Instructions: Make sure your questions and answers are on different pages. Do not include your name or any other identifying information. I will know that information from Canvas.

Question 1: What is one advantage of using Newton's method for optimization instead of gradient descent?

Question 2: Why will gradient descent always converge to the global minimum when the loss function is convex?

Question 3: Why are the second partial derivatives of the loss function stored in a matrix, rather than a vector like the first partial derivatives?

Answer Question 1: One advantage Newton’s method has over gradient descent is that it utilizes the second derivative of the loss function to converge to a local minimum more quickly.

Answer Question 2: Gradient descent will always converge to the global minimum on a convex function because a convex function only has one local minimum, which must be the global minimum. Since gradient always converges to a local minimum, that local minimum will be the global minimum in a convex function.

Answer Question 3: The second partial derivatives form a matrix because at the level of second partial derivatives, a small change in a single weight value has the potential to change both its own and all of the other weights’ first partial derivatives with respect to the loss. Therefore, with d weights, every weight needs d second partial derivatives, one for each weight’s first partial derivative. This results in a d x d matrix of second partial derivatives. For first partial derivatives, a small change in a single weight value only results in a potential change in the loss function’s value. So, you only need a vector of d first partial derivatives, one for each weight.