Exercises - Model-free prediction and control

Reinforcement Learning

1. Consider an environment with state space $\mathcal{S} = \{X, Y, Z\}$ where Z is a terminating state. There action space is $\mathcal{A} = \{\text{UP}, \text{DOWN}\}$. Below you can assume that both $V(s)$ and $Q(s, a)$ are initialized to 0, and use the discount factor $\gamma = 0.9$.

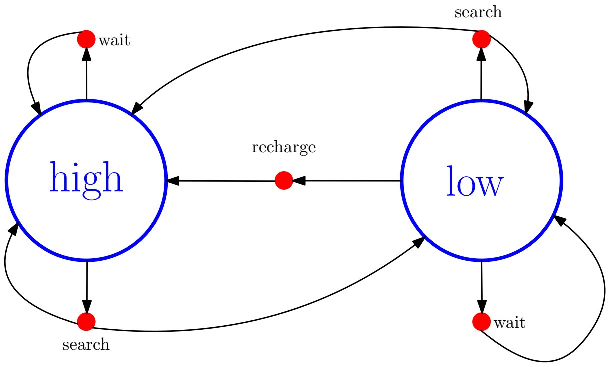
   We obtain the following episode from the environment (given in the format $S_0, A_0, R_1, S_1, A_1, R_2, \ldots$):

   $$X, \text{UP}, +100, Y, \text{DOWN}, -100, X, \text{UP}, +10, Z.$$

   (a) Assume that the episode was generated by following a fixed policy $\pi(a|s)$. What will the estimate $V(s)$ be after the episode if we use Monte-Carlo with $\alpha_t = 1/N(S_t)$? Compute the estimate both for the first-visit and every-visit version of MC.

   (b) Consider the same situation as in (a), but now we use TD(0) with step size $\alpha = 0.5$.

2. In this exercise we consider the recycling robot environment that is described in Example 3.3 in the textbook. The robot can be in two different states, either high battery level (h) or low battery level (ℓ). In h it can either search (s) or wait (w). In ℓ it can also recharge (r) itself. We consider a discount factor $\gamma \in (0, 1)$. Figure 2.1 shows a representation of the environment, but here we assume that we do not know the transition probabilities or the rewards given.



Figur 2.1: The recycling robot

	s	w	r
ℓ	0.5	1	2
h	2.5	1	

Table 2.1: The initial Q-function for the recycling robot.

t	s_t	a_t	r_{t+1}
0	h	w	+1
1	h	s	+5
2	ℓ	s	-3
3	h	w	

Table 2.2: Data received from the recycling robot while using Q-learning.

(a) SARSA was used, starting from the initial Q-function in Table 2.1, with a discount factor $\gamma = 0.9$ and step size $\alpha = 0.1$. The data received for three time steps are shown in Table 2.2. Compute the estimated $Q(s, a)$ after time-step 2.

(b) Consider the same setting as in (a), but now Q-learning was used instead. What will the estimated $Q(s, a)$ be after time-step 2?

3. Explain why Q-learning is considered to be an off-policy method.

4. Data has been collected following a policy π, and at time t the state S_t was visited. Below a number of different potential targets that can be used in training for estimating $v_\pi(s)$ are given. For each target, determine if it is an unbiased or biased estimate of $v_\pi(S_t)$, whether it can be implemented using only experience and if it would have relatively low or high variance compared to other methods.

   (a) $G_t = R_{t+1} + \gamma R_{t+2} + \cdots \gamma^{T-1} R_T$.

   (b) $R_{t+1} + \gamma v_\pi(S_{t+1})$.

   (c) $R_{t+1} + \gamma V(S_{t+1})$ where $V(S_{t+1})$ is an estimate of $v_\pi(S_{t+1})$.

   (d) (*) $R_{t+1} + \gamma R_{t+2} + \gamma^2 v_\pi(S_{t+2})$.

   (e) (*) $R_{t+1} + \gamma R_{t+2} + \gamma^2 V(S_{t+2})$ where $V(S_{t+2})$ is an estimate of $v_\pi(S_{t+2})$.

