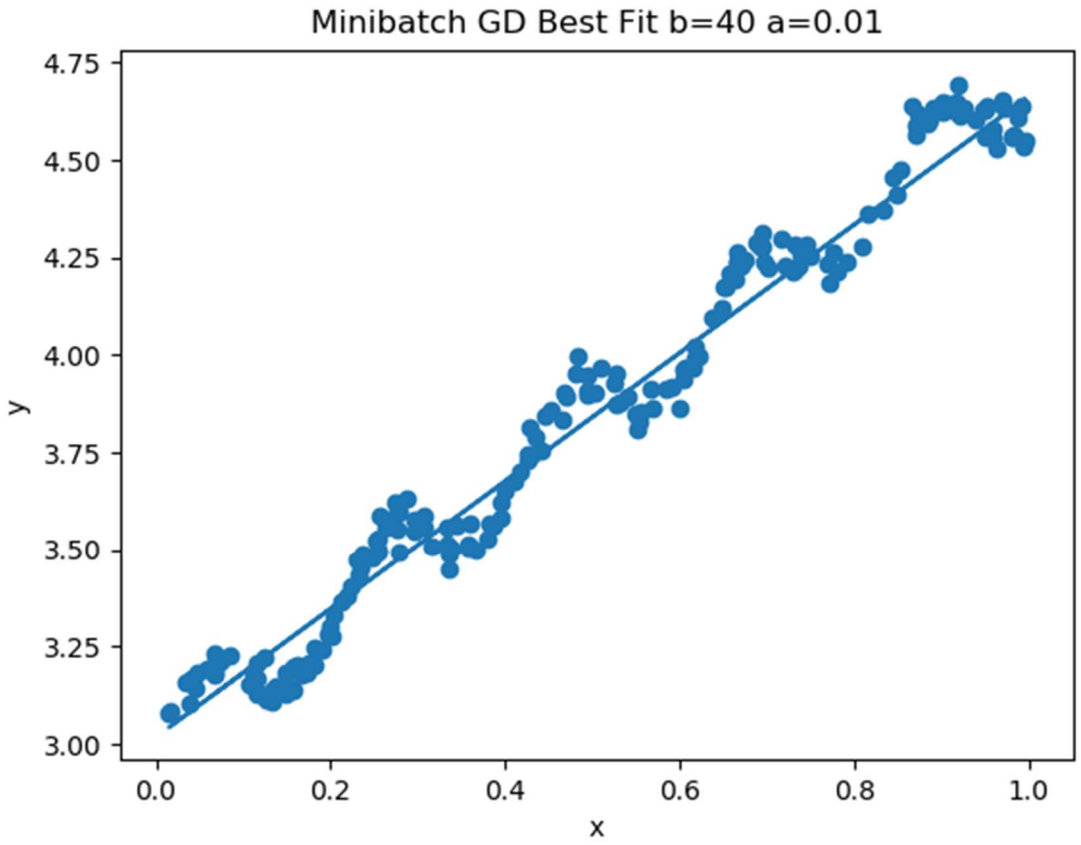

Minibatch GD Best Fit (a=0.01)
4.75
4.50
4.25
4.00
3.75
3.50
3.25
3.00
y
0.0
0.2
0.4
0.6
0.8
1.0
x


Minibatch GD Epoch vs Mean Training Loss-- Low Batch Size (b=5;a=0.01)
loss
epoch
0
100
200
300
400
500
0
20
40
60
80
100


Minibatch GD Epoch vs Mean Training Loss-- High Batch Size (b=100;a=0.0
500
400
300
200
100
0
loss
0
20
40
60
80
100
epoch



a=0.01; **b=40 there was an error in the title**


Minibatch GD Best Fit b=40 a=0.01
4.75
4.50
4.25
4.00
3.75
3.50
3.25
3.00
y
0.0
0.2
0.4
0.6
0.8
1.0
x

Effect of Learning Rate on Gradient Descent:

If the rate is set too low, the loss will converge too slowly and it will take a larger number of epochs to begin converging. If the rate is too high, overfitting will occur, and the error will balloon. If the rate is optimal, it will converge within 100 epochs.

Effect of Learning Rate on Stochastic Gradient Descent:

Similar effects are observed as with the standard gradient descent algorithm, but the optimal learning rate seems to be an order of magnitude larger for the stochastic version.

Effect of Learning Rate on Mini-Batch Gradient Descent:

As with the previous two, If the rate is set too low, the loss will converge too slowly and it will take a larger number of epochs to begin converging. If the rate is too high, overfitting will occur, and the error will balloon. If the rate is optimal, it will converge within 100 epochs. The learning rates that satisfy these three conditions tend to be similar to normal gradient descent when the batch size is set at 20.

Effect of Batch Size on Mini-Batch Gradient Descent:

Although there is clearly some difference in the graphs, it doesn't balloon the error in the same way overfitting does. In summary, it has some effect on convergence, but will not totally ruin your regression if it's chosen to be sufficiently large.

Exam Question

(3.1) The MSE on the training data for a perfect fit to the training data (e.g. by making use of the normal equations) would be zero (I have made the assumption all points are collinear by eyeballing it)