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?

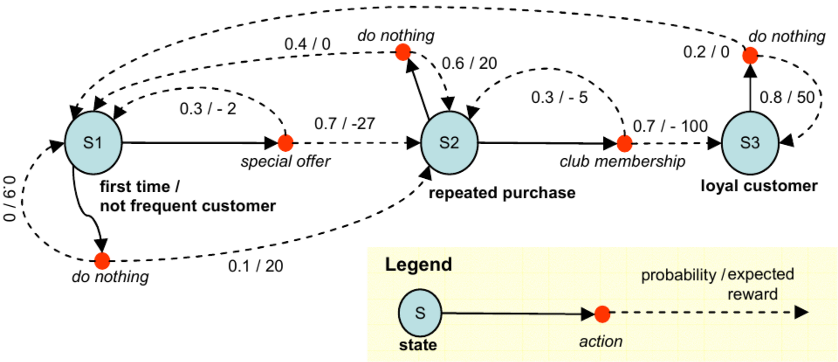


Figure 9.1: MDP corresponding to the advertising problem.

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:

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

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:

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

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.

Solutions

Answer of exercise 9.1

1. TRUE This is the difference with dynamic programming approaches;
2. FALSE In model based RL we do not have a known model of the environment, but we built an approximation of the model of the environment by basing on the data. This statement is true for model free RL approaches, which do not learn an MDP explicitly;
3. FALSE It is possible to keep the exploration policy fixed and use off-policy RL;
4. FALSE For instance, in the case of model-free prediction we might use data coming from previously executed policies and process them as a single batch;
5. TRUE We might resort to function approximation if the state space is continuous or if the complexity of the problem does not allow us to solve the problem.

Answer of exercise 9.2

1. TD since it might operate even without having complete episodes;
2. MC since is able to use episodes with different length and TD since it might be updated after each transition;
3. MC (and TD(∞)) since it needs the cumulative reward at the end of the episode;
4. TD since it updates the value function after each instantaneous reward;
5. TD since it makes computes the estimates by using the information learned so far. This also applies to MC in its incremental version with $\alpha \neq \frac{1}{n}$, n being the number of states processed so far;
6. MC (first-visit) since you will get only one sample per state in each episode (assuming to have fewer states than transitions in an episode);
7. TD and MC (every-visit) since the more the episodes are long the more the sample we have;
8. MC and TD since both returns the value for each state given the observed policy;
9. TD since it uses a bootstrap strategy (and MC with $\alpha \neq \frac{1}{n}$);
10. TD since it explicitly uses it to compute the state value function;

11. MC (first-visit) since the computed value is an unbiased estimator of the state value function;
12. TD and MC (every-visit) since they make use of biased state values and dependent samples, respectively.

Answer of exercise 9.3

1. TRUE/FALSE With every-visit we have a sample per states in the episode, while if we consider first-visit we will count the occurrence of a state only once, thus, we have at most a number of samples equal to the number of distinct states $|S|$;
2. TRUE Even if it is a biased estimator for the expected value of a state, by resorting to every-visit MC we are gathering more samples, thus, we reduce the variance, which in the case we have only few samples is crucial;
3. TRUE The presence of stochasticity increases the variance of the estimates, thus, the estimated values are more and more uncertain;
4. TRUE MC does not use any information about the transition model of the MDP, thus no assumption on the transition model is considered when using MC. Instead, TD explicitly makes use of the Markovian property of MDPs.

Answer of exercise 9.4

1. TRUE The estimated return in the TD equation is nothing else but the right hand side of the Bellman expectation equation in the case we have a single action and we know the transition we perform;
2. TRUE It requires only a single transition to update the value of a state, while for instance MC needs to have a complete episode before updating the estimation of the values of the MDP;
3. FALSE Since its update are performed at each transition, we do not require to have finite episodes. This is different from MC for which we need to wait for the end of the episode;
4. TRUE By considering TD we are introducing some bias (in the estimates of the state values), but the fact that they are computed by averaging more transitions (i.e., we are considering multiple state, action, reward tuples) decreases the variance of the estimates. Moreover, we know that the estimate is consistent, thus, if we consider an infinite number of transitions it becomes unbiased.