5. (*) Let us now consider targets for estimating $q_\pi(s, a)$. For the two targets below, discuss if they are unbiased or biased and whether they can be implemented in practice.

   (a) $R_{t+1} + \gamma \mathbb{E}_\pi[q_\pi(S_{t+1}, A_{t+1})|S_{t+1}] = R_{t+1} + \gamma \sum_a \pi(a|S_{t+1}) q_\pi(S_{t+1}, a)$.

   (b) $R_{t+1} + \gamma \mathbb{E}_\pi[Q(S_{t+1}, A_{t+1})|S_{t+1}] = R_{t+1} + \gamma \sum_a \pi(a|S_{t+1}) Q(S_{t+1}, a)$, where $Q(s, a)$ is an estimate of $q_\pi(s, a)$.

   Discuss if the variance of this target is higher or lower than the SARSA-target when the policy $\pi(a|s)$ is stochastic.

   (c) The target in (b) can be either off- or on-policy. Explain why this is the case.

   (d) Show that the Q-learning target is a special case of the target in (b).

Solutions

1.

(a) We have $S_0 = X$, $S_1 = Y$, $S_2 = X$ and $S_3 = Z$. The returns are

$$\begin{aligned} G_3 &= 0 \\ G_2 &= 10 + \gamma G_3 = 10 \\ G_1 &= -100 + \gamma G_2 = -91 \\ G_0 &= 100 + \gamma G_1 = 18.1. \end{aligned}$$

In first visit Monte-Carlo we just take the return from the first-visit in each state. For X the first visit is at $t = 0$ and for Y the first visit is at $t = 1$. Hence, since this is the first episode, $V(X) = G_0 = 18.1$ and $V(Y) = G_1 = -91$. However, since we also visit X at $t = 2$, the every-visit version will give $V(X) = 0.5(G_0 + G_2) = 0.5(18.1 + 10) = 14.05$ and $V(Y) = -91$.

(b) Now we instead use TD.

- For $t = 0$ we have $S_0 = X$, $R_1 = 100$, $S_1 = Y$, so the update is

  $$V(X) \leftarrow V(X) + \alpha\left[R_1 + \gamma V(Y) - V(X)\right] = 0 + 0.5\left[100 + 0 - 0\right] = 50.$$

  So now $V(X) = 50$ and $V(Y) = 0$.
- For $t = 1$ we have $S_1 = Y$, $R_2 = -100$ and $S_2 = X$. The update is

  $$V(Y) \leftarrow V(Y) + \alpha\left[R_2 + \gamma V(X) - V(Y)\right] = 0 + 0.5\left[-100 + 0.9 \times 50 - 0\right] = -27.5.$$

  Now $V(X) = 50$ and $V(Y) = -27.5$.
- For $t = 2$ we have $S_2 = X$, $R_3 = 10$ and $S_3 = Z$. The update is

  $$V(X) \leftarrow V(X) + \alpha\left[R_3 + \gamma V(Z) - V(X)\right] = 50 + 0.5\left[10 + 0 - 50\right] = 30.$$

  Hence, after the episode we have $V(X) = 30$ and $V(Y) = -27.5$.

2.

(a)

- At $t = 0$ we have $S_0 = h$, $A_0 = w$, $R_1 = 1$, $S_1 = h$, $A_1 = s$. The update is thus

  $$\begin{aligned} Q(h, w) &\leftarrow Q(h, w) + \alpha\left[1 + \gamma Q(h, s) - Q(h, w)\right] \\ &= 1 + 0.1\left[1 + 0.9 \times 2.5 - 1\right] = 1.255 \end{aligned}$$

  The new Q-table is thus

	s	w	r
ℓ	0.5	1	2
h	2.5	1.225	

- At $t = 1$ we have $S_1 = h$, $A_1 = s$, $R_2 = 5$, $S_2 = \ell$, $A_2 = s$. The update is

  $$\begin{aligned} Q(h, s) &\leftarrow Q(h, s) + \alpha\left[5 + \gamma Q(\ell, s) - Q(h, s)\right] \\ &= 2.5 + 0.1\left[5 + 0.9 \times 0.5 - 2.5\right] = 2.795 \end{aligned}$$

