DS 100: Principles and Techniques of Data Science **Date: July 24, 2019**

Discussion #9

Name:

Loss Functions

1. Recall the loss functions discussed during lecture. θ represents our estimate of our parameter, and y represents a data point. Discuss the advantages and drawbacks of each of the following loss functions:

   (a) Squared loss: $L(\theta, y) = (y - \theta)^2$

   (b) Absolute Loss: $L(\theta, y) = |y - \theta|$

   (c) Huber Loss: $L_\alpha(\theta, y) = \begin{cases} (y-\theta)^2 & |y-\theta| < \alpha \\ \alpha(|y-\theta| - \alpha/2) & \text{otherwise} \end{cases}$

Loss Minimization

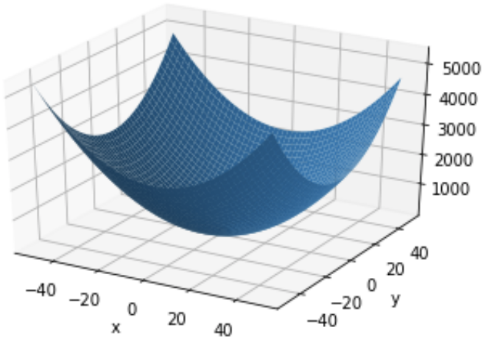
2. Consider the following loss function:

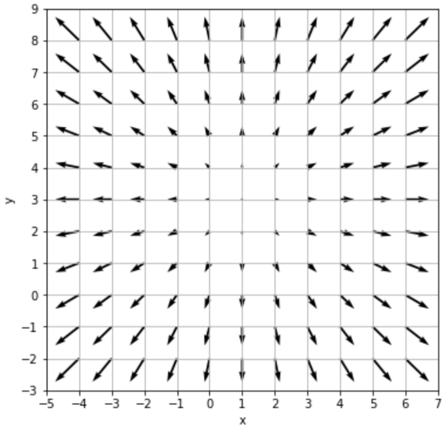
$$L(\theta, x) = \begin{cases} 4(\theta - x) & \theta \geq x \\ x - \theta & \theta < x \end{cases}$$

   Given a sample of $x_1, ..., x_n$, find the optimal θ that minimizes the the average loss.

Gradients

3. On the left is a 3D plot of $f(x, y) = (x-1)^2 + (y-3)^2$. On the right is a plot of its **gradient field**. Note that the arrows show the relative magnitudes of the gradient vector.





(a) From the visualization, what do you think is the minimal value of this function and where does it occur?

(b) Calculate the gradient $\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \end{bmatrix}^T$.

(c) When $\nabla f = \mathbf{0}$, what are the values of x and y?

4. In this question, we will explore some basic properties of the gradient.

   Note: In this class, we use the following conventions:

   - x represents a scalar
   - X represents a random variable
   - **x** represents a vector
   - **X** represents a matrix or a random vector (context will tell)

   (a) Determine the derivative of $f(x) = a_0 + a_1 x$ and gradient of $g(x_1, x_2) = a_0 + a_1 x_1 + a_2 x_2$.

   (b) Suppose $\mathbf{x} = \begin{bmatrix} x_1 & x_2 & ... & x_n \end{bmatrix}^T$, and $h(\mathbf{x}) = \mathbf{a}^T\mathbf{x}$, where **a, x** $\in \mathbb{R}^n$. Determine ∇h.

   (c) Determine the gradient of $f(\mathbf{x}) = \mathbf{x}^T\mathbf{x}$. *(Hint: f is a scalar-valued function. How can you write $\boldsymbol{x}^T\boldsymbol{x}$ as a sum of scalars?)*

Gradient Descent Algorithm

5. Given the following loss function and $\mathbf{x} = (x_i)_{i=1}^n$, $\mathbf{y} = (y_i)_{i=1}^n$, θ^t, explicitly write out the update equation for θ^{t+1} in terms of x_i, y_i, θ^t, and α, where α is the step size.

$$L(\theta, \mathbf{x}, \mathbf{y}) = \frac{1}{n} \sum_{i=1}^{n} \left(\theta^2 x_i^2 - log(y_i) \right)$$

6. (a) In your own words, describe how to use the update equation in the gradient descent algorithm.

   (b) Say that x and y are your model parameters and f as defined in question 1 is your loss function. Describe in your own words what happens "visually" as the gradient descent algorithm runs.

Convexity

Convexity allows optimization problems to be solved more efficiently and for global optimums to be realized. Mainly, it gives us a nice way to minimize loss (i.e. gradient descent). There are three ways to informally define convexity.

a. Walking in a straight line between points on the function keeps you above the function. This works for any function.

b. The tangent line at any point lies below the function (globally). The function must be differentiable.

c. The second derivative is non-negative everywhere (aka ”concave up” everywhere). The function must be twice differentiable.

7. Find a counterexample for the claim that the composition of two convex functions is also convex. $h = g(f(x))$