Answer of exercise 9.5

1. Since we have multiple instances of the same state in a single episode and their value is not the same, the two approach will provide different results.

2. Every-visit MC:

$$\begin{aligned} V(A) &= \frac{6+3+1+1}{4} = \frac{11}{4} \\ V(B) &= \frac{3+1+2}{3} = 2 \end{aligned}$$

   First-visit MC:

$$\begin{aligned} V(A) &= \frac{6+3+1}{3} = \frac{10}{3} \\ V(B) &= \frac{3+2}{2} = \frac{5}{2} \end{aligned}$$

3. TD:

$$\begin{aligned}
&V(A) \leftarrow V(A) + 0.1(3 + V(B) - V(A)) = 0 + 0.1(3 + 0 - 0) = 0.3 \\
&V(B) \leftarrow V(B) + 0.1(2 + V(B) - V(B)) = 0 + 0.1(2 + 0 - 0) = 0.2 \\
&V(B) \leftarrow V(B) + 0.1(1 + V(C) - V(B)) = 0.2 + 0.1(1 + 0 - 0.2) = 0.28 \\
&V(A) \leftarrow V(A) + 0.1(2 + V(A) - V(A)) = 0.3 + 0.1(2 + 0.3 - 0.3) = 0.5 \\
&V(A) \leftarrow V(A) + 0.1(1 + V(C) - V(A)) = 0.5 + 0.1(1 + 0 - 0.5) = 0.55 \\
&V(B) \leftarrow V(B) + 0.1(1 + V(A) - V(B)) = 0.28 + 0.1(1 + 0.55 - 0.28) = 0.407 \\
&V(A) \leftarrow V(A) + 0.1(1 + V(C) - V(A)) = 0.55 + 0.1(1 + 0 - 0.55) = 0.595
\end{aligned}$$

Answer of exercise 9.6

We know that if $\sum_{i\geq 0} \alpha_i = \infty$ and $\sum_{i\geq 0} \alpha_i^2 < \infty$ the estimator μ_i is consistent, thus, it converges to the real mean of the approximating problem when $i \to \infty$.

Thus the former choice $\alpha_i = \frac{1}{i}$ will provide a consistent estimator, while the latter $\alpha_i = \frac{1}{i^2}$ will be not guarantee to converge.

Answer of exercise 9.7

Any time you are able to use DP for solving a problem you might also consider to find the corresponding approximate solution with RL, thus, DP $\Rightarrow$ RL.

1. RL: we do not have the complete knowledge of the environment we are in, thus we can not resort to DP.

2. DP and RL: we have complete information about the transition model and of the reward function. We might resort to RL if we consider continuous state and/or action space, which is usually difficult to handle with DP.

3. RL: in this case the transition is known (single state), but we do not know the reward associated to each action. In the specific, we might use MAB techniques to solve it.

4. RL: we have complete information, given a fixed strategy of the opponent, but usually we do not resort to DP for computational complexity issues.

Answer of exercise 9.8

1. FALSE If we use a greedy policy it might happen that we are not exploring some actions at all, while they are the optimal ones. This is especially true if the reward you gain or the transitions from the states are stochastic.

2. TRUE The GLIE property requires that if we are considering an arbitrarily long learning process it will visit all the states an infinite number of times:

3. FALSE If we consider an off-policy learning method we are able to converge to the optimal policy even if we do not run the learned policy;

4. FALSE If we consider an off-policy learning method we might extract information about the optimal policy even if we consider data coming from sub-optimal ones.

