June 19, 2022

# 5 Exercises

## 5.1 Question

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?

### Answer

No answer provided.

## 5.2 Question

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?

### Answer

The results would be the same because a state can be visited only once.

## 5.3 Question

What is the backup diagram for Monte Carlo estimation of q pi ?

### Answer

The backup diagram for MC is given at page 95. An empty circle and a black filled circle represent a state action pair.

### 5.4 Question

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.

### Answer

In section 2.4 average return is calculated iteratively using only current average and number of occurrences.

Average return for each state-action pair can be calculated similarly.

$$Q_(S, A) = Q_(S, A) + (G - Q_(S, A))/N(S, A) \tag{1}$$

### 5.5 Question

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?

### Answer

There is no separate behaviour policy. Thus importing sampling ratio is 1.

First visit case:

$v(s) = 1 * 10/1 = 10$

Every visit case:

$v(s) = 1*10+1*9+1*8+1*7+1*6+1*5+1*4+1*3+1*2+1*1/10 = 55/10 = 5.5$

Results are similar if weighted estimator is used.

### 5.6 Question

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?

**Answer**

$$Q(s,a) = \frac{\sum_{t \in \tau(s,a)} Pt : (T(t) - 1) G_t]}{\sum_{t \in \tau(s,a)} Pt : (T(t) - 1)} \tag{2}$$