Reinforcement Learning: An Introduction
Attempted Solutions
Chapter 4

Scott Brownlie & Rafael Rui

1 Exercise 5.1

Consider the diagrams on the right in Figure 5.1. Why does the estimated value function jump up for the last two rows in the rear? Why does it drop off for the whole last row on the left? Why are the frontmost values higher in the upper diagrams than in the lower?

The estimated value function jumps up for the last two rows in the rear because these rows correspond to the player having 20 or 21, which are the only two values on which the player sticks. The player cannot lose when he sticks on 21 and is unlikely to lose when sticking on 20. For values less than 20 the player always hits, whereas the dealer only hits if he has less than 17, which appears a better strategy, hence why the value function is low for values less than 20.

The value function drops off for the whole last row on the left because this row corresponds to the dealer showing an ace. The ace is the best card to have because it can either be counted as 1 or 11 and thus increases the the dealer's chances of winning.

The frontmost values are higher in the upper diagrams because they correspond to the player having a usable ace. In this case the player has less chance of going bust. For example, if he has an ace and a 1, giving a total of 12, and he hits and receives a 10, the ace will then be counted as 1 and the player will still be in the game.

2 Exercise 5.2

Suppose every-visit MC was used instead of first-visit MC on the blackjack task. Would you expect the results to be very different? Why or why not?

The only way that the player can have the same sum more than once in a single episode is if he has a usable ace. For example, if he has an ace and a 5 then his sum will be 16. If he then receives a 6 then, to avoid going bust, the ace will be counted as 1 and his sum will be $1+5+6 = 12$. There is now the possibility that he once again ends up with a sum of 16 (by receiving a 4, for example). However, the same sum with and without a usable ace are classed as different states, hence it is impossible for the player to visit the same state more than once in a single episode and the estimated value functions will be exactly the same.

3 Exercise 5.3

What is the backup diagram for Monte Carlo estimation of q_π?

4 Exercise 5.4

The pseudocode for Monte Carlo ES is inefficient because, for each state–action pair, it maintains a list of all returns and repeatedly calculates their mean. It would be more efficient to use techniques similar to those explained in Section 2.4 to maintain just the mean and a count (for each state–action pair) and update them incrementally. Describe how the pseudocode would be altered to achieve this.

We would initialise $N(s, a) \leftarrow 0$ for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$, and then on each first visit to the pair S_t, A_t we would increment $N(S_t, A_t)$ by 1 and update $Q(S_t, A_t)$ as follows:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{1}{N(S_t, A_t)} \Big[G - Q(S_t, A_t) \Big].$$

5 Exercise 5.5

Consider an MDP with a single nonterminal state and a single action that transitions back to the nonterminal state with probability p and transitions to the terminal state with probability $1 - p$. Let the reward be $+1$ on all transitions, and let $\gamma = 1$. Suppose you observe one episode that lasts 10 steps, with a return of 10. What are the first-visit and every-visit estimators of the value of the nonterminal state?

The first-visit estimate of the value of the nonterminal state is 10. The every-visit estimate is

$$\frac{1}{10} \Big(10 + 9 + 8 + \cdots + 1 \Big) = 5.5.$$

6 Exercise 5.6

What is the equation analogous to (5.6) for action values $Q(s, a)$ instead of state values $V(s)$, again given returns generated using b?

It is

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

where $\mathcal{T}(s, a)$ denotes the set of all time steps in which action a is taken in state s.