Problem Set 5: Solutions

Hadi Bagdadi

Question 1

(a) Fill in the Q-value updates

$\alpha = 1$ for simplicity:

$$Q(s,a) \leftarrow r + \gamma \max_{a'} Q\big(s', a'\big),$$

where

$$r = \text{immediate reward}, \quad \gamma = 0.9, \quad s' = \text{resulting state after taking action } a \text{ from } s.$$

Step 1: $Q(B1, \text{Up}) = 0 \;+\; 0.9 \cdot 0 = 0.$

Step 2: $Q(A1, \text{Right}) = 0 \;+\; 0.9 \cdot 0 = 0.$

Step 3: $Q(A2, \text{Right}) = 0 \;+\; 0.9 \cdot 3 = 2.7.$

Step 4: $Q(A3, \text{Down}) = 2.7 \;+\; 0.9 \cdot 5 = 7.2.$

Step 5: $Q(B3, \text{Left}) = 7.2 \;+\; 0.9 \cdot 10 = 16.2.$

Step 6: $Q(B1, \text{Up}) = 0 \;+\; 0.9 \cdot 0 = 0.$

Step 7: $Q(A1, \text{Right}) = 0 \;+\; 0.9 \cdot 0 = 0.$

Step 8: $Q(A2, \text{Right}) = 0 \;+\; 0.9 \cdot 3 = 2.7.$

Step 9: $Q(A3, \text{Down}) = 2.7 \;+\; 0.9 \cdot 5 = 7.2.$

Step 10: $Q(B3, \text{Left}) = 7.2 \;+\; 0.9 \cdot 10 = 16.2.$

(b) Effect of changing γ to 0.99 and 0.01

- $\gamma = 0.99$**:** A larger discount factor means the agent values future rewards more strongly. As a result, it will favor actions that lead to higher long-term returns, even if they are further away. In this grid, the agent is more likely to take a longer path if that leads to a higher total reward by eventually reaching `B2` with more accumulated reward.
- $\gamma = 0.01$**:** A very small discount factor means the agent is almost myopic (short-sighted). It heavily prioritizes immediate reward and discounts future rewards very quickly. The resulting policy will tend to pick actions that give quicker or more certain immediate rewards rather than a long path with bigger delayed reward. Thus, the path to `B2` might be more direct (or in some cases, the agent may even fail to reach `B2`) if it perceives little immediate reward on the way.

(c) Constructing a policy from the given Q-values

Suppose the numeric values on each arrow in Figure 1 are actually the learned Q-values (instead of being the rewards). To construct a policy, we choose from each state the action with the highest Q-value. For instance:

- **Starting in A2:** If the largest Q-value from `A2` is “Right” (leading to `A3`), we choose that.
- **In A3:** We compare Q(`A3,Down`), Q(`A3,Left`), etc., and pick the action with the highest value. Suppose “Down” to `B3` is the highest, so we choose that.
- **In B3:** We again pick whichever action has the highest Q-value (e.g., if “Left” going to `B2` is highest, we pick that).

Continuing in this manner yields the policy (i.e., the best action to take from each cell) that maximizes the learned Q-values.