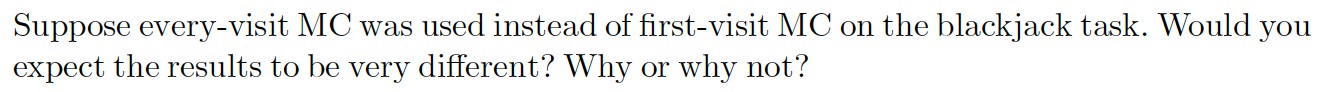
**Reinforcement Learning CAP 6629, Homework 3**

**Professor Xiangnan Zhong**

**Matthew Acs, Fall 2023**

Question 1 (10 points):



No, I would not expect the results to be very different. Every-visit MC updates using average returns for every time a state is visited in an episode, while first-time MC only updates using average returns for the first time a state is visited in an episode. Therefore, every-visit and first visit will not produce significantly different results because the only scenario in which states can be revisited is if an ace is switched from being an 11 to being a 1. This is not a common scenario, thus the difference between every-visit MC and first-visit MC is not significant. The difference between every-visit MC and first-visit MC is more significant in tasks that contain many revisited states, such as in a gridworld that allows for moving back to a previous state.

Question 2 (15 points):



A diagram of a structure

Description automatically generated

, S0

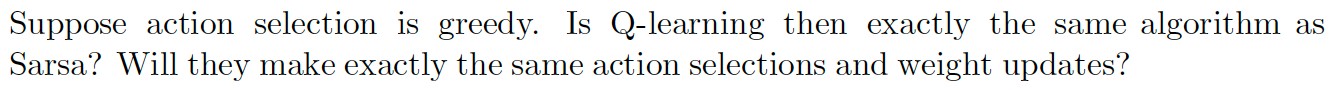
**Backup diagram for Monte Carlo estimation of q**

Question 3 (10 points):



Off-policy control methods use a behavior policy b to generate experience that is used to learn the value of the target policy π. Q-learning is considered an off-policy control method because it uses a different policy to generate actions for agent exploration from the policy that is being updated. Q-learning updates its q-values using the next states q value with the greedy action; however, the behavior policy does not follow this. On the other hand, an on-policy method such as SARSA, updates its q-values using the next states q value with the current policies next action. On policy control methods update the target policy using actions derived from the current policy.

Question 4 (15 points):



No, Q-learning and SARSA are different algorithms even if action selection is greedy. This is because Q-learning utilizes the maximum Q values of the next state in its update while SARSA chooses the next action and utilizes that in its update. Thus, Q learning updates its Q values before taking the next action while SARSA does not. This means that a Q learning algorithm may choose a different action based on the fact that it updated earlier than SARSA. For instance, when an action makes the agent move to the same state (i.e., crashing into a wall in a gridworld example), Q learning will choose an action based on an updated Q-value while SARSA will not. Thus, they make different actions selections and weight updates.