Answer of exercise 9.9

To provide a solution for the described scenario we need to solve a control problem. We need to use an online method since we do not have trajectories which are complete. Moreover, one might not resort to a dynamic programming approach, since we do not have a full description of the environment, but only trajectories. Finally, we need to have an off-policy method since we can only rely on trajectories collected in the past. Therefore:

1. Not a viable option since Monte Carlo requires complete trajectories;

2. Not a viable option since we would require a complete description of the environment;

3. Not a viable option since it requires to follow the policy provided by it. (Using importance sampling might be a solution)

4. Viable option since it is off-policy, online and uses only transitions.

Answer of exercise 9.10

1. We start with the initial values $Q(A,l) = Q(A,r) = Q(B,l) = Q(A,r) = 0$. Then, we compute the *Q-learning* update $Q(S_t, a_t) = (1-\alpha)Q(S_t,a_t) + \alpha(R_t + \gamma \max_{a \in \{l,r\}} Q(S_{t+1}, a))$ for every step t in the episode:

   a) $Q(A,l) \leftarrow 0.5Q(A,l) + 0.5(1 + Q(A,\cdot)) = 0.5 \cdot 0 + 0.5 \cdot 1 = 0.5$

   b) $Q(A,l) \leftarrow 0.5Q(A,l) + 0.5(1 + Q(A,l)) = 0.5 \cdot 0.5 + 0.5 \cdot 1.5 = 1$

   c) $Q(A,r) \leftarrow 0.5Q(A,r) + 0.5(0 + Q(B,\cdot)) = 0.5 \cdot 0 + 0.5 \cdot 0 = 0$

   d) $Q(B,r) \leftarrow 0.5Q(B,r) + 0.5(10 + Q(B,\cdot)) = 0.5 \cdot 0 + 0.5 \cdot 10 = 5$

   e) $Q(B,l) \leftarrow 0.5Q(B,l) + 0.5(0 + Q(A,l)) = 0.5 \cdot 0 + 0.5 \cdot 1 = 0.5$

   f) $Q(A,r) \leftarrow 0.5Q(A,r) + 0.5(0 + Q(B,r)) = 0.5 \cdot 0 + 0.5 \cdot 5 = 2.5$

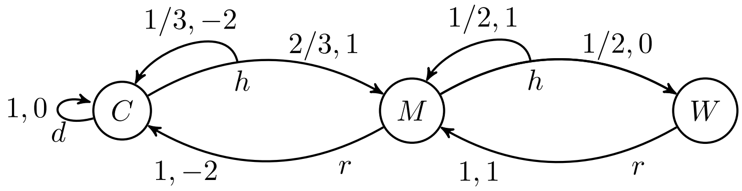
   g) $Q(B,l) \leftarrow 0.5Q(B,l) + 0.5(0 + Q(A,r)) = 0.5 \cdot 0.5 + 0.5 \cdot 2.5 = 1.75$

   which results in $Q(A,l) = 1, Q(A,r) = 2.5, Q(B,l) = 1.75, Q(B,r) = 5$.

2. We select the greedy policy $\pi(S) \in \arg\max_{a \in \{l,r\}} Q(S,a)$ w.r.t. the state-action values obtained with *Q-learning*, which gives $\pi(A) = r, \pi(B) = r$.

3. Yes, since the episodes contains states in which the policy is stochastic (unless in every states all the actions are optimal).

Answer of exercise 9.11

1. 

2. The transition are stochastic since some of the actions are leading to two different new states, e.g., the action heat (h) in the state cold (c). The reward is stochastic as well, since heating in the medium state provided once the reward 1 and once -1.

3. This setting suggests it is an MDP prediction problem, either using directly the original episodes (using MC or TD) or one might use the estimated model and use DP techniques to solve it in an exact way, due to the limited dimension of the problem.

