9 Reinforcement Learning

9.1 Questions

Exercise 9.1

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 update the exploration policy over time while learning the optimal policy;
4. Since RL sequentially decides 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 9.2

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 9.3

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 9.4

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 9.5

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:

$$(A, h, 3) \to (B, r, 2) \to (B, h, 1) \to (C)$$
$$(A, h, 2) \to (A, h, 1) \to (C)$$
$$(B, r, 1) \to (A, h, 1) \to (C)$$

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 of the value function estimated by the two aforementioned methods.
3. Compute the value function by resorting to TD. Assume to have a discount factor $\gamma = 1$, to start from zero values for each state, and $\alpha = 0.1$.

Exercise 9.6

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 9.7

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 9.8

Tell if the following statements are true or false.

1. To converge to the optimal policy we can iteratively use an MC estimation step and a greedy policy improvement step;
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 9.9

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 9.10

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 is still exploring in the given environment?

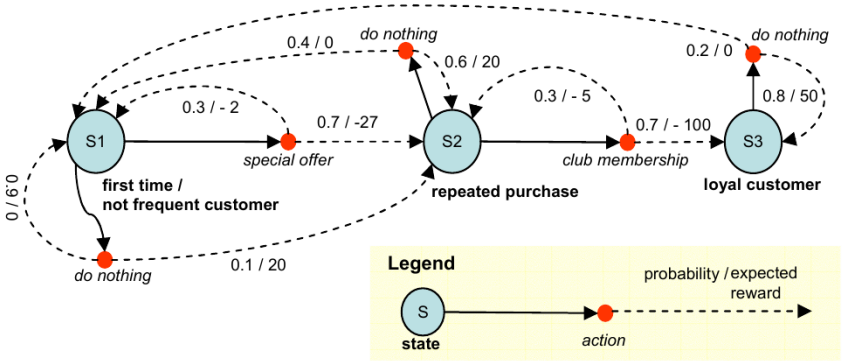

Figure 9.1: MDP corresponding to the advertising problem.

Exercise 9.11

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.

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

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.

Exercise 9.12

Consider the case in which you have an estimated state/action value of $Q(s, a) = 3$ you perform action a, gain a reward of $R_t = 1$ and reach state s'. In s' you could perform only actions a'_1 and a'_2 ($Q(s', a'_1) = 1$ and $Q(s', a'_2) = 2$). Moreover, if you would use the current policy π you would choose action a'_1 in state s'.

Consider a discount factor $\gamma = 0.5$ and a learning rate $\alpha = 0.5$. Tell if the following values are consistent with the use of SARSA and/or Q–learning algorithms after the update:

1. $Q(s, a) = 2.75$
2. $Q(s, a) = 2.25$
3. $Q(s, a) = 3$
4. $Q(s, a) = -2.5$

Motivate the answers you provided.

Exercise 9.13

Assume to have an MDP with four states $\mathcal{S} = \{H, M, L, F\}$ (F is terminal), two actions $\mathcal{A} = \{r, w\}$ and a discount factor $\gamma = 1$. Given the following trajectories:

$$(H, r, 2) \rightarrow (L, r, 3) \rightarrow (M, r, 2) \rightarrow (F)$$
$$(H, w, 2) \rightarrow (H, r, 3) \rightarrow (M, w, 1) \rightarrow (F)$$

1. Compute the values of the different states by resorting to first-visit and every-visit MC.
2. Using a learning rate $\alpha = 0.5$, compute the state values by resorting to TD. Assume to start from zero values for each state.
3. Can you tell if the previously defined MDP is deterministic or stochastic?

Exercise 9.14

Consider the MDP modeling an advertising problem in Figure 9.1. where on the transition probabilities and the rewards are specified on the edges.

1. Provide the formulation of the Bellman expectation for V equations for the MDP in the figure in the case we consider the policy: $\pi(s_1; dn) = 1$ and $\pi(s_2; dn) = 1$ and with discount factor $\gamma = 0.5$.
2. Compute the value of state 2, i.e., $V(s_2)$ (justify your computations).

Exercise 9.15

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

1. Applying MC estimation on a single episode, you extract a number of samples for the value function equal to the length of the episode;

2. Applying MC estimation on a single episode, you extract a number of samples for the value function less or equal to the number of states of the MDP;
3. TD cannot be used in the case we are analysing an MDP with no terminal state;
4. MC every visit is a consistent, but biased, estimator for the state value function of the MDP.

Exercise 9.16

Consider the MDP in Figure 9.1.

