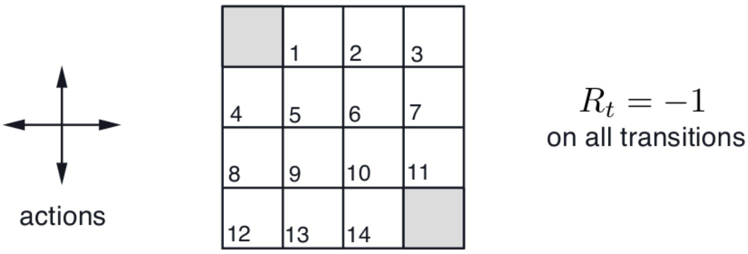
Note: The references mentioned in the exercises refer to the textbook (Sutton and Barto) in the 2nd edition.

3 Dynamic Programming

22. Exercise 4.1 In Example 4.1, if π is the equiprobable random policy, what is $q_\pi(11, down)$?

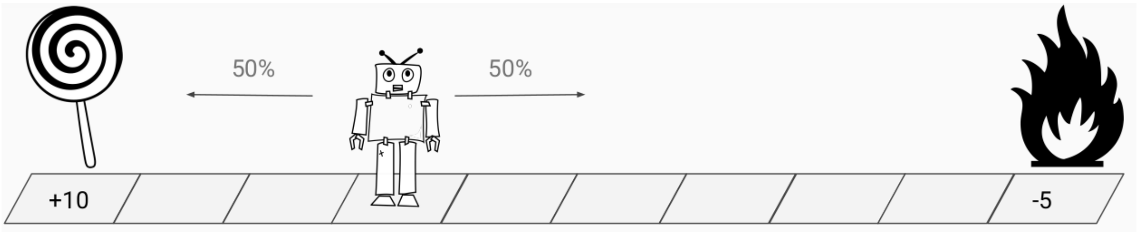


Example 4.1

23. Exercise 4.2 In Example 4.1, suppose a new state 15 is added to the gridworld just below state 13, and its actions, $left$, up, $right$, and $down$, take the agent to states 12, 13, 14, and 15, respectively. Assume that the transitions from the original states are unchanged. What, then, is $v_\pi(15)$ for the equiprobable random policy? Now suppose the dynamics of state 13 are also changed, such that action down from state 13 takes the agent to the new state 15. What is $v_\pi(15)$ for the equiprobable random policy in this case?

24. Implementation Task: 1-D Gridworld

    Consider the following one-dimensional "gridworld": You are on a route consisting of 10 states. The state on the left side is a terminal state with a reward of $+10$ and the state on the right is also a terminal state width a reward of -5. You can move left and right.



1D Gridworld

    Write an implementation of Dynamic Programming for estimating the values of the above states under the equiprobable random policy.

    Then use policy improvement to find the optimal policy.

25. Exercise 4.3 What are the equations analogous to (4.3), (4.4), and (4.5) for the action-value function q_π and its successive approximation by a sequence of functions q_0 , q_1 , q_2 ,... ? (Textbook p. 74)

26. Exercise 4.5 How would policy iteration be defined for action values? Give a complete algorithm for computing q_* , analogous to that on page 80 for computing v_* . Please pay special attention to this exercise, because the ideas involved will be used throughout the rest of the book.

27. Exercise 4.6 Suppose you are restricted to considering only policies that are ϵ-soft, meaning that the probability of selecting each action in each state, s, is at least $\epsilon/|A(s)|$. Describe qualitatively the changes that would be required in each of the steps 3, 2, and 1, in that order, of the policy iteration algorithm for v_* (Textbook p. 80).

28. Exercise 4.10 What is the analog of the value iteration update (4.10) for action values, $q_{k+1}(s,a)$?