Róbert Leó Jónsson (robertt20@ru.is))

Chapter 5

Exercise 5.1

The estimated value function jumps up for the last two rows because there, we have cards that sum up to 20 or 21, so we are very likely to win.

It drops off in the whole last row on the left because we are less likely to win if the dealer has an ace.

The frontmsot values are higher in the upper diagrams because aces are useful in blackjack.

Exercise 5.2

It seems like the same state cannot be visited twice in the same game in blackjack - the card's sums will always change after a move.

Exercise 5.3

It is the same as for the value function. It is a path graph, where the nodes alternate between states and actions.

Exercise 5.4

We would keep an integer of how many times we have visited the state-action pair (s, a), for each state, and each action. Then, instead of appending R to the list and repeatedly taking the mean, we would instead use the iterative mean update equation we encountered before.

Exercise 5.5

The rewards are $(R_1, \ldots, R_{10}) = (1, \ldots, 1)$, so the return G_t is

$$\begin{aligned} G_t &= \sum_{i=t+1}^{10} R_i = \sum_{i=t+1}^{10} 1 && (1) \\ &= 10 - t && (2) \end{aligned}$$

Thus, for the first-visit case, our estimate is

$$v(S) = G_0 = 10 - 0 = 10 \qquad (3)$$

However, for the every-step case, the estimate is

$$
\begin{aligned}
v(S) &= \frac{1}{10}\sum_{t=0}^{9} G_t = \frac{1}{10}\sum_{t=0}^{9} 10 - t && (4)\\
&= \frac{1}{10}\sum_{t=0}^{9} 10 - \frac{1}{10}\sum_{t=0}^{9} t && (5)\\
&= 10 - \frac{1}{10}\frac{9(9-1)}{2} && (6)\\
&= 10 - 3.6 = 6.4 && (7)\\
& && (8)
\end{aligned}
$$

Exercise 5.6

Given a starting state S_t and an action A_t, the probability of the subsequent state-action trajectory $S_{t+1}, A_{t+1}, \ldots, S_T$ occuring under any policy π is

$$
\begin{aligned}
&p(S_{t+1} \mid S_t, A_t)\pi(A_{t+1} \mid S_{t+1}) \cdots p(S_T \mid S_{T-1}, A_{T_1}) && (9)\\
&= \frac{1}{\pi(A_t \mid S_t)} \prod_{k=t}^{T-1} \pi(A_k \mid S_k) p(S_{k+1} \mid S_k, A_k) && (10)
\end{aligned}
$$

Thus, instead of $\rho_{t:T-1}$, we usee

$$
\begin{aligned}
\phi_{t:T-1} &= \frac{b(A_t \mid S_t)}{\pi(A_t \mid S_t)} \rho_{t:T-1} && (11)\\
&= \frac{b(A_k \mid S_k)}{\pi(A_k \mid S_k)} \prod_{k=t}^{T-1} \frac{\pi(A_k \mid S_k)}{b(A_k \mid S_k)} && (12)\\
&= \prod_{k=t+1}^{T-1} \frac{\pi(A_k \mid S_k)}{b(A_k \mid S_k)} && (13)\\
&= \rho_{t+1:T-1} && (14)
\end{aligned}
$$

Then the right transformation is

$$
\mathbb{E}\left[\rho_{t+1:T-1} G_t \mid S_t = s, A_t = a\right] = q_\pi(s, a) \qquad (15)
$$

And thus equation (5.6) becomes

$$
Q(s, a) = \frac{\sum_{t \in \mathcal{T}(s)} \rho_{t+1:T(t)-1} G_t}{\sum_{t \in \mathcal{T}(s)} \rho_{t+1:T(t)-1}} \qquad (16)
$$