hw 2 Solution

chenyu wu

Oct 2023

1 Sutton and Barto exercises

3.7 (8/6 pts)

What is going wrong with robot maze learning?

This is an open question. Here I only provide one aspect to answer this question.

The communication is not effective. There is no restriction to the time of running out of the maze. The agent does not get a penalty as they use a very long time to get out of the maze, which will make the agent indifferent to whether to get out of the maze faster or slower. Therefore, trainers should apply some penalty according to the time the robot is used.

3.8 (8/6 pts)

Suppose $\gamma = 0.5$ and the following sequence of rewards is received $R_1 = -1, R_2 = 2, R_3 = 6, R_4 = 3,$ and $R_5 = 2$, with $T = 5$. What are $G_0, G_1, \cdots, G_5$?

According to the hint we have:

$$
\begin{aligned}
G_5 &= 0 \\
G_4 &= R_5 = 2 \\
G_3 &= R_4 + \gamma G_4 = 3 + 1 = 4 \\
G_2 &= R_3 + \gamma G_3 = 6 + 2 = 8 \\
G_1 &= R_2 + \gamma G_2 = 2 + 4 = 6 \\
G_0 &= R_1 + \gamma G_1 = -1 + 3 = 2
\end{aligned}
$$

3.9 (8/6 pts)

Suppose $\gamma = 0.9$ and the reward sequence is $R_1 = 2$ followed by an infinite sequence of 7s. What are G_1 and G_0?

We know that

$$\begin{aligned} G_1 &= 7 + \gamma 7 + \gamma^2 7 + \cdots \\ &= 7(1 + \gamma + \gamma^2 + \cdots) \\ &= 7\frac{1}{1-\gamma} = 70 \\ G_0 &= 2 + \gamma G_1 \\ &= 2 + 63 = 65 \end{aligned}$$

3.12 (8/6 pts)

Give an equation for v_π in terms of q_π and π

$$v_\pi(s) = \sum_{a \in \mathcal{A}(s)} \pi(a|s) q_\pi(s, a)$$

3.13 (8/6 pts)

Give an equation for q_π in terms of v_π and the four argument p.

$$q_\pi(s, a) = \sum_{s' \in \mathcal{S}, r \in \mathcal{R}} (r + \gamma v_\pi(s')) p(s', r|s, a)$$

3.15 (8/6 pts)

Only the intervals between the rewards are important. Suppose now we have a new reward setting $R_i = R_i + c$. Then according to equation (3.8) we know that the new discounted return is

$$\tilde{G}_t = \sum_{k=0}^{\infty} \gamma^k \tilde{R}_{t+k+1} = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} + \sum_{k=0}^{\infty} \gamma^k c = G_t + \sum_{k=0}^{\infty} \gamma^k c.$$

Then we can see the new value function should solve

$$\begin{aligned} \tilde{v}_\pi(s) &\doteq \mathbb{E}_\pi[\tilde{G}_t | S_t = s] \\ &= \mathbb{E}_\pi[G_t | S_t = s] + \sum_{k=0}^{\infty} \gamma^k c \\ &= v_\pi(s) + \frac{c}{1-\gamma} \end{aligned}$$

We conclude $v_c = \frac{c}{1-\gamma}$.

3.17 (8/6 pts)

What is the Bellman equation for action values, that is, for q_π? It must give the action value $q_\pi(s, a)$ in terms of the action values, $q_\pi(s', a')$, of possible successors to the state-action pair (s,a).

We can derive this from the results of 3.12 and 3.13

$$q_\pi(s,a) = \sum_{s',r} p(s',r|s,a) \left[r + \gamma \sum_{a'} \pi(a'|s') q_\pi(s',a') \right]$$

3.25 (8/6 pts)

Give an equation for v_* in terms of q_*

$$\begin{aligned} v_*(s) &= \sum_a \pi(a|s) q_*(s,a) \\ &= \max_{a \in \mathcal{A}(s)} q_*(s,a) \end{aligned}$$

3.26 (8/6 pts)

Give an equation for q_* in terms of v_* and the four argument p

$$q_*(s,a) = \sum_{s',r} p(s',r|s,a)[r + \gamma v_*(s')]$$

3.27 (8/6 pts)

Give an equation for π_* in terms of q_*

$$\pi_*(a|s) = \begin{cases} 1 & a = \arg\max_a q_*(s,a) \\ 0 & \text{o.w.} \end{cases}$$

4.5 (10 pts)

