1 a) Show that $MSE(w) = ||y - Xw||^2$

For Multivariate regression the equation for the dependent variable is given by

$$y = w_0 + w_1 x^{(1)} + w_2 x^{(2)} + w_3 x^{(3)} + \dots\dots + w_d x^{(d)}$$

where d is the number of attributes (colums) of the dataset.

Assume the dataset is not normalized, so we do have an intercept value.

Equation for y can be written as,

$$y = \sum_{i=1}^{d} w_i x^{(i)}$$

Mean square error can be written as, $MSE(w) = \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i w)^2$

Ignore $1/n$ because its just a scaling factor.

$$MSE(w) = \sum_{i=1}^{n} (y_i - x_i w)^2$$

$$MSE(w) = (y_1 - x_1 w)^2 + (y_2 - x_2 w)^2 + \dots (y_n - x_n w)^2$$

This can be written using an L2-norm as

$$MSE(w) = ||y - Xw||^2$$

X matrix is $\begin{bmatrix} 1 \cdots\cdots & x^{(1)} & \cdots\cdots \\ 1 \cdots\cdots & x^{(2)} & \cdots\cdots \\ & \vdots & \\ 1 \cdots\cdots & x^{(n)} & \cdots\cdots \end{bmatrix}$ order $\rightarrow n \times (d+1)$

n - number of observations

d - number of attributes of every single observation

w matrix is $\begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{bmatrix}$ order $\rightarrow (d+1) \times 1$

y matrix is of the order $(n \times 1)$

The coordinates of w represent the different coefficient values in a d-dimensional vector space. Each coefficient maps to one of the d dimentions.

Prove that $\hat{w} = (X^T X)^{-1} X^T y$

(1) b) $y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}_{n \times 1}$ $\quad w = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{bmatrix}_{(d+1) \times 1}$ $\quad X = \begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \cdots & x_d^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \cdots & x_d^{(2)} \\ \vdots & \vdots & & & \\ 1 & x_1^{(n)} & x_2^{(n)} & \cdots & x_d^{(n)} \end{bmatrix}_{n \times (d+1)}$

$$Xw = \begin{bmatrix} w_0 + w_1 x_1^{(1)} \cdots + w_d x_d^{(1)} \\ w_0 + w_1 x_2^{(2)} \cdots + w_d x_d^{(2)} \\ \vdots \\ w_0 + w_1 x_2^{(n)} \cdots + w_d x_d^{(n)} \end{bmatrix}$$

Error, $e(w) = y - \langle x . w \rangle$

$$MSE(w) = \frac{1}{n} \sum e_i^2(w)$$

$$= \frac{1}{n} e^T . e$$

This is because, $[e_1 \ e_2 \cdots e_n] \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = e_1^2 + e_2^2 + \cdots e_n^2$

Substitute the value of e, $e = y - xw$

$$\therefore MSE(w) = \frac{1}{n} . (y - xw)^T (y - xw)$$

$$MSE(w) = \frac{1}{n} \left[(y^T - x^T w^T)(y - xw) \right]$$

$$= \frac{1}{n}\left[y^T y - y^T x w - x^T w^T y + x^T w^T x w \right]$$

$$\frac{d(MSE(w))}{d(w)} = \frac{1}{n}\left[0 - y^T x - x^T y + 2x^T x w \right]$$

But, $(x^T y)^T = y^T x$

$$\frac{d(MSE)}{d(w)} = \frac{1}{n} \cdot \left[0 - y^T x - y^T x + 2x^T x w \right]$$

$$= \frac{1}{n} \cdot \left[-2y^T x + 2x^T x w \right]$$

$$\frac{d(MSE)}{d(w)} = \frac{2}{n}\left[x^T x w - y^T x \right]$$

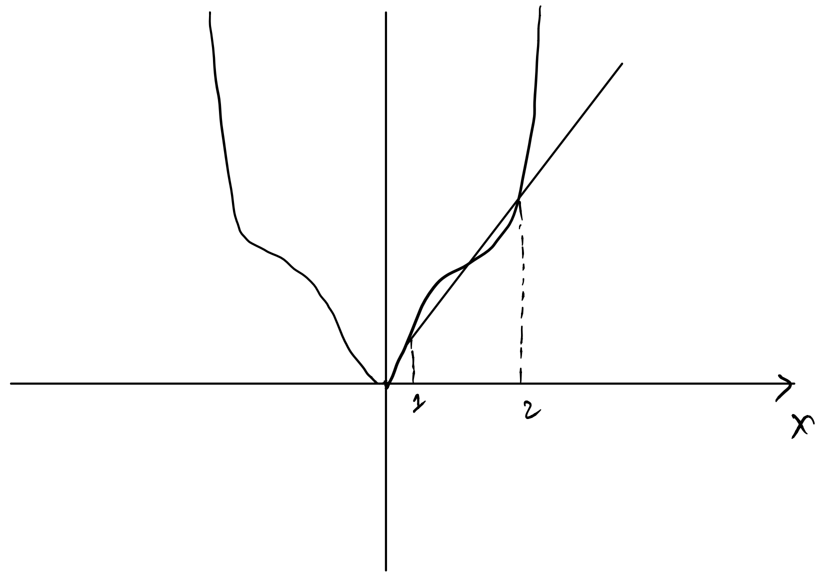
for optimal value of w, equate the above equation to zero

$$x^T x \hat{w} - y^T x = 0$$

$\hat{w} \rightarrow$ optimal value of w $\quad \boxed{\hat{w} = y^T x (x^T x)^{-1}}$

(2) No, convexity is not a necessary condition for gradient descent to successfully train a model.

For the example given, $f(x) = x^2 + 3\sin^2 x$,
the curve looks something like this.
Here we can see that the curve is not convex because the line drawn for points 1 and 2 does not lie above $f(x)$ curve.



While smoothness is a pre-requisite for gradient descent, that is not the case for convexity.

Gradient descent algorithm gives us an optimal (minimum) value for the above curve despite not being convex as long as the curve is smooth and there is a global minima.