Question 1: There are multiple iterative methods for optimization. Give one example and briefly explain what it is and how it can be useful.

Question 2: What is one restriction that prevents gradient descent from working?

Question 3: For strongly convex functions, briefly describe what happens to gradient descent with a constant step size?

Answer 1: One example is newton’s method. This is where you initialize the parameters at some value and decrease that value of the empirical risk iteratively by running the utilizing the following formula: $wt+1 = wt - (\nabla^2 f(wt))^{-1} \nabla f(wt)$. This can be a very useful method because it can converge more quickly than gradient descent due to it being a second-order optimization method. It essentially uses the second derivative of f rather than the first.

Answer 2: If the learning rate is too large or the value of the gradient is zero, then gradient descent will not work by decreasing the value of the objective at each iteration.

Answer 3: For strongly convex functions, which is basically a function that has a very smooth and bowl-shaped structure that guarantees that there is a unique optimal solution, gradient descent with a constant step size converges exponentially quickly to the optimum. This is also known or called convergence at a linear rate.