  The new Q-table is thus

	s	w	r
ℓ	0.5	1	2
h	2.795	1.225	

- At $t = 2$ we have $S_2 = \ell$, $A_2 = s$, $R_{t+1} = -3$ and $S_3 = h$, $A_3 = w$. The update is

  $$\begin{aligned} Q(\ell, s) &\leftarrow Q(\ell, s) + \alpha\left[-3 + \gamma Q(h, w) - Q(\ell, s)\right] \\ &= 0.5 + 0.1\left[-3 + 0.9 \times 1.225 - 0.5\right] = 0.26 \end{aligned}$$

  We thus get the Q-table

	s	w	r
ℓ	0.26	1	2
h	2.795	1.225	

(b)

- At $t = 0$ we have $S_0 = h$, $A_0 = w$ and $R_1 = +1$, and $S_1 = h$. The Q-update is

  $$\begin{aligned} Q(h,w) &\leftarrow Q(h,w) + \alpha\left[1 + \gamma \max_a Q(h,a) - Q(h,w)\right] \\ &= 1 + 0.1\,[1 + 0.9 \times 2.5 - 1] = 1.225 \end{aligned}$$

  Our new Q-table is thus

	s	w	r
ℓ	0.5	1	2
h	2.5	1.225	

- At $t = 1$ we have $S_1 = h$, $A_1 = s$, $R_{t+1} = 5$, $S_{t+1} = \ell$.
  The update is

  $$\begin{aligned} Q(h,s) &\leftarrow Q(h,s) + \alpha\left[5 + \gamma \max_a Q(\ell,a) - Q(h,s)\right] \\ &= 2.5 + 0.1\,[5 + 0.9 \times 2 - 2.5] = 2.93. \end{aligned}$$

  Our new Q-table is thus

	s	w	r
ℓ	0.5	1	2
h	2.93	1.225	

- At $t = 2$ we have $S_2 = \ell$, $A_2 = s$, $R_{t+1} = -3$ and $S_3 = h$. The update is

  $$\begin{aligned} Q(\ell,s) &\leftarrow Q(\ell,s) + \alpha\left[-3 + \gamma \max_a Q(h,a) - Q(\ell,s)\right] \\ &= 0.5 + 0.1\,[-3 + 0.9 \times 2.93 - 0.5] = 0.4137. \end{aligned}$$

  And we thus arrive at the Q-table:

	s	w	r
ℓ	0.4137	1	2
h	2.93	1.225	

3. In Q-learning the target is $R_{t+1} + \gamma Q(S_{t+1}, A')$ where $A' = \arg\max_a Q(S_{t+1}, a)$. However, this action will not in general equal A_{t+1} that comes from the policy we use to collect data. Hence Q-learning evaluates a policy that is different from the policy used to collect data.

4. Note that the target is unbiased if $v_\pi(S_t) = \mathbb{E}_\pi[\text{Target}|S_t]$.

(a) By the definition of $v_\pi(s)$ we know that $v_\pi(S_t) = \mathbb{E}_\pi[G_t|S_t]$, so the target is unbiased. It can also be computed using only experience. However, the value of all future R_{t+k} will be stochastic and depend on all future states S_{t+k} we visit. Taking a weighted sum of all these random future rewards will typically result in a relatively large variance.

(b) By the Bellman equation we know that $v_\pi(S_t) = \mathbb{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1})|S_t]$, hence the target is unbiased. However, it cannot be computed from experience alone since we need to know $v_\pi(s)$. It can also be expected to have relatively low variance since the only stochastic part is the reward R_{t+1} and what the next state S_{t+1} is.

