**Problem 1**

**1.1 CS337: Theory**

1.1. CS 337: Theory

$$\nabla mse(w,b) = \begin{pmatrix} \partial mse(w,b)/\partial w \\ \partial mse(w,b)/\partial b \end{pmatrix}$$

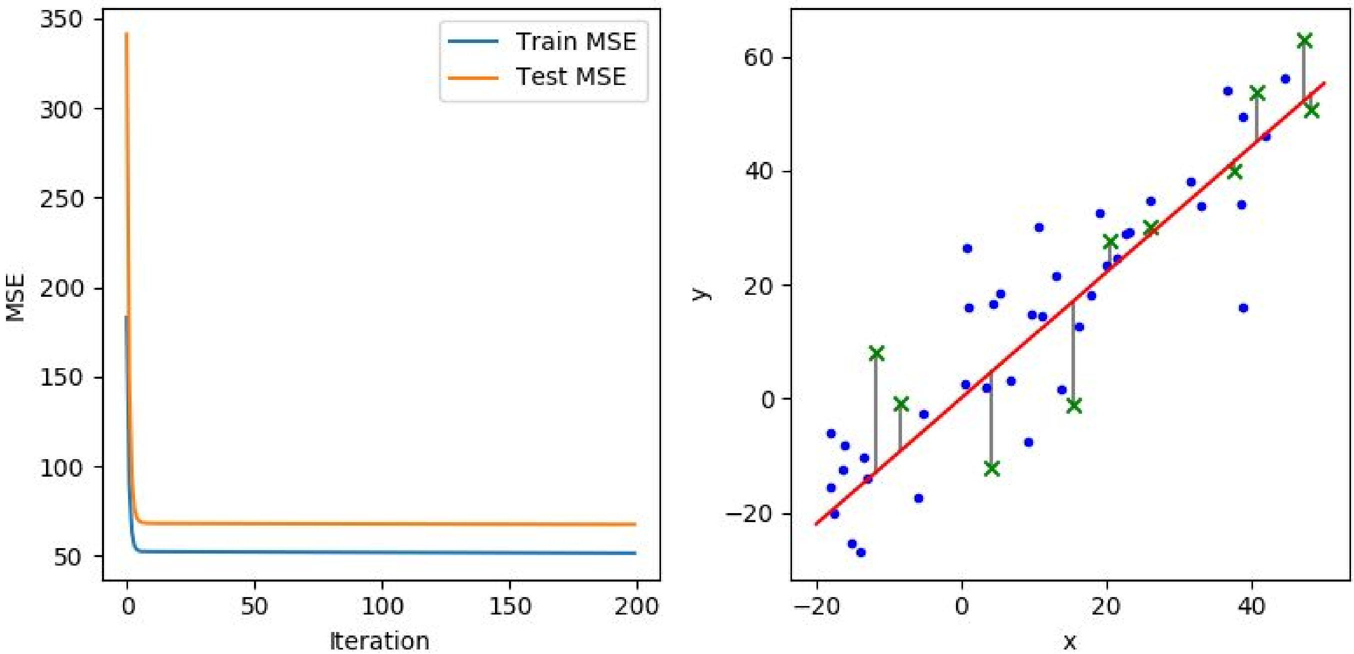
$$mse(w,b) = \frac{1}{2N}\sum_{i=1}^{N}\left((wx_i+b)-y_i\right)^2$$

$$\frac{\partial mse(w,b)}{\partial w} = \frac{1}{2N}\sum_{i=1}^{N} 2\left((wx_i+b)-y_i\right)x_i = \frac{1}{N}\sum_{i=1}^{N}\left[\left((wx_i+b)-y_i\right)x_i\right]$$

$$\frac{\partial mse(w,b)}{\partial b} = \frac{1}{2N}\sum_{i=1}^{N} 2\left((wx_i+b)-y_i\right) = \frac{1}{N}\sum_{i=1}^{N}\left[(wx_i+b)-y_i\right]$$

$$\therefore \nabla mse(w,b) = \begin{bmatrix} \frac{1}{N}\sum_{i=1}^{N}\left[x_i\left((wx_i+b)-y_i\right)\right] \\ \frac{1}{N}\sum_{i=1}^{N}\left[\left((wx_i+b)-y_i\right)\right] \end{bmatrix}$$

**1.2 CS335: Lab**
**(d)**




As we see in the first figure, both train and test mse flatten out with increase in iteration number. Thus parameters $(w, b)$ are close to $(w_{optimal}, b_{optimal})$ as the gradient is close to 0. The train mse is less than the test mse as the data is fitted on the train data

For the linear prediction model, we observe a lot of deviation from the curve resulting in high test_mse error
The red line represents the linear curve
Blue dots: train points
Green cross: test points with grey lines denoting deviation