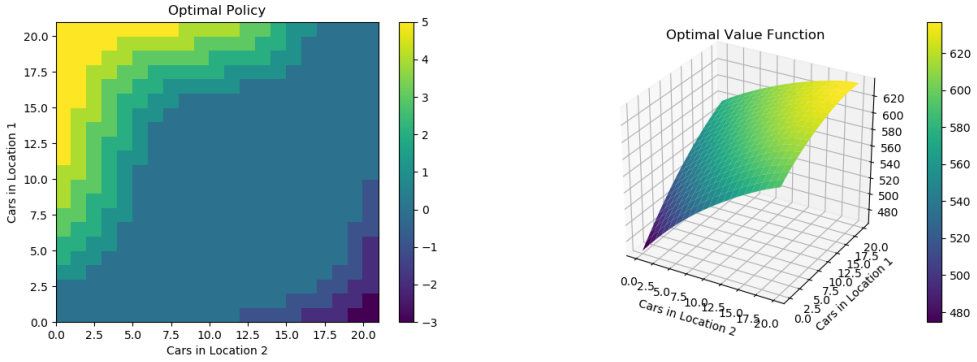
How would policy iteration be defined for action values?

1. Initialization
   $q(s,a) \in \mathbb{R}$ and $\pi(s) \in \mathcal{A}(s)$ arbitrarily for all $s \in \mathcal{S}$.
2. Policy Evaluation
   Loop:
   $\Delta \leftarrow 0$
   Loop for each $(s,a) \in \mathcal{S} \times \mathcal{A}(s)$:
   $q \leftarrow q(s,a)$
   $q(s,a) \leftarrow \sum_{s',r} p(s',r|s,\pi(s))[r + \gamma \sum_{a'} q(s',a')]$
   $\Delta \leftarrow \max(\Delta, |q - q(s,a)|)$
   until $\Delta < \theta$ (a small positive number determining the accuracy of estimation)

3. Policy Improvement
   $policy - stable \leftarrow true$
   For each $(s, a) \in \mathcal{S} \times \mathcal{A}$:
   $old - action \leftarrow \pi(s)$
   $\pi(s) \leftarrow \arg\max_a q(s, a)$
   If $old - action \neq \pi(s)$, then $policy - stable \leftarrow false$
   If $policy - stable$, then stop and return $q \approx q_*$, and $\pi \approx \pi_*$; else go to 2

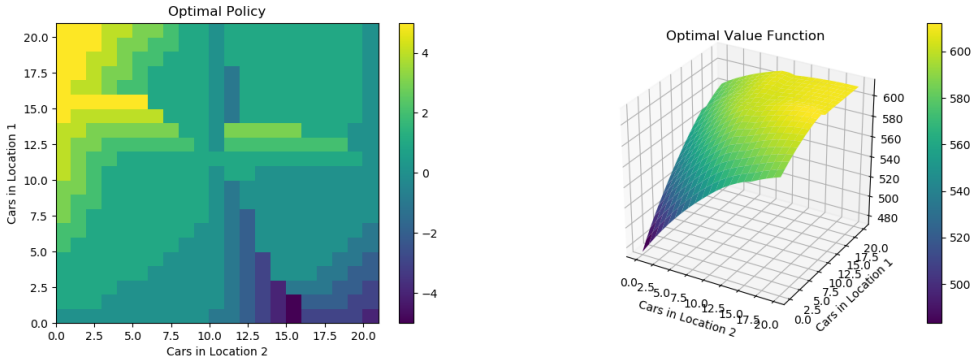
4.7 (10 pts)

See Figure 1 and Figure 2



(a) Policy (b) Value

Figure 1: Example 4.2 policy and value function



(a) Policy (b) Value

Figure 2: Example 4.2 policy and value function

2 Problem 2 (10 pts)

(a):
To see this, we can simply rearrange the summation. We first have

$$\begin{aligned} v_\pi(s,\gamma) &= \sum_{k=0}^{\infty} \gamma^k \mathbb{E}_\pi \left[R_{k+1} | S_0 = s \right] \\ &\doteq \sum_{k=0}^{\infty} \gamma^k C_{k+1} \end{aligned}$$

Then we check the $\tilde{v}_\pi(s,\gamma)$:

$$\begin{aligned} \tilde{v}_\pi(s,\gamma) &= \sum_{t=1}^{\infty} \gamma^{t-1}(1-\gamma) \mathbb{E}_\pi \left[\sum_{k=1}^{T} R_k | S_0 = s, T = t \right] \\ &= \sum_{t=1}^{\infty} \gamma^{t-1}(1-\gamma) \sum_{k=1}^{T} \mathbb{E}_\pi \left[R_k | S_0 = s \right] \\ &= \sum_{t=1}^{\infty} \gamma^{t-1}(1-\gamma) \sum_{k=1}^{T} C_k \\ &= \left[(1-\gamma)(1+\gamma+\gamma^2+\cdots) \right] C_1 + \left[(1-\gamma)(\gamma+\gamma^2+\cdots) \right] C_2 + \cdots \\ &= \sum_{k=1}^{\infty} \gamma^{k-1} C_k \\ &= v_\pi(s,\gamma) \end{aligned}$$