Answer of exercise 9.12

The update rules in this case of SARSA and Q–learning are, respectively:

$$\begin{aligned}Q(s,a) &\leftarrow Q(s,a)+\alpha(R_t+\gamma Q(s',a'_1)-Q(s,a))\\ &=3+0.5(1+0.5\cdot 1-3)=2.25\\ Q(s,a) &\leftarrow Q(s,a)+\alpha(R_t+\gamma \max_{a'} Q(s',a')-Q(s,a))\\ &=3+0.5(1+0.5\cdot 2-3)=2.5\end{aligned}$$

Thus:

1. Inconsistent with both;
2. Consistent with SARSA;
3. Inconsistent with both;
4. Inconsistent with both.

Answer of exercise 9.13

MC First visit	MC Every visit
$V(H)=\frac{7+6}{2}=\frac{13}{2}$	$V(H)=\frac{7+6+4}{3}=\frac{17}{3}$
$V(L)=5$	$V(L)=5$
$V(M)=\frac{2+1}{2}=\frac{3}{2}$	$V(M)=\frac{2+1}{2}=\frac{3}{2}$

The TD update rule is:

$$\begin{aligned}V(s_t) &\leftarrow V(s_t)+\alpha(R_t+\gamma V(s_{t+1})-V(s_t))=V(s_t)\\ V(s_t) &\leftarrow V(s_t)+0.5(R_t+V(s_{t+1})-V(s_t))\end{aligned}$$

thus we have:

$$\begin{aligned}V(H) &\leftarrow 0+0.5(2+0-0)=1\\ V(L) &\leftarrow 0+0.5(3+0-0)=1.5\\ V(M) &\leftarrow 0+0.5(2+0-0)=1\\ V(H) &\leftarrow 1+0.5(2+1-1)=2\\ V(H) &\leftarrow 2+0.5(3+1-2)=3\\ V(M) &\leftarrow 1+0.5(1+0-1)=1\end{aligned}$$

Since there is a state action pair with different rewards i.e., (H,r), the MDP can not be deterministic.

Answer of exercise 9.14

1. The Bellman equation for a state is of the form:

$$V^{\pi}(s) = \sum_{a} \pi(s,a) \left(R(s,a) + \gamma \sum_{s'} P(s'|s,a) V^{\pi}(s') \right)$$

where $R(s,a)$ is the expected immediate reward. In its matricial form the equation (system of equations) is:

$$V = \begin{bmatrix} 2 \\ 12 \\ 40 \end{bmatrix} + \frac{1}{2} \begin{bmatrix} 0.9 & 0.1 & 0 \\ 0.4 & 0.6 & 0 \\ 0.2 & 0 & 0.8 \end{bmatrix}$$

2. Considering the first two equations and multiplying both sides by 20 we have:

$$20V(s_1) = 40 + 9V(s_1) + V(s_2)$$

$$20V(s_2) = 240 + 4V(s_1) + 6V(s_1)$$

leading to $V(s_1) = \frac{16}{3}$ and $V(s_2) = \frac{56}{3}$.

Answer of exercise 9.15

1. FALSE: if we are using MC first visit since we are not counting twice the same state, TRUE: if we are using MC every visit, which computes a sample for the value function for each state visited during the episode.

2. TRUE: for MC first visit, FALSE: for MC every visit, for the same reasons of question 1.

3. FALSE: the TD update requires only to have a new visited state and, differently from MC, it does not require to have the complete episode to be applied.

4. TRUE: it is biased since it is considering the same visited state multiple times, but if the number of samples is large enough it converges to the unbiased estimates (consistent estimator), e.g., provided by MC first visit.

Answer of exercise 9.16

1. A discount factor of $\gamma = 1$ is not advisable for an infinite time horizon MDP, since it could lead to infinite reward. Moreover, using $\gamma = 1$ we are not able to provide a closed form solution to the Bellman equation.

2. We saw during classes that even using $\gamma = 0.9$ we would use the most myopic strategy $(do\ nothing, do\ nothing, do\ nothing)$. If we have an even smaller discount factor could only have an different optimal strategy.

