1.1 ~~$\frac{1}{N}((wx+b)-y)(x)$~~ $\nabla mse(w,b) = \begin{pmatrix} \frac{1}{N}\sum((wx+b)-y)(x) \\ \frac{1}{N}\sum((wx+b)-y) \end{pmatrix}$

2.1 a) $\hat{Y} = XW$

b) $mse = \frac{1}{2N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 = \frac{1}{2N}(XW-Y)^T(XW-Y)$

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

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

$$= \frac{1}{2N}\left(\frac{\partial}{\partial W}(W^TX^TXW - W^TX^TY - Y^TXW + Y^TY)\right)$$

$$= \frac{1}{2N}\left(2X^TXW - X^TY - \frac{\partial}{\partial W}(Y^TXW)^T\right)$$

$$= \frac{1}{2N}\left(2X^TXW - X^TY - \frac{\partial}{\partial W}(W^TX^TY)\right)$$

$$= \frac{1}{2N}(2X^TXW - X^TY - X^TY) = \frac{X^TXW - X^TY}{N}$$

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

$$= mse_{ols} + \lambda W^TW$$

$$\frac{\partial}{\partial W} mse = \frac{\partial}{\partial W} mse_{ols} + \lambda\frac{\partial}{\partial W} W^TW$$

$$= \frac{X^TXW - X^TY}{N} + 2\lambda W$$

3.1 Let $R = \begin{bmatrix} r_1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & r_2 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & r_i & \cdots & 0 \\ 0 & 0 & 0 & \cdots & \cdots & r_n \end{bmatrix}_{n \times n}$

$$E(W) = \frac{1}{2N} \sum_{i=1}^{N} r_i^2 (y_i - W^T x_i)^2$$

$$= \frac{1}{2N} (RY - RXW)^T (RY - RXW)$$

$$\frac{\partial E(W)}{\partial W} = \frac{1}{2N} \frac{\partial}{\partial W} (Y^T R^T RY - Y^T R^T RXW - W^T X^T R^T RY + W^T X^T R^T RX$$

$$= \frac{1}{2N} \left(\frac{\partial}{\partial W} (-YR^T RXW)^T - X^T R^T RY + 2X^T R^2 XW \right)$$

$$= \frac{1}{2N} \left(\frac{\partial}{\partial W} (-W^T X^T R^T RY^T) - X^T R^2 Y + 2X^T R^2 XW \right)$$

$$= \frac{1}{2N} \left(-X^T R^2 Y - X^T R^2 Y + 2X^T R^2 XW \right)$$

$$= \frac{1}{N} \left(X^T R^2 XW - X^T R^2 Y \right) = 0$$

$$\Rightarrow W = (X^T R^2 X)^{-1} (X^T R^2 Y)$$

4.2 The closed form of OLS doesn't have a solution if the columns of X are linearly dependent. Yes, the gradient descent converges to a sol. in that case, a closed form solution doesn't exist.

The modification I did to the data was to convert X to I(4*4)

4.1 The program was not executing properly because, it was calculating $(X^TX)^{-1}$ for an X, whose columns are linearly dependent. So, I modified the data in such a way that the columns are linearly independent and no.of columns ≤ no.of rows.

$|X^TX| \neq 0$ iff X has ~~not~~ full column rank (given no.of columns ≤ no.of rows)

If X has full column rank and $|X^TX| = 0$,

Assume $X^TXv = 0 \Rightarrow v^T X^TXv = 0 \Rightarrow (Xv)^TXv = 0$

$\Rightarrow ||Xv||^2 = 0 \Rightarrow Xv = 0$, a contradiction

If $|X^TX| \neq 0$ and X has non-full column rank,

Assume $Xv = 0 \Rightarrow X^TXv = 0 \Rightarrow |X^TX| = 0$, a contradiction

Hence, proved

~~If~~ Let no.of columns be n and no.of rows be m.

