**Data 100, Spring 2023**

Discussion #7 Solutions

# Dive into Gradient Descent

1. We want to minimize the loss function $L(\theta) = (\theta_1 - 1)^2 + |\theta_2 - 3|$. While you may notice that this function is not differentiable everywhere, we can still use gradient descent wherever the function *is* differentiable!

   Recall that for a function $f(x) = k|x|$, $\frac{df}{dx} = k$ for all $x > 0$ and $\frac{df}{dx} = -k$ for all $x < 0$.

   (a) What are the optimal values $\hat{\theta}_1$ and $\hat{\theta}_2$ to minimize $L(\theta)$? At that point $\hat{\theta}$, what is the gradient $\nabla L = \begin{bmatrix} \frac{\partial L}{\partial \theta_1} & \frac{\partial L}{\partial \theta_2} \end{bmatrix}^T \Big|_{\theta_1 = \hat{\theta}_1, \theta_2 = \hat{\theta}_2}$?

   **Solution:** By inspection, neither the square loss nor the absolute loss can be smaller than 0. Hence, the minimizing values are $\theta_1 = 1$ and $\theta_2 = 3$.

   At this point, $\frac{\partial L}{\partial \theta_1}\Big|_{\theta_1 = 1} = 0$, but $\frac{\partial L}{\partial \theta_2}\Big|_{\theta_2 = 3}$ is undefined!

   Staff Notes: Students may be confused by this if they are coming from a math background, but let them know that this function is technically non-differentiable at $\theta_2 = 3$, but that this has no bearing on the rest of the question, it just motivates why we may want to use gradient descent.

   (b) Suppose we initialize our gradient descent algorithm randomly at $\theta_1 = 2$ and $\theta_2 = 5$. Calculate the gradient $\nabla L = \begin{bmatrix} \frac{\partial L}{\partial \theta_1} & \frac{\partial L}{\partial \theta_2} \end{bmatrix}^T \Big|_{\theta_1 = 2, \theta_2 = 5}$ at the specified $\theta_1$ and $\theta_2$ values.

   **Solution:** For $\theta_2 > 3$:

   $$\begin{bmatrix} \frac{\partial L}{\partial \theta_1} & \frac{\partial L}{\partial \theta_2} \end{bmatrix}^T = \begin{bmatrix} 2(\theta_1 - 1) & 1 \end{bmatrix}^T$$

   Thus, the gradient is $\begin{bmatrix} 2 & 1 \end{bmatrix}^T$.

   (c) Apply the first gradient update with a learning rate $\alpha = 0.5$. In other words, calculate $\theta_1^{(1)}$ and $\theta_2^{(1)}$ using the initializations $\theta_1^{(0)} = 2$ and $\theta_2^{(0)} = 5$.

   **Solution:** Applying the gradient step:

   $$\theta^{(1)} = \theta^{(0)} - \alpha \nabla L = \begin{bmatrix} 2 - 0.5(2) \\ 5 - 0.5(1) \end{bmatrix} = \begin{bmatrix} 1 \\ 4.5 \end{bmatrix}$$

(d) How many gradient steps does it take for $\theta_1$ and $\theta_2$ to converge to their optimal values obtained in part (a) assuming we keep a constant learning rate of $\alpha = 0.5$? In other words, what is the value of t when $\theta^{(t)} = \begin{bmatrix}\hat{\theta}_1 & \hat{\theta}_2\end{bmatrix}^T$.

*Hint:* After part (c), what is the derivative $\frac{\partial L}{\partial \theta_1}$ evaluated at $\theta_1^{(1)}$?

**Solution:** Note that the derivative with respect to $\theta_1$ is 0 at $\theta_1^{(1)} = 1$ since it is the optimal solution! Then, we essentially only update $\theta_2$ where the partial derivative is always 1 (until we reach the optimal solution - then our derivative is undefined)! Every time, the partial derivative of $\theta_2$ is 1 - so the update is simply:

$$\theta_2^{(i+1)} = \theta_2^{(i)} - 0.5$$

Hence, to update this from 5 to 3, we must take 4 gradient steps (i.e. from 5 to 4.5, 4.5 to 4, 4 to 3.5, 3.5 to 3).

