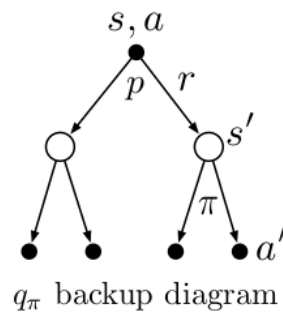


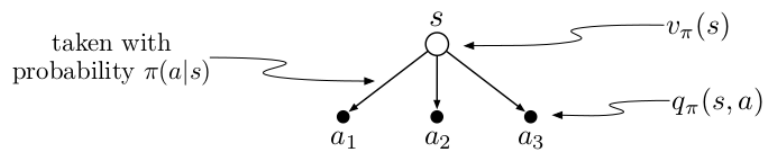
Note: The references mentioned in the exercises refer to the textbook (Sutton and Barto) in the 2nd edition.

## 2 Markov Decision Processes

7. Exercise 3.2 Is the MDP framework adequate to usefully represent all goal-directed learning tasks? Can you think of any clear exceptions?
8. Exercise 3.4 Give a table analogous to that in Example 3.3 (textbook p. 52), but for  $p(s', r|s, a)$ . It should have columns for  $s$ ,  $a$ ,  $s'$ ,  $r$ , and  $p(s', r|s, a)$ , and a row for every 4-tuple for which  $p(s', r|s, a) > 0$ .
9. Exercise 3.6 Suppose you treated pole-balancing as an episodic task but also used discounting, with all rewards zero except for  $-1$  upon failure. What then would the return be at each time? How does this return differ from that in the discounted, continuing formulation of this task?
10. Exercise 3.7 Imagine that you are designing a robot to run a maze. You decide to give it a reward of  $+1$  for escaping from the maze and a reward of zero at all other times. The task seems to break down naturally into episodes - the successive runs through the maze - so you decide to treat it as an episodic task, where the goal is to maximize expected total reward (3.7). After running the learning agent for a while, you find that it is showing no improvement in escaping from the maze. What is going wrong? Have you effectively communicated to the agent what you want it to achieve?
11. Exercise 3.8 Suppose  $\gamma = 0.5$  and the following sequence of rewards is received  $R_1 = -1$ ,  $R_2 = 2$ ,  $R_3 = 6$ ,  $R_4 = 3$ , and  $R_5 = 2$ , with  $T = 5$ . What are  $G_0$ ,  $G_1$ , ...,  $G_5$ ? Hint: Work backwards.
12. Exercise 3.9 Suppose  $\gamma = 0.9$  and the reward sequence is  $R_1 = 2$  followed by an infinite sequence of 7s. What are  $G_1$  and  $G_0$ ?
13. Exercise 3.11 If the current state is  $S_t$ , and actions are selected according to stochastic policy  $v_\pi$ , then what is the expectation of  $R_{t+1}$  in terms of  $\pi$  and the four-argument function  $p$  (3.2)?
14. Exercise 3.14 The Bellman equation (3.14) must hold for each state for the value function  $v_\pi$  shown in Figure 3.2 (right) of Example 3.5. Show numerically that this equation holds for the center state, valued at  $+0.7$ , with respect to its four neighboring states, valued at  $+2.3$ ,  $+0.4$ ,  $-0.4$ , and  $+0.7$ . (These numbers are accurate only to one decimal place.)
15. Exercise 3.15 In the gridworld example, rewards are positive for goals, negative for running into the edge of the world, and zero the rest of the time. Are the signs of these rewards important, or only the intervals between them? Prove, using (3.8), that adding a constant  $c$  to all the rewards adds a constant,  $v_c$ , to the values of all states, and thus does not affect the relative values of any states under any policies. What is  $v_c$  in terms of  $c$  and  $\gamma$ ?
16. Exercise 3.16 Now consider adding a constant  $c$  to all the rewards in an episodic task, such as maze running. Would this have any effect, or would it leave the task unchanged as in the continuing task above? Why or why not? Give an example.
17. Exercise 3.17 What is the Bellman equation for action values, that is, for  $q_\pi$ ? It must give the action value  $q_\pi(s, a)$  in terms of the action values,  $q_\pi(s_0, a_0)$ , of possible successors to the state-action pair  $(s, a)$ . Hint: the backup diagram to the right corresponds to this equation. Show the sequence of equations analogous to (3.14), but for action values.