- Provide the optimal policy for a discount factor of $\gamma = 1$;
- Provide the optimal policy for a discount factor of $\gamma = 0.5$ (you can justify your answer basing on what has been shown during the lectures and exercise sessions);
- Provide the equations which computes the state-value function of state $S2$ in the case we follow a policy:

  $$(do\ nothing, club\ membership, do\ nothing)$$

  and a discount factor of $\gamma = 0.1$ (you are not required to invert a matrix).

Exercise 9.17

Which method would you choose to solve the following **control problems**:

- Advertisement problem (the one with three states we saw during classes);
- Atari Games (hint: we are able to generate as many episodes as we want);
- Poker;
- Black Jack.

Motivate your answer.

Exercise 9.18

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

$$(A, h, 3) \rightarrow (B, r, 2) \rightarrow (C, h, 1) \rightarrow (D)$$
$$(C, h, 2) \rightarrow (A, h, 1) \rightarrow (D)$$
$$(B, r, 1) \rightarrow (A, h, 1) \rightarrow (D)$$

1. Do you think that a total reward maximization ($\gamma = 1$) is possible in this MDP?
2. Compute the approximation of the state-value function of the MDP by using MC first-visit and every-visit.
3. Assume to consider a discount factor $\gamma = 0.5$. Compute the state-value function by resorting to TD(0). Assume to start from zero values for each state and $\alpha = 0.5$.

Exercise 9.19

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

1. Reinforcement Learning (RL) techniques use a tabular representation of MDPs to handle continuous state and/or action spaces;
2. We can use data coming from sub-optimal policies to learn the optimal one;
3. In RL we always estimate the model of the environment;
4. In RL we require to have full knowledge of the environment.

Exercise 9.20

Tell if the following statements about Reinforcement Learning (RL) are true or false. Motivate your answers.

1. The value function estimation provided by $TD(0)$ is equivalent to the Monte Carlo one.
2. The use of an ε-greedy policy in control RL problems is required to incentivize exploitation.
3. Eligibility traces are used to distribute the instantaneous reward over multiple time steps.
4. Importance sampling allows to re-use experience generated by old policies in RL algorithms.

Exercise 9.21

Tell if the following statements about Reinforcement Learning are true or false, and motivate your answers.

1. For policy evaluation, if I have a simulator of the RL task, I can use DP, but not Monte Carlo or TD.
2. Off-policy learning, Exploring Starts, and Soft-policies are three ways to deal with the Exploration-Exploitation dilemma.
3. The sparsity of the reward can be a problem for on-policy algorithms.
4. SARSA, as Value Iteration in DP, is based on the Bellman Expectation Equation.
5. Applying a TD approach to a problem, I can better exploit the markovianity of the state.
6. A Markov Decision Process can always be solved analytically.
7. Differently from Q-learning, SARSA cannot handle the exploration-exploitation trade-off.
8. In an MDP a stochastic policy cannot be optimal.

Exercise 9.22

Tell whether the following statements about MDP and RL are true or false. Motivate your answers.

1. Policy Evaluation always outputs the optimal value function.
2. The value function may decrease on some steps of Policy Iteration, but in the end, the algorithm outputs the optimal one.
3. Employing a discount factor in the computation of the cumulative return in MDPs is only a mathematical trick to ensure the convergence of the return.
4. Dealing with the exploration-exploitation tradeoff is more crucial in SARSA than Q-learning.
5. The optimal policy of a Multi-Armed Bandit problem can be found with Value-Iteration.
6. Given an MDP with a certain reward function, there is only a single policy that is optimal for it, and, for each optimal policy, there is only a single reward function for which it is optimal.

7. In an MDP, for each state, we can always choose an action that is optimal, independently from the time, and from the past history.
8. It is always better to perform Policy Evaluation by solving a linear system instead of using Dynamic Programming.
9. All sequential decision problems can be modeled as MDPs.
10. As many policies can be optimal, there can be multiple optimal value functions in an MDP.

Exercise 9.23

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

```
1  while m < M:
2    ns, r = env.transition_model(a)
3    na = eps_greedy(s, Q, eps)
4    Q[s, a] = Q[s, a] + alpha * (r + env.gamma * Q[ns, na] - Q[s, a])
5    m = m + 1
6    s = ns
7    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 9.24

Indicate whether the following statements about Monte Carlo and Temporal Difference are true or false. Motivate your answers.

1. If I am trying to evaluate a policy on a small number of interactions, it is generally better to use TD rather than MC methods.
2. MC evaluation would be a reasonable choice for providing an online estimation of a policy on a stream of data.
3. The MC every-visit evaluation suffers a smaller bias in comparison to MC first-visit evaluation.

4. The TD evaluation method is consistent, which means that it will surely converge to the optimal value function if sufficient data is available.