Writing this all out:

$$\theta^{(2)} = \theta^{(1)} - \alpha \nabla L = \begin{bmatrix} 1 - 0.5(0) \\ 4.5 - 0.5(1) \end{bmatrix} = \begin{bmatrix} 1 \\ 4 \end{bmatrix}$$

$$\theta^{(3)} = \theta^{(2)} - \alpha \nabla L = \begin{bmatrix} 1 - 0.5(0) \\ 4 - 0.5(1) \end{bmatrix} = \begin{bmatrix} 1 \\ 3.5 \end{bmatrix}$$

$$\theta^{(4)} = \theta^{(3)} - \alpha \nabla L = \begin{bmatrix} 1 - 0.5(0) \\ 3.5 - 0.5(1) \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

Notice that every time, we reduce $\theta_2$ by 0.5 as expected, so the number of gradient steps is 4.

# One-hot Encoding

2. In order to include a qualitative variable in a model, we convert it into a collection of Boolean vectors. These Boolean vectors contain only the values 0 and 1. For example, suppose we have a qualitative variable with 3 possible values, call them $A$, $B$, and $C$, respectively. For concreteness, we use a specific example with 10 observations:

$$[A, A, A, A, B, B, B, C, C, C]$$

We can represent this qualitative variable with 3 Boolean vectors that take on values $1$ or $0$ depending on the value of this qualitative variable. Specifically, the values of these 3 Boolean vectors for this dataset are $x_A$, $x_B$, and $x_C$, arranged from left to right in the following design matrix, where we use the following indicator variable:

$$x_{i,k} = \begin{cases} 1 & \text{if } i\text{-th observation has value } k \\ 0 & \text{otherwise.} \end{cases}$$

Furthermore, let $\vec{y}$ represent any vector of outcome variables, and $y_i$ is the value of said outcome for the $i$-th subject. This representation is also called one-hot encoding. It should be noted here that $\vec{x_A}$, $\vec{x_B}$, $\vec{x_C}$, and $\vec{y}$ are all vectors.

$$\mathbb{X} = \begin{bmatrix} | & | & | \\ \vec{x_A} & \vec{x_B} & \vec{x_C} \\ | & | & | \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

We will show that the fitted coefficients for $\vec{x_A}$, $\vec{x_B}$, and $\vec{x_C}$ are $\bar{y}_A$, $\bar{y}_B$, and $\bar{y}_C$, the average of the $y_i$ values for each of the groups, respectively.

(a) Show that the columns of $\mathbb{X}$ are orthogonal, (i.e., the dot product between any pair of column vectors is 0).

**Solution:** The argument is the same for any pair of $\mathbb{X}$'s columns so we show the orthogonality for one pair, $\vec{x_A} \cdot \vec{x_B}$.

$$\begin{aligned} \vec{x_A} \cdot \vec{x_B} &= \sum_{i=1}^{10} x_{A,i} x_{B,i} \\ &= \sum_{i=1}^{4}(1 \times 0) + \sum_{i=5}^{7}(0 \times 1) + \sum_{i=8}^{10}(0 \times 0) \\ &= 0 \end{aligned}$$

(b) Show that

$$\mathbb{X}^T \mathbb{X} = \begin{bmatrix} n_A & 0 & 0 \\ 0 & n_B & 0 \\ 0 & 0 & n_C \end{bmatrix}$$

Here, $n_A$, $n_B$, $n_C$ are the number of observations in each of the three groups defined by the levels of the qualitative variable.

