Alena Guseva

916268698

Dr. Ghulam Rasool

Reinforcement Learning

10.17.2019

Please answer: Exercise 4.3, 4.5, 4.8

Programming Write a program to solve Jack's car rental problem. Reproduce plots shown in Fig. 4.2 (page 81). Write a program to solve Gambler's problem. Reproduce plots shown in Fig. 4.3 (page 84).

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 $q_0, q_1, q_2, \ldots$?
□

$$
\begin{aligned}
v_\pi(s) &\doteq \mathbb{E}_\pi[G_t \mid S_t = s] \\
&= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t = s] && \text{(from (3.9))} \\
&= \mathbb{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s] && (4.3) \\
&= \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)\Big[r + \gamma v_\pi(s')\Big], && (4.4)
\end{aligned}
$$

$$
\begin{aligned}
q_\pi(s,a) = \mathbb{E}[G_t|G_t = s, A_t = a] &= \mathbb{E}_\pi[R_{t+1} + \gamma V_\pi(S_{t+1})|S_t = s, A_t = a] \\
&= \mathbb{E}_\pi[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1})|S_t = s, A_t = a] \\
&= \mathbb{E}_\pi[R_{t+1} + \gamma \textstyle\sum_{a'} \pi(a'|s')\, q_\pi(s',a')|S_t = s, A_t = a] \\
&= \sum_{s',r} p(s',r|s,a)[r + \gamma \sum_{a\prime} q_\pi(s',a')\pi(a'|s')]
\end{aligned}
$$

$$
\begin{aligned}
v_{k+1}(s) &\doteq \mathbb{E}_\pi[R_{t+1} + \gamma v_k(S_{t+1}) \mid S_t = s] \\
&= \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)\Big[r + \gamma v_k(s')\Big], && (4.5)
\end{aligned}
$$

$$
q_{k+1}(s,a) = \sum_{s',r} p(s',r|s,a)[r + \gamma \sum_{a\prime} q_k(s',a')\pi(a'|s')]
$$

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. □

Algorithm:

1. Inputs: $q_*(s,a) = 0, for\ all\ s \in S\ and\ a \in A$
2. **the policy evatuation step**

   Repeat

   $\Delta \leftarrow 0$;

   For each $(s,a) \in (S \times A)$ do

   $q \leftarrow q_*(s,a)$;

   $$q_*(s,a) \leftarrow \sum_{s',r} p(s',r|s,a)[r + \gamma \sum_{a'} (s',a')\pi(a'|s')];$$

   $\Delta \leftarrow \max\,(\Delta, |q - q_*(s,a)|)$

   Until $\Delta < \theta\ (small\ positive\ number)$;

3. **To improve policy**

   Policy-stable ←true.

   For each $s \epsilon S$ do

   Old-action ← $\pi(s)$

   $\pi(s) \leftarrow argmax_a\, q_*(s,a)$;

   If old-action $\neq \pi(s), then\ policy - state\ \leftarrow false$.

   If policy-stable, then stop and return $q \approx q_*, and\ \pi \approx \pi_*$;

   Else go to 2.

Exercise 4.8 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? □

At step 50, it doesn't matter if you lose 1 or lose 49, so you might as well go for as much as possible to maximize the potential value. on the other hand, on step 51 you could lose 1 then get back to step 50 and try again.