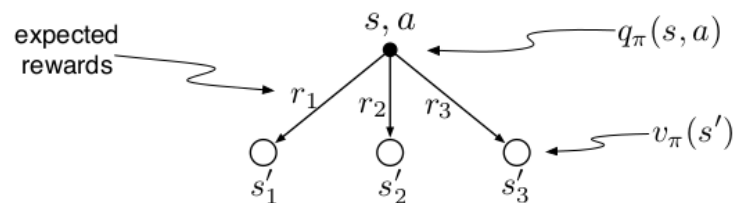


18. Exercise 3.18 The value of a state depends on the values of the actions possible in that state and on how likely each action is to be taken under the current policy. We can think of this in terms of a small backup diagram rooted at the state and considering each possible action:



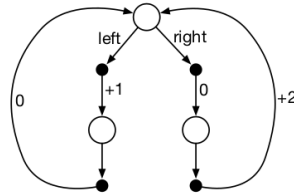
Give the equation corresponding to this intuition and diagram for the value at the root node,  $v_\pi(s)$ , in terms of the value at the expected leaf node,  $q_\pi(s, a)$ , given  $S_t = s$ . This equation should include an expectation conditioned on following the policy,  $\pi$ . Then give a second equation in which the expected value is written out explicitly in terms of  $\pi(a|s)$  such that no expected value notation appears in the equation.

19. Exercise 3.19 The value of an action,  $q_\pi(s, a)$ , depends on the expected next reward and the expected sum of the remaining rewards. Again we can think of this in terms of a small backup diagram, this one rooted at an action (state-action pair) and branching to the possible next states:

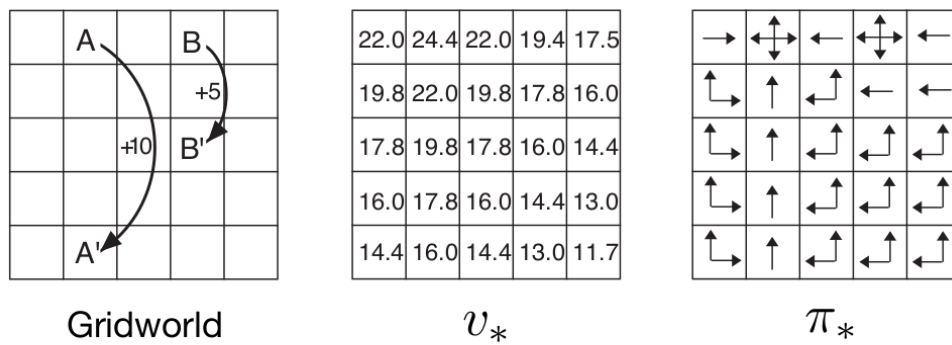


Give the equation corresponding to this intuition and diagram for the action value,  $q_\pi(s, a)$ , in terms of the expected next reward,  $R_{t+1}$ , and the expected next state value,  $v_\pi(S_{t+1})$ , given that  $S_t = s$  and  $A_t = a$ . This equation should include an expectation but not one conditioned on following the policy. Then give a second equation, writing out the expected value explicitly in terms of  $p(s', r|s, a)$  defined by (3.2), such that no expected value notation appears in the equation.

20. Exercise 3.22 Consider the continuing MDP shown below. 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. There are exactly two deterministic policies,  $\pi_{\text{left}}$  and  $\pi_{\text{right}}$ . What policy is optimal if  $\gamma = 0$ ? If  $\gamma = 0.9$ ? If  $\gamma = 0.5$ ?



21. Exercise 3.24 Figure 3.5 (below) gives the optimal value of the best state of the gridworld as 24.4, to one decimal place. Use your knowledge of the optimal policy and (3.8) to express this value symbolically, and then to compute it to three decimal places.



**Figure 3.5:** Optimal solutions to the gridworld example.