Assignment 2 Question 3

Sheen Thusoo

Part b)

Although batch-sequential, batch-stochastic, and ordinary gradient descent are all optimization methods for minimizing an objective function, there are slight differences that can be contrasted. The ordinary gradient descent algorithm involves going through each unit in the population P and calculating the gradient at this unit for every iteration. This differs from batch-sequential and batch-stochastic gradient descent, as for these algorithms the gradient is calculated at a batch of units in the population for every iteration. This is possible because when we sum over the population, the gradient can be split into N (where N is the size of the population) smaller and independent gradient calculations which can be computed in any order. This is valuable when N is extremely large as the ordinary gradient descent algorithm would take a very long time to find a solution. With batch-sequential and batch-stochastic, the units are split into H non-overlapping batches of size M (such that $H * M = N$) and the gradient of each batch can be calculated in parallel. Thus, batch-sequential and batch-stochastic gradient descent have a shorter computation time. Since we are estimating the gradient using a subset (or batch) of the population, we usually use a fixed step size, λ instead of optimizing the step size using line search. Batch-sequential and batch-stochastic gradient descent differ in the way in which the subsets or batches are chosen. In the batch-sequential, we sequentially move through the H batches we created and update the estimated theta value **after each batch** rather than after all batches. If it takes us more than H iterations to converge to a solution, we iterate through the batches sequentially. In batch-stochastic gradient descent, we randomly select samples (or batches) from the population and compute the gradient at each iteration. It is similar to batch-sequential in that it also updates the theta value after each batch, but it differs in that the batches are randomly selected. When comparing each of these algorithms on a graph visually, the path of ordinary gradient descent is a smooth line towards the direction of the minimum of the function. The path of batch-sequential gradient descent is in a slightly more noisy line (which moves in different directions) until it reaches the minimum of the objective function. The path of stochastic gradient descent is a highly noisy line (going in many different directions) that moves in the general direction of the solution. The paths of batch-sequential and batch-stochastic are more noisy since we compute the gradient at a sample (batch) of the population rather than each and every point. Batch-stochastic is even more noisy since we randomly select the sample (batch) at which to calculate the gradient.