CMPUT 365: Introduction to Reinforcement Learning, Winter 2023
Worksheet #3: Value Functions and Bellman Equations

Manuscript version: *#0ade60-dirty* - 2023-04-05 12:25:28-06:00

Question 1. (*Exercise 3.12 S&B*, with more details) Let $\pi : \mathcal{S} \to \mathcal{M}_1(\mathcal{A})$ be a memoryless policy in a finite MDP $M = (\mathcal{S}, \mathcal{A}, \mathcal{R}, p)$ and consider the total discounted value criteria. Recall that the value $v_\pi(s)$ for state s under policy π is the expected total discounted reward an agent will receive when starting from state s and then executing policy π. Recall also that the value $q_\pi(s, a)$ for the state-action pair (s, a) under policy π is the expected total discounted reward an agent will receive when starting from state s, taking first action a and then executing policy π for the remaining time steps. Express $v_\pi(s)$ as a function of the action values $q_\pi(s, a)$ and the action probabilities $\pi(a|s)$. Prove that the relation you propose holds.

Question 2. In this question, you will take a word specification of an MDP, and write the formal terms and determine the optimal policy.

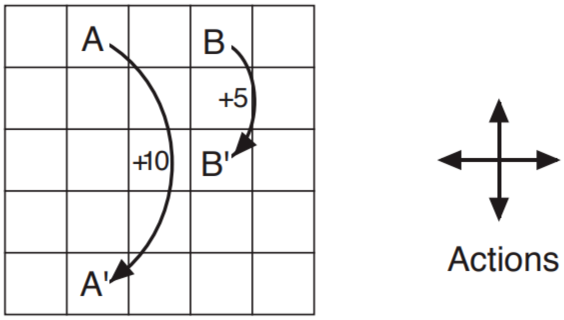
Suppose there are two actions. In the problem, the agent always starts in the same state, s_0. From this state, if it takes action 1 it transitions to a new state s_1 and receives reward 10; if it takes action 2 it transitions to a new state s_2 and receives reward 5. From s_1, if it takes action 1 it receives a reward of 5 and terminates; if it takes action 2 it receives a reward of 10 and terminates. From s_2 if it takes action 1 it receives a reward of 10 and terminates; if it takes action 2 it receives a reward of 5 and terminates. Assume the agent cares equally about rewards regardless of the timestep that they are incurred.

1. Draw the MDP for this problem. Is this an episodic or continuing problem? What is the value of the discount factor γ?

2. Assume the policy is $\pi(a = 1|s_i) = 0.3$ for all $s_i \in \{s_0, s_1, s_2\}$. What is $\pi(a = 2|s_i)$? And what is the value function for this policy? In other words, find $v_\pi(s)$ for all three states.

3. What is the optimal policy in this environment?

Question 3. Consider the gridworld and value function in the figure below. Using your knowledge of the transition dynamics and the values (numbers in each grid cell), write down the policy corresponding to taking the greedy action with respect to the values in each state. Create a grid with the same dimension as the figure and draw an arrow in each square denoting the greedy action. The meaning of the arrows on the left-hand side figure is that if one takes any action in state A, one lands on state A' while receiving a reward of 10. Similarly, any action taken in state B leads to state B' while a reward of 5 is incurred. Otherwise, the actions work as usual, including that on the boundary an action that would lead out of the cell will just result in the next state to be the same as the state where the action was taken.



3.3	8.8	4.4	5.3	1.5
1.5	3.0	2.3	1.9	0.5
0.1	0.7	0.7	0.4	-0.4
-1.0	-0.4	-0.4	-0.6	-1.2
-1.9	-1.3	-1.2	-1.4	-2.0

Question 4. (*Exercise 3.22 S&B*, modified by adding a question and made consistent with the terminology used in class) Consider the continuing MDP shown on the bottom. The only decision to be made is that in the top state, where two actions are available, left and right. The numbers show the rewards that are received deterministically after each action.

(a) List and describe all the possible *deterministic memoryless* policies in this MDP.

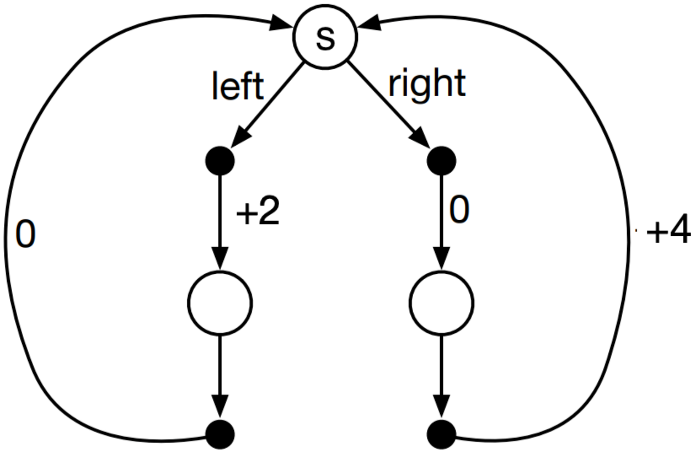
(b0) Assume an agent starts at state s and then chooses *left* for the next five steps when there is a decision to be made. Could it be that the agent was following a memoryless policy when taking these decisions? Explain your answer.

(b1) Assume an agent starts at state s and then happens to choose *left* for the next two steps when there is a decision to be made and then happens to choose *right* for the next three steps when there is a decision to be made. Could it be that the agent was following a memoryless policy when taking these decisions? Explain your answer.

(b2) Is the following policy a memoryless policy for this MDP? Choose *left* for five steps when there is a decision to be made, then *right* for five steps when there is a decision to be made, then *left* for five steps when there is a decision to be made, and so on? Explain your answer.

(c) Give a memoryless policy that is optimal if $\gamma = 0$! Do the same for $\gamma = 0.9$! Do the same for $\gamma = 0.5$!

(d) For each of the values of γ above, write down the optimal value of state s to two decimal places. Show your work.



Question 5. Fix $\alpha \geq 0$ and $\beta \in \mathbb{R}$. Let $M = (\mathcal{S}, \mathcal{A}, \mathcal{R}, p)$ be an MDP. Let $M' = (\mathcal{S}, \mathcal{A}, \mathcal{R}', p')$ be the MDP we get from M if all rewards $r \in \mathcal{R}$ of M are replaced by $\alpha r + \beta$. Thus,

$$\mathcal{R}' = \{\alpha r + \beta \, : \, r \in \mathcal{R}\}$$

and for $(s, a, r', s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{R} \times \mathcal{S}$,

$$p'(\alpha r' + \beta, s'|s, a) = p(r', s'|s, a)\,.$$

1. (**) Let π^* be an optimal policy for M. Is π^* an optimal policy for M'? Why or why not?
2. (*) Let $\alpha > 0$, and let π^* be now an optimal policy for M'. Is π^* an optimal policy for M? Why or why not?
3. What is the answer to the last question when $\alpha = 0$?

Question 6. Let π be a memoryless policy in a finite MDP $M = (\mathcal{S}, \mathcal{A}, \mathcal{R}, p)$ under the discounted total expected reward criterion, Express the action-value function q_π in terms of v_π. The formula will also include p and π. Prove the correctness of your formula.

Question 7. A genie comes to you and promises that you will have access to either the optimal state value function v_*, or the optimal state-action value function q_* of an MDP where you need to know how to act optimally (but not both). Which one will you choose and why?
