c. Define and explain the difference between on-policy and off-policy algorithms. What

are the advantages and disadvantages of each?

[6 marks]

On policy

- Learning occurs on the target policy itself

- 'learning by doing'

- Continually estimate for and change some behaviour policy \pi

Off policy

- Learning comes from an alternate policy which is then used to learn the optimal policy

- Kind of works the way that our brains do - similar to 'learning by observation'

An on-policy algorithm aims to learn using the target policy itself with the algorithm choosing actions using a chosen policy (such as epsilon greedy) and can be said to be ‘learning by doing.’ Once an action has been selected, it is then used to update the estimated costs associated with the policy. The algorithm then converges with probability 1 when exploration is reduced over time and all state-action pairs are visited an infinite number of times.

Compared to an on-policy algorithm, off-policy algorithms do not use the action taken to update. Instead, the action used to update will always choose the best possible action and as a result, unlike on-policy algorithms does not need the action taken at St+1. However, this results in the algorithm not considering potential negative costs associated with exploration. Since the actions chosen are the best possible actions, the learning comes from an alternate policy which is then used to learn to optimal policy and is similar to how our brains “learn by observation.”