3. $V = (I - \gamma P)^{-1}R = \left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \frac{1}{10} \begin{bmatrix} 0.9 & 0.1 & 0 \\ 0 & 0.3 & 0.7 \\ 0.2 & 0 & 0.8 \end{bmatrix} \right)^{-1} \begin{bmatrix} 2 \\ -71.5 \\ 40 \end{bmatrix}$

Answer of exercise 9.17

The first problem has few states and the transition model and the rewards are known, therefore can be solved:

- Use the closed form solution over each possible deterministic policy (brute force)
- Use the approximate solution provided by the iterative application of the Bellman Expectation operator interleaved by greedy policy improvement steps (DP - policy iteration);
- Use the approximate solution provided by the iterative application of the Bellman Optimality operator (DP - value iteration).

In the other problems we do not have information about the transition (it is stochastic or unknown) or the rewards. Therefore, we need to use some RL control technique, like SARSA or Q-learning.

Answer of exercise 9.18

1. The MDP corresponding to the provided episodes might have either a finite time horizon or an indefinite time horizon (there is no way of discriminating between the two cases by looking at a finite number of episodes). In the former case the use of a total reward maximization is a viable option, in the latter one no.

2. Since there are no repeated states the FV and EV Monte Carlo estimation coincides.

$$V(A) = \frac{6+1+1}{3} = \frac{8}{3} \qquad V(B) = \frac{3+2}{2} = \frac{5}{2} \qquad V(C) = \frac{1+3}{2} = 2$$

3. The formula for the $TD(0)$ update is the following:

$$V(S) \leftarrow V(S) + \alpha(R(S) + \gamma V(S') - V(S))$$

where S' is the next state we visit. Therefore:

$$
\begin{aligned}
V(A) &\leftarrow 0 + 0.5(3 + 0.5 \cdot 0 - 0) = \frac{3}{2}\\
V(B) &\leftarrow 0 + 0.5(2 + 0.5 \cdot 0 - 0) = 1\\
V(C) &\leftarrow 0 + 0.5(1 + 0.5 \cdot 0 - 0) = \frac{1}{2}\\
V(C) &\leftarrow \frac{1}{2} + 0.5\left(2 + 0.5 \cdot \frac{3}{2} - \frac{1}{2}\right) = \frac{13}{8}\\
V(A) &\leftarrow \frac{3}{2} + 0.5\left(1 + 0.5 \cdot 0 - \frac{3}{2}\right) = \frac{5}{4}\\
V(B) &\leftarrow 1 + 0.5\left(1 + 0.5 \cdot \frac{5}{4} - 1\right) = \frac{21}{16}\\
V(A) &\leftarrow \frac{5}{4} + 0.5\left(1 + 0.5 \cdot 0 - \frac{5}{4}\right) = \frac{9}{8}
\end{aligned}
$$

Answer of exercise 9.19

1. FALSE A common solution to solve problems with a continuous states/actions is to use a functional approximation approach on the modeling of the value/state-action value function.

2. TRUE This is the definition of the so-called off-policy learning strategies.

3. FALSE For instance in Q-learning we only model the state-action value function (model-free approach).

4. FALSE RL is a data-driven approach and therefore it requires to have only episodes/-transitions from the environment or a simulator to generate them.

Answer of exercise 9.20

1. FALSE: $TD(0)$ is equivalent to Temporal difference.

2. FALSE: the fact that it chooses a random policy with probability ε incentivizes the exploration of possibly sub-optimal action by the policy.

3. TRUE: it allows to update more than one value functions after each step.

4. TRUE: it reweights the samples to w.r.t. a sampling policy and a target one, to give the correct importance to each step.

Answer of exercise 9.21

1. FALSE: DP requires the knowledge MDP model, instead a simulator can be used to simulate experience and thus can be used with MC and TD.

2. TRUE: These three approaches allows to keep exploring while trying to find optimal policy

3. TRUE: Despite reward sparsity is a general issue for RL algorithms, an on-policy algorithm might not be able to explore enough to reach states with higher reward, while an off-policy algorithm might be more aggressive in the exploration.

