Deep Learning HW1

1. Prove Normal Equations:
   Given a training set $\mathcal{S} = \{\mathbf{X}, \mathbf{y}\}$, a linear hypothesis class $\{h_{\boldsymbol{\theta}}(\mathbf{x}) = \sum_{j=1}^{N} \theta_j x_j\}$ and the mean squared error loss function:

$$\mathcal{L} = \frac{1}{2M} \sum_{i=1}^{M} \left(h_{\boldsymbol{\theta}}(\mathbf{x_i}) - y_i\right)^2$$

   prove that $\boldsymbol{\theta}$ that minimizes $\mathcal{L}$ satisfies:

$$\mathbf{X}^T \mathbf{X} \boldsymbol{\theta} = \mathbf{X}^T \mathbf{y}$$

   where: $\mathbf{x_i}, \boldsymbol{\theta} \in \mathbb{R}^N$, $\mathbf{y} \in \mathbb{R}^M$, $\mathbf{X} = \begin{bmatrix} — & \mathbf{x_1}^T & — \\ — & \mathbf{x_2}^T & — \\ & \vdots & \\ — & \mathbf{x_M}^T & — \end{bmatrix}$, $M \geq N$

$Solution$:

Let us derive $\mathcal{L}$ w.r to $\theta$.

$$\mathcal{L} = \frac{1}{2M} \sum_{i=1}^{M} (h_\theta(x_i) - y_i)^2 = \frac{1}{2M} \sum_{i=1}^{M} \left( \sum_{j}^{N} \theta_j x_{ij} - y_i \right)^2 = \frac{1}{2M} \sum_{i=1}^{M} \left\| \theta X_i^T - y_i \right\|^2 = \frac{1}{2} \|X\theta - y\|^2$$

$$\frac{\partial \mathcal{L}}{\partial \theta} = X^T (X\theta - y)$$

$$\frac{\partial \mathcal{L}}{\partial \theta} = 0 \; iff \; X^T (X\theta - y) = 0$$

$$\rightarrow X^T X \theta = X^T y$$

As required.

2. Unique solution:

Show that a unique solution for linear regression exists iff the features are not linearly dependent. Namely, show that a unique solution:

$$\underset{\boldsymbol{\theta}}{\operatorname{argmin}} \mathcal{L} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

exists iff $\mathbf{X}$ has full column rank.

Assume X is full rank, we know that $rk(X) = rk(X^T)$ and since $X^TX$ is square (matrix multiplication definition) That would suggest that $X^TX$ is also invertible since it's a composition of full rank transformation.

Taking our optimality condition from Q1 that states that $X^TX\theta = X^Ty$ and applying our new found knowledge we could rearrange the terms to get $(X^TX)^{-1}X^TX\theta = (X^TX)^{-1}X^Ty \; iff \; (X^TX)^{-1}X^Ty = \theta$. Meaning that the optimal point of $\mathcal{L}$ with respect to $\theta$ is when $\theta = (X^TX)^{-1}X^Ty$.

Assume that a unique solution $\theta = (X^TX)^{-1}X^Ty$ exists. We know from basic linear algebra that rank $X$ = rank $X^TX$ and since $X^TX$ is invertible, it is full rank, and so is $X$.

Short proof that $rk(X) = rk(X^TX)$. Proving that $Null(X^TX) \subset Null(X)$ suffices since the other direction is trivial, and the statement will follow.

Let $u$ be a vector, such that $X^TXu = 0$ and let $v = Xu$. $u^T \overbrace{X^TXu}^{0} = v^Tv = 0$. That could only be if $v$ is the 0 vector. And so $Null(X^TX) \subset Null(X)$.