(c) In this case $V(S_{t+1})$ can be any estimate, and unless $V(S_{t+1}) = v_\pi(S_{t+1})$ there will be a bias in this target. That is, since we replace everything after R_{t+1} in the G_t with an estimated value of the future return from S_{t+1} we will have a bias in the target. However, by replacing $v_\pi(S_{t+1})$ with the estimate $V(S_{t+1})$ we can compute without knowing the true value function.

(d) Note that

$$\begin{aligned} &\mathbb{E}_\pi[R_{t+1} + \gamma R_{t+2} + \gamma^2 v_\pi(S_{t+2})|S_t] = E_\pi[R_{t+1} + \gamma(R_{t+2} + \gamma v_\pi(S_{t+2}))|S_t] = \\ &\qquad \mathbb{E}_\pi[R_{t+1} + \gamma \mathbb{E}_\pi[R_{t+2} + \gamma v_\pi(S_{t+2})|S_{t+1}]|S_t] = \mathbb{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1})|S_t] = v_\pi(S_t) \end{aligned}$$

hence the target is unbiased. Here we have used the law of total expectation to see that

$$\mathbb{E}_\pi\left[\mathbb{E}_\pi[R_{t+2} + \gamma v_\pi(S_{t+2}) | S_{t+1}] \,|S_t\right] = \mathbb{E}_\pi[R_{t+2} + \gamma v_\pi(S_{t+2}) | S_t]$$

However, to compute it we need the true $v_\pi(S_{t+1})$. For variance, the target uses $R_{t+1}, R_{t+2}, S_{t+1}$ and S_{t+2}. Hence it can be expected that the variance is higher than for the target in (b), but lower than for the MC-target in part (a).

(e) As in (c) this will be biased unless $V(S_{t+2}) = v_\pi(S_{t+2})$, but now we can compute it. It can be expected that the bias is lower than in (c), since we take two steps with observed rewards before replacing the rest of the return with an estimated value.

5.

(a) We get

$$\mathbb{E}_\pi[R_{t+1} + \gamma\mathbb{E}_\pi[q_\pi(S_{t+1}, A_{t+1})|S_{t+1}]|S_t, A_t] = \mathbb{E}_\pi[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1})|S_t, A_t] = q_\pi(S_t, A_t),$$

so it is unbiased. We cannot use it in practice, since it requires that we know $q_\pi(S_{t+1}, A_{t+1})$.

(b) In this case the target is biased since we use an estimate $Q(S_{t+1}, A_{t+1})$. But we can compute it in practice, even though we take the expectation over the next action, since we know the policy $\pi(a|s)$.

The target will typically have lower variance than the SARSA-target $R_{t+1} + \gamma Q(S_{t+1}, A_{t+1})$ if $\pi(a|s)$ is stochastic. To see this, note that A_{t+1} in SARSA is a random variable, and will thus add variance to the SARSA-target. The target suggested here instead uses the average over all possible next actions according to the policy $\pi(a|s)$.

Note: The method we get with this target is called Expected SARSA. We have not discussed it during the lectures, but you can read more about it in Section 6.6 of the textbook.

(c) The target in (b) can be used to estimate q_π. Note that this target works either if the experience is collected using the same policy π or a different behavior policy μ. If the same policy π is used for data collection, then it is an on-policy method, if a different policy is used for data collection, then it is off-policy.

(d) Q-learning uses the target $R_{t+1} + \gamma \max_a Q(S_{t+1}, a)$. If we let $\pi(a|s)$ be greedy with respect to Q, i.e.,

$$\pi(a|s) = \begin{cases} 1 & \text{if } a = \arg\max_{a'} Q(s, a') \\ 0 & \text{otherwise} \end{cases}$$

then

$$\sum_a \pi(a|S_{t+1})Q(S_{t+1}, a) = Q(S_{t+1}, \arg\max_a Q(S_{t+1}, a)) = \max_a Q(S_{t+1}, a),$$

so the target in (b) becomes the same as the target in Q-learning. Hence, if we use a behavioral policy $\mu = \varepsilon\text{-greedy}(Q)$ an let $\pi = \text{greedy}(Q)$ with the target in (b) then we get Q-learning.