Value Functions & Bellman Equations

Rupam Mahmood

January 31, 2020

Important equations

$$v_\pi(s) = \sum_a \pi(a \mid s) q_\pi(s, a)$$

$$v_\pi(s) = \sum_a \pi(a \mid s) \sum_{s',r} p(s', r \mid s, a)[r + \gamma v_\pi(s')]$$

$$q_\pi(s, a) = \sum_{s',r} p(s', r \mid s, a) \left[r + \gamma \sum_{a'} \pi(a' \mid s') q_\pi(s', a') \right]$$

$$v_*(s) = \max_a q_*(s, a)$$

$$v_* = \max_a \sum_{s',r} p(s', r \mid s, a)[r + \gamma v_*(s')]$$

?

Bellman equation with expected reward $r(s, a)$

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)[r + \gamma v_\pi(s')]$$

$$\sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)r = \sum_a \pi(a|s) \sum_r r \sum_{s'} P(S_{t+1} = s', R_{t+1} = r | S_t = s, A_t = a)$$

$$= \sum_a \pi(a|s) \sum_r rP(R_{t+1} = r | S_t = s, A_t = a)$$

due to the law of total probabilities

$$= \sum_a \pi(a|s) E[R_{t+1} | S_t = s, A_t = a]$$

$$= \sum_a \pi(a|s) r(s, a)$$

Therefore, $v_\pi(s) = \sum_a \pi(a|s) \left[r(s, a) + \gamma \sum_{s'} p(s'|s, a) v_\pi(s') \right]$

Worksheet 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 you have a problem with two actions. 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 long term reward as about immediate reward.

(a) Draw the MDP for this problem. Is it an episodic or continuing problem? What is γ?

(b) 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.

(c) What is the optimal policy in this environment?

$$v_\pi(s) = \sum_a \pi(a \mid s) \sum_{s',r} p(s', r \mid s, a)[r + \gamma v_\pi(s')]$$

Worksheet question

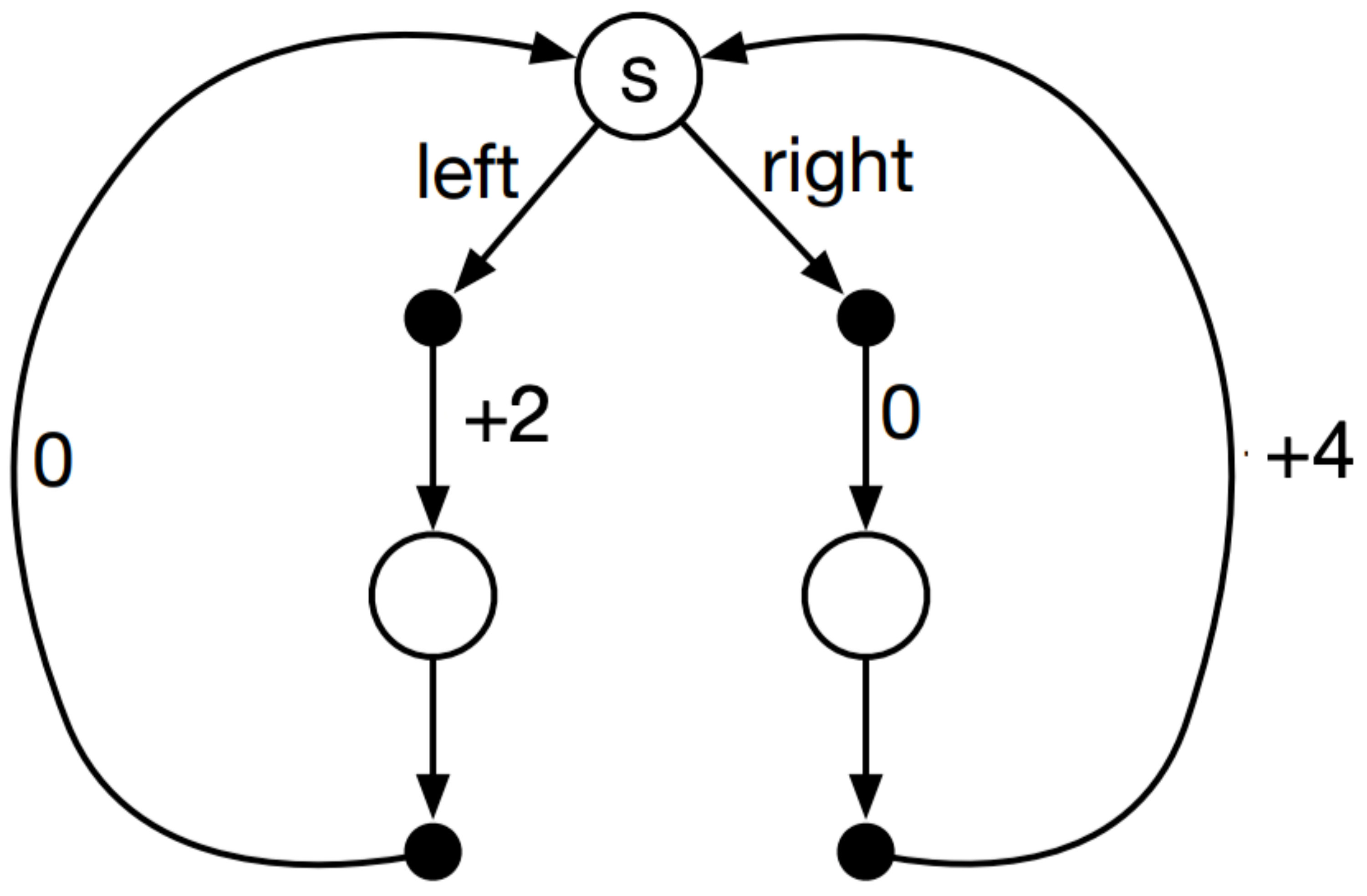
4. (*Exercise 3.22 in* 2^{nd} *ed.*) Consider the continuing MDP shown on the bottom. The only decision to be made is that in the top state, where two actions are available, left and right. The numbers show the rewards that are received deterministically after each action.

