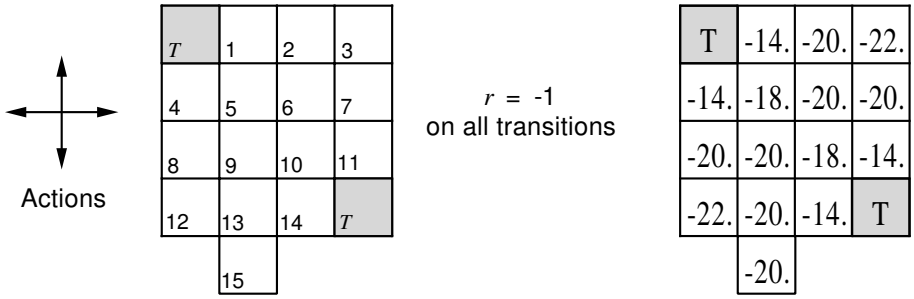
Answers to Exercises
Reinforcement Learning: Chapter 4

Exercise 4.1 If π is the random policy, what is $Q^\pi(11, \texttt{down})$? What is $Q^\pi(7, \texttt{down})$?

Answer: $Q^\pi(11, \texttt{down}) = -1$. $Q^\pi(7, \texttt{down}) = -15$.

Exercise 4.2 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?



Answer: In the case where none of the other states have their outgoing transitions changed, then the new state's value under the random policy is

$$\begin{aligned} V^\pi(15) &= E_\pi\{r_{t+1} + V^\pi(s_{t+1}) | s_t = s\} \\ &= -1 + \frac{1}{4}V^\pi(12) + \frac{1}{4}V^\pi(13) + \frac{1}{4}V^\pi(14) + \frac{1}{4}V^\pi(15) \end{aligned}$$

Plugging in the asymptotic values for $V_\infty = V^\pi$ for states 12, 13, and 14 from Figure 4.1 (and above, right) and solving for $V^\pi(15)$ yields

$$V^\pi(15) = -1 - \frac{1}{4}22 - \frac{1}{4}20 - \frac{1}{4}14 - \frac{1}{4}V^\pi(15)$$

$$V^\pi(15)\left(1 - \frac{1}{4}\right) = -15$$

$$V^\pi(15) = -20$$

If the dynamics of state 13 also change, then it turns out that the answer is the same! This can be most easily seen by hypothesizing that $V^\pi(15) = -20$ and then checking that all states still satify the Bellman equation for V^π.

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 approximation by a sequence of functions $Q_0, Q_1, Q_2, \ldots$?

Answer:

$$\begin{aligned} Q^\pi(s,a) &= E_\pi\{r_{t+1} + \gamma Q^\pi(s_{t+1}, a_{t+1}) | s_t = s, a_t = a\} && (4.3) \\ &= \sum_{s'} \mathcal{P}^a_{ss'} \left[\mathcal{R}^a_{ss'} + \gamma \sum_{a'} \pi(s', a') Q^\pi(s', a') \right] && (4.4) \end{aligned}$$

$$\begin{aligned} Q_{k+1}(s,a) &= E_\pi\{r_{t+1} + \gamma Q_k(s_{t+1}, a_{t+1}) | s_t = s, a_t = a\} \\ &= \sum_{s'} \mathcal{P}^a_{ss'} \left[\mathcal{R}^a_{ss'} + \gamma \sum_{a'} \pi(s', a') Q_k(s', a') \right] && (4.5) \end{aligned}$$

Exercise 4.5 How would policy iteration be defined for action values? Give a complete algorithm for computing Q^*, analogous to Figure 4.3 for computing V^*. Please pay special attention to this exercise because the ideas involved will be used throughout the rest of the book.

