Question 1: Does Gradient Descent always produce the global minimum of the loss function? Why or why not?

Question 2: What are the tradeoffs and differences between Gradient Descent and Newton's Method?

Question 3: What are some possible pitfalls of Gradient Descent?

Answer 1: Gradient Descent does not always determine the global minimum of the loss function. Gradient Descent will stop when the gradient first approaches zero, so the algorithm will not continue past finding the local optimum. Unless the function contains a single local optimum, Gradient Descent does not guarantee the global optimum.

Answer 2: Gradient Descent relies only on first order partial derivatives, while Newton's Method uses second order partial derivatives. In addition, Gradient Descent allows the user to define its own parameter, the learning rate, while Newton's Method does not use a learning rate, or a user defined parameter. While Newton's Method might converge faster to the optimum than Gradient Descent since it uses second order partial derivatives, loss functions that have undefined or unusable second order partial derivatives cannot utilize Newton's Method. In addition, the use of a learning rate gives the user more control over tuning and shaping the model.

Answer 3: If the weights are initialized so that the gradient begins close to a local minimum and far from the global minimum, Gradient Descent will stop the nearest local minimum, instead of finding the global minimum. As a result, Gradient Descent will determine an ineffective set of parameters based on the effectiveness of the nearest and first local minimum that Gradient Descent finds.