(b):
From a similar approach:

$$\begin{aligned} \tilde{v}_\pi(s,\gamma) &= \sum_{t=1}^{\infty} (t-1)\gamma^{t-2}(1-\gamma)^2 \sum_{k=1}^{t} \mathbb{E}_\pi [R_k | S_0 = s] \\ &= \sum_{t=1}^{\infty} (t-1)\gamma^{t-2}(1-\gamma)^2 \sum_{k=1}^{t} C_k \\ &= \sum_{k=1}^{\infty} C_k \sum_{t=k}^{\infty} (t-1)\gamma^{t-2}(1-\gamma)^2 \\ &= \sum_{k=1}^{\infty} C_k \sum_{t=k}^{\infty} (t-1)\gamma^{t-2}(1-\gamma)^2 \\ &= \sum_{k=1}^{\infty} C_k [(k-1)\gamma^{k-2}(1-\gamma) + \gamma^{k-1}] \\ &= v_\pi(s,\gamma) + (1-\gamma)\frac{d}{d\gamma} v_\pi(s,\gamma) \end{aligned}$$

3 Problem 3 (10 pts)

3.1 (a) Write down the DPE:

The DPE of this problem is defined as

$$\begin{aligned} V(X_n) &\doteq \max\left\{(1-\exp(X_N))^+, \gamma\mathbb{E}\left[V(X_{N+1})|X_N\right]\right\} \\ &\doteq \max\left\{(1-\exp(X_N))^+, \frac{\gamma}{2}V(X_N+1)+\frac{\gamma}{2}V(X_N-1)\right\} \end{aligned}$$

3.2 (b) Should z be positive or negative?

Since the value $(1-\exp(x))^+$ is decreasing with x and equal to 0 is $x \geq 0$, we can see that it will not be optimal when $x \geq 0$. Thus, we conclude that z should be negative.

3.3 (c) The structural form of $v(x)$:

Since we stop when $X_n \leq z$, we can see that the solution $v(x) = (1-e^x)$ for $x \leq z < 0$ and $v(x) = \frac{\gamma}{2}V(x+1) + \frac{\gamma}{2}v(x-1)$ for $x > z$. Then the general solution for the latter difference equation is

$$v(x) = Ae^{\alpha x} + Be^{-\alpha x}, \forall x \geq z,$$

for some constant A and B, where

$$\alpha \doteq \log\frac{1-\sqrt{1-\gamma^2}}{\gamma}.$$

From the definition of α we can see that $\alpha < 0$. Since we want solution $v(x)$ be bounded, we can see that $B = 0$, and thus

$$v(x) = Ae^{\alpha x}, \forall x \geq z.$$

To solve this, we can apply the boundary condition $v(z) = (1-e^z)$:

$$1 - e^z = Ae^{\alpha z} \Rightarrow A = e^{-\alpha z}(1-e^z).$$

Therefore, we conclude the solution

$$v(x) = \begin{cases} 1-e^x & x \leq z \\ e^{\alpha(x-z)}(1-e^z) & x > z. \end{cases}$$

3.4 (d) Show that if $\gamma = 1$ then $v(x) = 1$ for all x:

Intuitive solution: Since $\gamma = 1$, there is no cost to do exploration. Therefore, one can always keep exploring as long as the function value does not reach its upper bound, i.e. $\lim_{x\to-\infty}(1-e^x) = 1$. However, we know that the process X_n is a simple symmetric random walk so there is no finite time for $X_n = -\infty$. Therefore, this implies no optimal stopping time.

To see this rigorously:

It is clear that $v(x) \leq 1$ for all x. But for an arbitrary $j \in \mathbb{N}$, let $\sigma_j \doteq \inf\{n \geq 0 : X_n \leq j\}$, then σ_j is finite, and we have, for $x \geq -j$,

$$v(x) = \mathbb{E}[(1 - \exp(X_{\sigma_j}))^+] = (1 - \exp(-j)) \to 1$$

as $j \to \infty$. This implies that $v(x) \geq 1$, and hence $v(x) = 1$. The value function clearly satisfies the DPE

$$v(x) = \max\{(1 - e^x)^+, [v(x+1) + v(x-1)]/2\}.$$

But since the stopping time

$$\inf\{n \geq 0 : v(X_n) = (1 - \exp(X_n))^+\} = \infty,$$

there is no optimal stopping time.