8 Reinforcement Learning

8.1 Questions

Exercise 8.1

Tell if the following statements are true or false and provide the adequate motivations to your answer.

1. In RL we do not require to have the model of the environment;
2. In RL we do not represent the model of the environment;
3. We need to use data coming from the optimal policy if we want to learn it;
4. Since RL sequentially decide the action to play at each time point, we cannot use information provided by historical data;
5. We can manage continuous space with RL.

Exercise 8.2

Tell if the following properties hold for MC or TD and motivate your answers.

1. Can be applied to infinite horizon ML;
2. Can be applied to indefinite horizon ML;
3. Needs an entire episode;
4. Works step by step (online);
5. Applies bootstrap;
6. The number of samples depends on the dimension of the MDP;
7. The number of samples depends on the length of the episodes;
8. Solves the prediction problem;

9. Reuse the information learned from past learning steps;
10. Makes use of the Markov property of the MDP;
11. Has no bias;
12. Has some bias.

Exercise 8.3

Tell if the following statements are true or false and motivate your answers.

1. With MC estimation you can extract a number of samples for the value function equal to the length of the episode you consider for prediction;
2. Generally, every-visit estimation is better if you use a small amount of episodes;
3. Stochasticity in the rewards requires the use of a larger number of episode to have precise prediction of the MDP value in the case we use MC estimation;
4. MC estimation works better than TD if the problem is not Markovian.

Exercise 8.4

Tell if the following statements are true or false and motivate your answers.

1. To compute the value of a state TD uses an approach similar to the one used in the Policy Evaluation algorithm;
2. TD updates its prediction as soon as a new tuple (state, action, reward, next state) is available;
3. TD cannot be used in the case there is no terminal state in the original MDP;
4. Since with TD we use values computed by averaging, we introduce less variance in the estimation than MC.

Exercise 8.5

Evaluate the value for the MDP with three states $\mathcal{S} = \{A, B, C\}$ (C is terminal), two actions $\mathcal{A} = \{h, r\}$ given the policy π, given the following trajectories:

$$
\begin{aligned}
&(A, h, 3) \to (B, r, 2) \to (B, h, 1) \to (C) \\
&(A, h, 2) \to (A, h, 1) \to (C) \\
&(B, r, 1) \to (A, h, 1) \to (C)
\end{aligned}
$$

1. Can you tell without computing anything if by resorting to MC with every-visit and first-visit approach you will have different results?
2. Compute the values with the two aforementioned methods.
3. Assume to consider a discount factor $\gamma = 1$. Compute the values by resorting to TD? Assume to start from zero values for each state and $\alpha = 0.1$.

Exercise 8.6

Comment on the use of α in the stochastic approximation problem to estimate an average value:

$$\mu_i = (1 - \alpha_i)\mu_{i-1} + \alpha_i x_i$$

Is $\alpha_i = \frac{1}{i}$ a valid choice? Is $\alpha = \frac{1}{i^2}$ meaningful?

Exercise 8.7

Consider the following problems and tell when the optimal policy can be found by resorting to RL or DP techniques:

1. Maze Escape
2. Pole balancing problem
3. Ads displacement
4. Chess

Exercise 8.8

Tell if the following statements are true or false.

1. To converge to the optimal policy we can even use MC estimation and a greedy policy;
2. To ensure convergence we should ensure that all the states are visited during the learning process;

3. It is not possible to learn the optimal policy by running a different policy on an MDP;
4. Information gathered from previous experience can not be included in the RL learning process.

Provide adequate motivations for your answers.

Exercise 8.9

You want to apply RL to train an AI agent to play a single-player videogame. The state of the game is fully observable and, at each step, the agent has to select an action from a discrete set of possibilities. The interaction ends as soon as the agent reaches the end of the level or fails. To optimize the policy for your AI, you have a set of recorded trajectories (i.e., sequences of state, action, and reward) of the AI agent playing the game following a suboptimal policy. Unfortunately, most of these trajectories are not complete (i.e., they do not cover all the interactions from the beginning of the level to either the end, or to a game-over state).

Indicate if the following methods can be applied to this problem, motivating your answer.

1. Monte Carlo Policy Iteration;
2. Value Iteration;
3. Sarsa;
4. Q-Learning.

Exercise 8.10

Consider the following snippet of code and answers to the questions below providing adequate motivations.

```
while m < M:
  ns, r = env.transition_model(a)
  na = eps_greedy(s, Q, eps)
  Q[s, a] = Q[s, a] + alpha * (r + env.gamma * Q[ns, na] - Q[s, a])
  m = m + 1
  s = ns
  a = na
```

1. What algorithm is this code implementing? What kind of problem is it addressing?
2. Explain the operations performed by the `eps_greedy` function.

3. What conditions do we need on `alpha` and `eps` to make the algorithm converge to a desirable solution?

4. How can we modify Line 4 to make the algorithm work off-policy?

Exercise 8.11

Consider the following episode obtained by an agent interacting with an MDP having two states $\mathcal{S} = \{A, B\}$ and two actions $\mathcal{A} = \{l, r\}$,

$$(A, l, 1) \rightarrow (A, l, 1) \rightarrow (A, r, 0) \rightarrow (B, r, 10) \rightarrow (B, l, 0) \rightarrow (A, r, 0) \rightarrow (B, l, 0) \rightarrow (A).$$

Answer to the following questions providing adequate motivations.

1. Execute the *Q-learning* algorithm on the given episode considering initial state-action values $Q(S, a) = 0$ for every state-action pair, learning rate $\alpha = 0.5$, and discount factor $\gamma = 1$.

2. Provide the best policy according to the output of *Q-learning*.

3. Do you think that the agent fully exploited the policy learned in the episode above? Make a consistent guess with the available information.

Exercise 8.12

We are given an Heating, Ventilation, and Air Conditioning (HVAC) in which the states are cold (c), medium (m), warm (w) temperature. We can perform three actions: heat (h), refrigerate (r), and do nothing (d). Assume to have the following partial episodes for the HVAC functioning.

$$(c, d, 0) \rightarrow (c, h, 1) \rightarrow (m, h, 1) \rightarrow (m, h, -1) \rightarrow (w, r, 1) \rightarrow (m, \cdot, \cdot) \rightarrow \dots$$
$$(m, r, -2) \rightarrow (c, h, -2) \rightarrow (c, h, 1) \rightarrow (m, h, 1) \rightarrow (m, h, 1) \rightarrow (w, \cdot, \cdot) \rightarrow \dots$$

where a tuple (S, A, R) correspond to the State, Action, and Reward at a specific time.

1. Model it as an MDP and draw the corresponding graphical representation, specifying the transition probabilities and rewards (estimated from the episodes) for each transition.

2. Can you tell if the reward of this process is stochastic or deterministic? And what about the transitions?

3. Assuming we want to evaluate the performance of the HVAC, tell which kind of problem we are in and suggest a technique to solve it.