Answer: Just as for state values, we would have an alternation of policy improvement and policy evalution steps, only this time in Q rather than in V:

$$\pi_0 \xrightarrow{\text{PE}} Q^{\pi_0} \xrightarrow{\text{PI}} \pi_1 \xrightarrow{\text{PE}} Q^{\pi_1} \xrightarrow{\text{PI}} \pi_2 \xrightarrow{\text{PE}} \cdots \xrightarrow{\text{PI}} \pi^* \xrightarrow{\text{PE}} Q^*$$

Each policy evaluation step, $\pi_i \xrightarrow{\text{PE}} Q^{\pi_i}$, would involve multiple iterations of equation (4.5) above, until convergence, or some other way of computing Q^{π_i}. Each policy improvement step, $Q^{\pi_i} \xrightarrow{\text{PI}} \pi_{i+1}$, would be a greedification with respect to Q^{π_i}, i.e.:

$$\pi_{i+1}(s) = \arg\max_a Q^{\pi_i}(s,a).$$

A boxed algorithm for policy iteration to Q^* is:

1. Initialization
   $\pi \leftarrow$ an arbitrary deterministic policy
   $Q \leftarrow$ an arbitrary function: $\mathcal{S} \times \mathcal{A}(s) \mapsto \Re$
   $\theta \leftarrow$ small positive number

2. Policy Evaluation
   Repeat
      $\Delta \leftarrow 0$
      For each $s \in \mathcal{S}$ and $a \in \mathcal{A}(s)$:
         $q \leftarrow Q(s,a)$
         $Q(s,a) \leftarrow \sum_{s'} \mathcal{P}^a_{ss'}[\mathcal{R}^a_{ss'} + \gamma Q(s', \pi(s'))]$
         $\Delta \leftarrow \max(\Delta, |q - Q(s,a)|)$
   until $\Delta < \theta$

3. Policy Improvement
   policy-improved $\leftarrow false$
   For each $s \in \mathcal{S}$:
      $b \leftarrow \pi(s)$
      $\pi(s) \leftarrow \arg\max_a Q(s,a)$
      If $b \neq \pi(s)$ then *policy-improved* $\leftarrow true$
   If *policy-improved*, then go to 2
                        else stop

In the "arg max" step, it is important that ties be broken in a consistent order.

Exercise 4.6 Suppose you are restricted to consideration only of algorithms that are ε-*soft*, meaning that the probability of selecting each action in each state, s, was at least $\frac{\varepsilon}{|\mathcal{A}(s)|}$. 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^* (Figure 4.3).

Answer: Step 3, the policy improvement step, would have to be changed such that the new policy is not the deterministic greedy policy, but the closest ε-soft policy. That is, all non-greedy actions would be given the minimal probability, $\frac{\varepsilon}{|\mathcal{A}(s)|}$, and all the rest of the probability would go to the greedy action. The check for termination would also need to be changed. Somehow we would have to check for a change in the action with the bulk of the probability.

Step 2, policy evalution, would need to be generalized to accomodate stochastic policies. A new equation analogous to (4.5) would be needed.

Step 1, initialization, would need be changed only to permit the initial policy to be stochastic.

Exercise 4.7 Why does the optimal policy for the gambler's problem have such a curious form? In particular, for capital of 50 it bets it all on one flip, but for capital of 51 it does not. Why is this a good policy?

Answer: In this problem, with $p = 0.4$, the coin is biased against the gambler. Because of this, the gambler want to minimize his number of flips. If he makes many small bets he is likely to lose. Thus, with a stake of 50 he can bet it all and have a .4 probability of winning. On the other hand, with stake of 51 he can do slightly better. If he bets 1, then even if he loses he still has 50 and thus a .4 chance of winning. And if he wins he ends up with 52. With 52 he can bet 2 and maybe end up with 54 etc. In these cases there is a chance he can get up to 75 without ever risking it all on one bet, yet he can always fall back (if he loses) on one big bet. And if he gets to 75 he can safely bet 25, possibly winning in one, while still being able to fall back to 50. It is this sort of logic which causes such big changes in the policy with small changes in stake, particularly at multiples of the negative powers of two.

Exercise 4.9 What is the analog of the value iteration equation (4.9) for action values, $Q_{k+1}(s, a)$?

Answer: Value iteration in action values is defined by

$$\begin{aligned} Q_{k+1}(s,a) &= E\Big\{ r_{t+1} + \gamma \max_{a'} Q_k(s_{t+1}, \alpha) \;\Big|\; s_t = s, a_t = a \Big\} \\ &= \sum_{s'} \mathcal{P}^a_{ss'} \Big[\mathcal{R}^a_{ss'} + \gamma \max_{a'} Q_k(s', a') \Big] \end{aligned}$$

for all $s \in \mathcal{S}$ and $a \in \mathcal{A}(s)$. For arbitrary Q_0, the sequence $\{Q_k\}$ converges to Q^*.