We also see the curve tries to fit the data well with nearly equal spread of points above and below the curve

**Also note that we didn't preprocess the data!**

**Problem 2**
**2.1 CS337: Theory**

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

(b) $$mse(W) = \frac{1}{2N}\sum_{i=1}^{N}(X_iW - Y_i)^2$$

$$\frac{\partial mse(W)}{\partial W} = \frac{1}{\cancel{2}N}\sum_{i=1}^{N}\cancel{2}(X_iW - Y_i)X_i$$

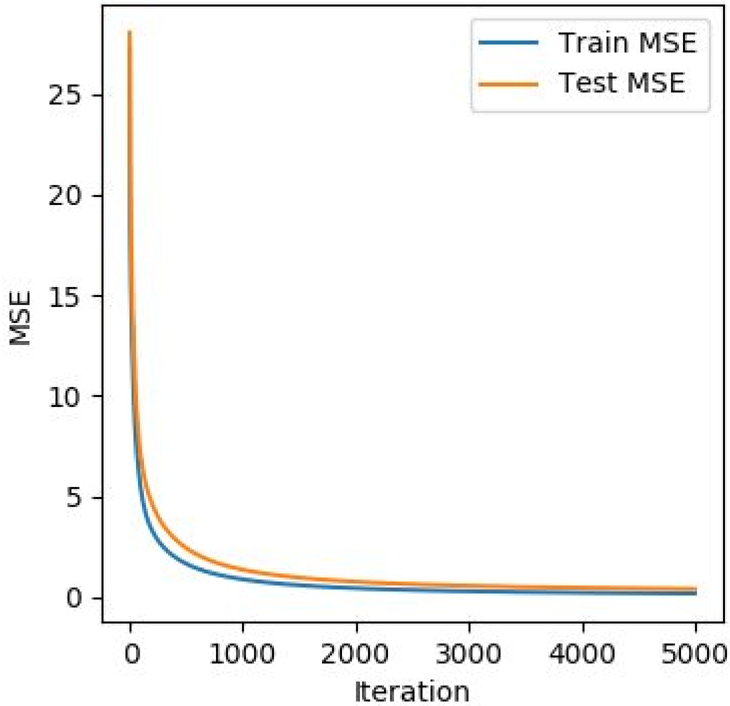
$$= \frac{1}{N}\sum_{i=1}^{N}(X_iW - Y_i)X_i \quad \left[\because \frac{\partial}{\partial W}(X_iW) = X_i\right]$$

$$= \frac{1}{N}[X^T(XW - Y)]$$

(c) $$mse(W) = \frac{1}{2N}\sum_{i=1}^{N}(X_iW - Y_i)^2 + \lambda\|W\|^2$$

$$\frac{\partial mse(W)}{\partial W} = \frac{1}{N}[X^T(XW-Y)] + 2\lambda W \quad \left[\because \frac{\partial}{\partial W}(\|W\|^2) = 2\lambda W\right]$$

**2.2 CS335: Lab**




**Problem 3**
**3.1: CS 337: Theory**

3 Weighted Linear Regression

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

$$= \frac{1}{2n} (Y - XW)^T R (Y - XW)$$

where : R is a diagonal matrix with $r_i^2$ on the diagonal and zeros everywhere

X is a matrix with $x_i$ on the rows

W is a column vector w

$$E(w) = \frac{1}{2n} [Y^T R Y - Y^T R X W - W^T X^T R Y + W^T X^T R X W]$$

$$\nabla_w E(w) = \frac{1}{2n} [- X^T R Y - X^T R Y + 2 X^T R X W] \quad (\because R^T = R)$$

equating to zeros, we get

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

[assuming $X^T R X$ is full column rank]

**Problem 4**

4. Failure Cases of Linear Regression

4.1. The matrix X is not full column rank hence $X^T X$ is also not full rank. $\Rightarrow (X^T X)^{-1}$ does not exist.

We observe that the columns X0 and X2 are dependent with X2 = 3X0. Hence X is not full column rank. Removing column X2 produces a full column rank X and hence we obtain the corresponding matrix W.

4.2 The closed form solution of OLS exists if $(X^T X)^{-1}$ exists
$\Rightarrow$ X is full column rank since $X^T X$ is full rank [inverse exists]
Hence the data matrix X has to be full column rank.

In case X is not full column rank, there exists dependency among columns of X. Hence there ~~exists~~ does not exist a unique solution W but there exists a optimal subspace of W. These all are global optimal W and the gradient descent converges to one such W after which gradient turns 0 This can also be seen by the fact that the objective function is convex hence the gradient descent converges to one of optimal solutions of W