4. FALSE: SARSA is based on Bellman Expectation Equation, while Value Iteration exploits Bellman Optimality Equation.

5. TRUE: TD does bootstrapping and, thus, it exploits the Markov property of the state.

6. FALSE: Even if we have a full knowledge of the MDP model, an analytical solution can be computationally infeasible. This is also true if we have a discount factor of $\gamma = 1$.

7. FALSE: In SARSA the exploration-exploitation trade-off is handled using epsilon-soft policies.

8. FALSE: In a finite MDP, for sure there is at least an optimal deterministic policy. This, however, does not prevent the existence of a stochastic optimal policy.

Answer of exercise 9.22

1. FALSE: Policy Evaluation goal is to find the correct value function corresponding to the policy we are evaluating. The optimal value function is the value function corresponding to the optimal policy instead: in order to obtain it, we have to solve the MDP (e.g., with Value-Iteration).

2. FALSE, the policy improvement theorem guarantees that, at every step, the new policy is better than the previous one.

3. FALSE, the discount factor can also be interpreted as the probability for an episode to terminate at each step, or as how much we value immediate rewards w.r.t. future ones.

4. TRUE, being SARSA an on-policy algorithm, the learned policy is used to collect samples too, hence it has to balance between obtaining high returns and exploring the state and action space. Q-learning is an off-policy algorithm, hence an explorative policy can be employed computing the greedy policy at the end of the learning process.

5. TRUE, provided that I know the reward distributions of each arm, however, in that case, it would be probably simpler to find the mean of each one and then to choose the best. (FALSE, in order to use value-iteration I have to know the model, hence the distributions, which are usually unknown to the learner in the MAB setting)

6. FALSE, for each MDP we are guaranteed that there is always at least one optimal policy, but this does not prevent having more optimal policies. Moreover, the same policy can still be optimal if we modify the reward function, sometimes this procedure can also speed up learning (reward shaping).

7. TRUE, this is guaranteed from the existence of a stationary, Markovian, and deterministic optimal policy for each (infinite horizon) MDP.

8. FALSE, when the state space is small we can use the closed-form solution. However, even if the linear system approach offers the exact solution, for very large problems it can be impractical to solve it, hence, we can resort to using the approximated solution offered by DP

9. FALSE, to model a problem as an MDP, we have to assume that the environment state is fully observable and Markovian, hence the current state of the environment should be completely determined by the current observation made by the agent. In some cases, we can "transform" the state of the environment to make it Markovian.

10. FALSE, the only optimal value function is the unique fixed point of the Bellman Optimal Operator, and all optimal policies share the same value function.

Answer of exercise 9.23

1. This snippet of code is implementing the main loop of the `SARSA` algorithm, which tackles the Reinforcement Learning control problem.

2. The `eps_greedy` function is implementing an epsilon greedy policy. Thus, it returns the action that maximizes the `Q` value in `s` with probability 1 - `eps`, or a random action with probability `eps`.

3. To ensure that the algorithm will eventually converge to the optimal policy, one should take `eps` that goes to zero, and a learning rate `alpha` that follows Robbins-Monro conditions.

4. We could change the update rule to implement the `Q-learning` algorithm, i.e.,

   ```
   Q[s, a] = Q[s, a] + alpha * (r + env.gamma * np.max(Q[ns, :])
   - Q[s, a]),
   ```

or apply the importance sampling correction to the samples we have.

Answer of exercise 9.24

1. TRUE. TD difference evaluation generally suffers a smaller variance in comparison to MC methods. If I can only access a few interactions, the variance would likely be the main issue for the value estimation.

2. FALSE. MC evaluation requires full episodes to provide an estimate of the value function, which is not a good fit for an online setting. Instead, one could use TD that employs bootstrapping to provide an online estimate of the value function.

3. FALSE. MC first-visit is unbiased but suffers from a large variance, whereas MC every-visit is slightly biased but it usually reduce the variance of the estimation.

4. FALSE. TD is a consistent policy evaluation method. This means that it will converge to the exact evaluation of the policy taking the data, which can be different from the optimal policy.