**Solution:** Here, we note that

$$\mathbb{X}^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

We also note that

$$\mathbb{X}^T \mathbb{X} = \begin{bmatrix} \vec{x_A}^T \vec{x_A} & \vec{x_A}^T \vec{x_B} & \vec{x_A}^T \vec{x_C} \\ \vec{x_B}^T \vec{x_A} & \vec{x_B}^T x_B & \vec{x_B}^T \vec{x_C} \\ \vec{x_C}^T \vec{x_A} & \vec{x_C}^T \vec{x_B} & \vec{x_C}^T \vec{x_C} \end{bmatrix}$$

Since we earlier established the orthogonality of the vectors in $\mathbb{X}$, we find $\mathbb{X}^T \mathbb{X}$ to be the diagonal matrix:

$$\mathbb{X}^T \mathbb{X} = \begin{bmatrix} n_A & 0 & 0 \\ 0 & n_B & 0 \\ 0 & 0 & n_C \end{bmatrix}$$

(c) Show that

$$\mathbb{X}^T \mathbb{Y} = \begin{bmatrix} \sum_{i \in A} y_i \\ \sum_{i \in B} y_i \\ \sum_{i \in C} y_i \end{bmatrix}$$

where $i$ is an element in group $A$, $B$, or $C$.

**Solution:** Note in the previous solution we found $\mathbb{X}^T$. The solution follows from recognizing that for a row in $\mathbb{X}^T$, e.g., the first row, we have

$$\sum_{i=1}^{10} x_{A,i} \times y_i = \sum_{i=1}^{4} y_i = \sum_{i \in \text{ group A}} y_i$$

(d) Use the results from the previous questions to solve the normal equations for $\hat{\theta}$, i.e.,

$$\begin{aligned} \hat{\theta} &= [\mathbb{X}^T\mathbb{X}]^{-1}\mathbb{X}^T\mathbb{Y} \\ &= \begin{bmatrix} \bar{y}_A \\ \bar{y}_B \\ \bar{y}_C \end{bmatrix} \end{aligned}$$

**Solution:** By inspection, we can find

$$[\mathbb{X}^T\mathbb{X}]^{-1} = \begin{bmatrix} \frac{1}{n_A} & 0 & 0 \\ 0 & \frac{1}{n_B} & 0 \\ 0 & 0 & \frac{1}{n_C} \end{bmatrix}$$

When we pre-multiply $\mathbb{X}^T\mathbb{Y}$ by this matrix, we get

$$\begin{bmatrix} \frac{1}{n_A} & 0 & 0 \\ 0 & \frac{1}{n_B} & 0 \\ 0 & 0 & \frac{1}{n_C} \end{bmatrix} \begin{bmatrix} \sum_{i \in A} y_i \\ \sum_{i \in B} y_i \\ \sum_{i \in C} y_i \end{bmatrix} = \begin{bmatrix} \bar{y}_A \\ \bar{y}_B \\ \bar{y}_C \end{bmatrix}$$

(e) *(Extra)* Show that if you augment your $\mathbb{X}$ matrix with an additional $\vec{1}$ bias vector as shown below, $\mathbb{X}^T\mathbb{X}$ is not full rank. Conclude that the new $\mathbb{X}^T\mathbb{X}$ is not invertible, and we cannot use the least squares estimate in this situation.

$$\mathbb{X} = \begin{bmatrix} | & | & | & | \\ \vec{1} & \vec{x_A} & \vec{x_B} & \vec{x_C} \\ | & | & | & | \end{bmatrix}$$

**Solution:**
We can show that $\mathbb{X}^T\mathbb{X}$ is equal to the following.

$$\mathbb{X}^T\mathbb{X} = \begin{bmatrix} n & n_A & n_B & n_C \\ n_A & n_A & 0 & 0 \\ n_B & 0 & n_B & 0 \\ n_C & 0 & 0 & n_C \end{bmatrix}$$

It can be observed that since $n_A + n_B + n_C = n$, the sum of the final 3 columns subtracted from the first column yields the zero vector $\vec{0}$. By the definition of linear dependence, we can conclude that this matrix is not full rank, and hence, we cannot invert it. As a result, we cannot compute our least squares estimate since it requires $\mathbb{X}^T\mathbb{X}$.