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 (often soft (non-deterministic) 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.

Unlike on-policy algorithms, 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. 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.”

Comparing the two approaches, off-policy algorithms offer more advantages if a good starting policy is not available as it encourages more exploration and does not require the action taken at St+1. However, this comes at the cost of increasing the time taken for the algorithm to run due to the extra exploration. Additionally, off-policy algorithms do not consider potential negative costs associated with exploration. On the other hand, on-policy would be more advantageous if a good policy is already available, but this approach may not explore other policies well and risk getting trapped in local minima.

A good example that shows the advantages and disadvantages of each approach is the cliff-walking example discussed in S&B. Both approaches use ε-greedy action selection with Sarsa as the on-policy and Q-learning as the off-policy algorithm. Initially, the on-policy approach performs significantly better as it chooses the safe path, the agent is unlikely to fall into the cliff; however, the trade-off for this is that the safe path is sub-optimal requiring a few extra steps. The off-policy approach performs worse at the start as it tries to use the optimal path, the agent occasionally falls off the cliff, resulting in a steep penalty. This difference in approach results in the different rewards per episode seen in the graph below. If ε is gradually reduced, both policies will converge to the optimal policy.

Diagram

Description automatically generated