If $n \geq m$,

~~Rank (X)~~ col. rank $(X) \leq n$

col.rank $(X^T) \leq m$

Since col rank $(X^TX)_{n \times n} \leq$ min. (col. rank (X^T), col.rank (X))

$\leq$ min. (m, n)

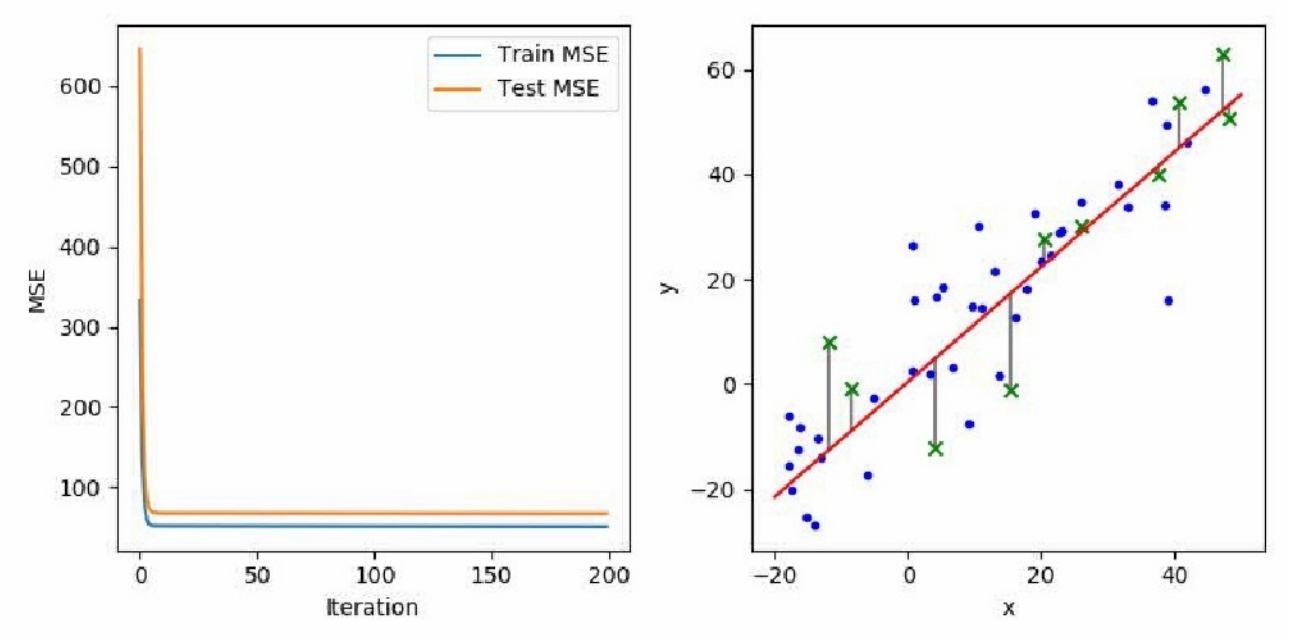
$\leq m < n \Rightarrow X^TX$ is non-invertible

4.2) (continued) Also, for OLS to not have a closed form sols no.of col.(X) ≥ no.of rows (X)

If there exists an x_0 such that, $X^TX x_0 = X^TY$, then there exict infinite solutions of the form $(x_0 + \lambda v)$, where λ is any scalar and v belongs to null space of X^TX (i.e. $X^TXv = 0$), as we are discussing the case when X^TX is not invertible. Hence, gradient descent converges to any of these solutions.

If a minima to which gradient descent converge exicts, it should satisfy the eq. $X^TXw_0 = X^TY$, where w_0 is a point of minima ($\because X^TXw_0 = X^TY$ is obtained by equating derivative to zero, which is the condition for minima). Hence we can say that multiple minima exict with same mse value (using above paragraph) and gradient descent, by the nature of the algorithm, converges to any one of them).

1.2 d) The blue points denote training data and the green points denote test data and the red line denotes the best fit line. The line best fits on the training data, as it is trained on the training data, but also fits the test data somewhat well. Some points have positive error and some have negative error, but the overall error is minimized as we used the square of the error as our measure.


Train MSE
Test MSE
MSE
Iteration
600
500
400
300
200
100
0
50
100
150
200
y
x
60
40
20
0
−20
−20
0
20
40

The vertical distance between the points and the line denote $\hat{y}-y$, where $\hat{y}$ is predicted by the line and y is the original value. As the line passes through the middle of the data, it also matches our intuition and prior knowledge that the mean reduces mean square error. But, we can also see that that the line is also influenced by outliers, since we treat all training data equally.