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

(b) Is the following policy valid for this MDP (i.e. does if fit our definition of a policy): Choose *left* for five steps, then *right* for five steps, then *left* for five steps, and so on? Explain your answer.

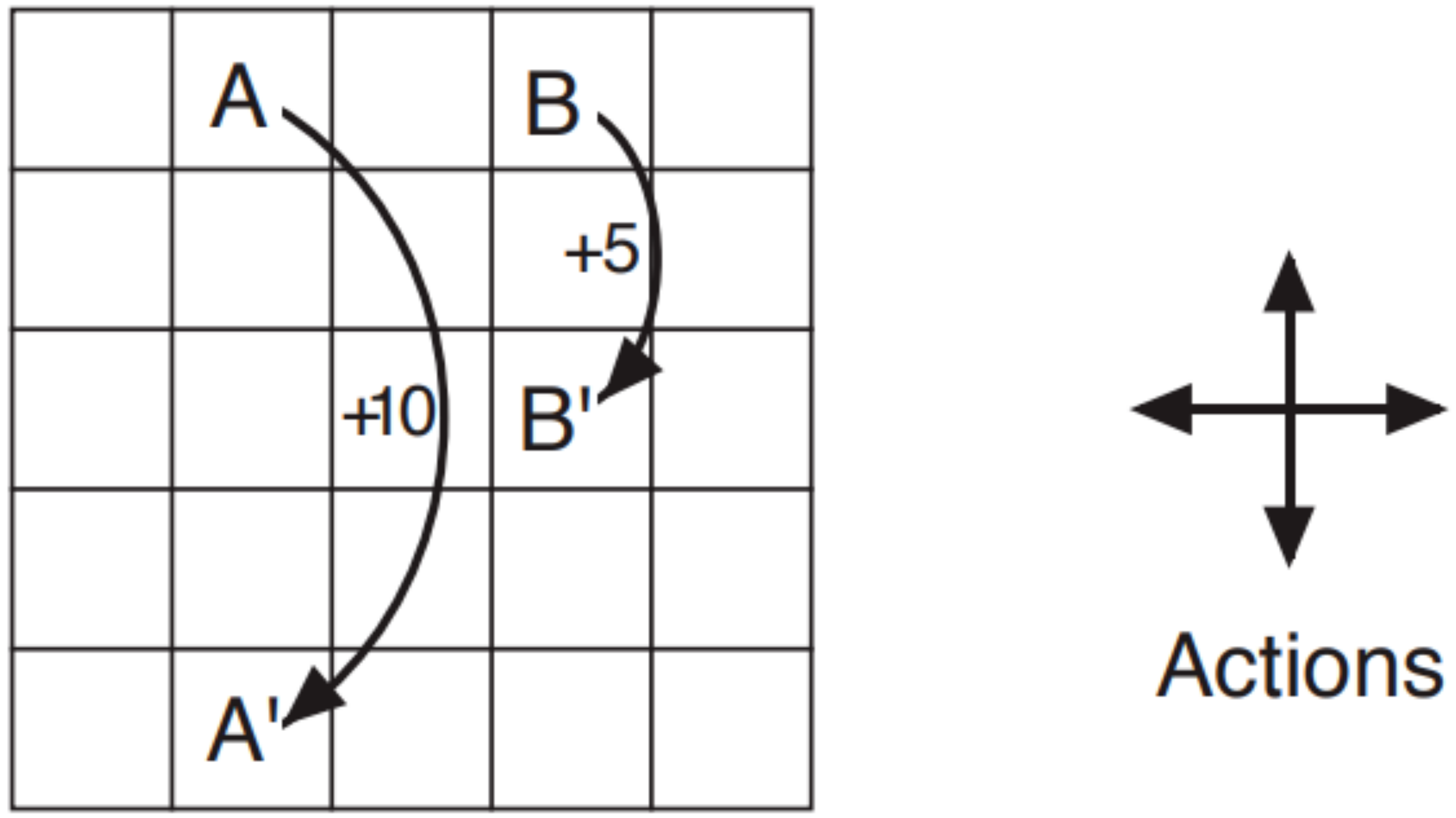
(c) What policy is optimal if $\gamma = 0$? If $\gamma = 0.9$? If $\gamma = 0.5$?

(d) For each possible policy, what is the value of state s? Write down the numeric value to two decimal places. *Hint*: write down the return under each policy starting in state s (don't forget γ). Simplify the infinite sum, using the fact that many rewards are zero. Then plug in the rewards and γ and compute the number.



Worksheet question

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.



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