Reinforcement Learning: HW 1

Siddhant Chaudhary

Contents

0.1 Problem 1.

Problem Statement. Consider the MDP $M = \langle S, A, P, R \rangle$, where $S = \{s_1, s_2, s_3\}$ and $A = \{a_1, a_2\}$, with the transition probabilities given by the

	s_1	s_2	s_3
(s_1, a_1)	1	0	0
(s_1, a_2)	0	0.5	0.5
(s_2, a_1)	0	1	0
(s_2, a_2)	0.3	0	0.7
(s_3, a_1)	0	0	1
(s_3, a_2)	0.1	0.9	0

and the rewards are given by

	a_1	a_2
s_1	1	2
s_2	0	3
s_3	1	4

(a) Write down the state space S, action space A, state transition matrix P and the reward vector R of M.
(b) How many policies does M have? Briefly explain your answer.
(c) Let π_2 denote the policy that takes action a_1 in states s_1 and s_2 and action a_2 in state s_3. Define π_2 using symbols. Draw π_2 as an MDP.

Solution. First, consider part (a). The state space is $S = \{s_1, s_2, s_3\}$ and the action space is $A = \{a_1, a_2\}$. We've already written down the state transition matrix P and the reward vector R.

Now, we come to part (b). A policy is just a map $\pi : S \to A$; so, to count the number of policies, we need to count the number of such maps. Clearly, $|S| = 3$ and $|A| = 2$, so there are $2^3 = 8$ such functions, and hence there are 8 policies for the MDP M.

Finally, we come to (c). The policy π_2 is defined using symbols as follows.

$$\pi(s_1) = a_1$$
$$\pi(s_2) = a_1$$
$$\pi(s_3) = a_2$$

The corresponding MDP for this policy can be easily drawn.

0.2 Problem 2.

Problem Statement. We have seen four definitions of V^π in class. Write down all four definitions formally using the right notation.

Solution. The first definition we saw of V^π for a policy π is the expected discounted reward over an infinite time horizon; formally, we defined for a state $s \in S$

$$V^\pi(s) := \mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, \pi(s_t)) \;\middle|\; s_0 = s\right]$$

We then looked at the matrix equation for V^π, which is the following.

$$V^\pi = R^\pi + \gamma P^\pi V^\pi$$

Here, R^π is the vector whose coordinates are the rewards $R(s, \pi(s))$ for $s \in S$ and P^π is the $|S| \times |S|$ matrix whose sth row is the probability vector associated to the state-action pair $(s, \pi(s))$. From the above, we can see that

$$(I - \gamma P^\pi)V^\pi = R^\pi$$

and it turns out that the matrix $I - \gamma P^\pi$ is invertible, which gives us

$$V^\pi = (I - \gamma P^\pi)^{-1} R^\pi$$

We can consider this to be another definition of V^π.

Next, we expanded the RHS of the equation $V^\pi = R^\pi + \gamma P^\pi V^\pi$ to obtain

$$\begin{aligned} V^\pi &= R^\pi + \gamma P^\pi R^\pi + \gamma^2 (P^\pi)^2 R^\pi + \cdots \\ &= \left(\sum_{t=0}^{\infty} \gamma^t (P^\pi)^t\right) R^\pi \end{aligned}$$

The matrix $\sum_{t=0}^{\infty} \gamma^t (P^\pi)^t$ is denoted D_π, and is called the *visitation frequency matrix*. So, we get

$$V^\pi = D^\pi R^\pi$$

This in turn can be treated as another definition of V^π.

Finally, let B_π be the Bellman backup operator for the policy π. We have seen in class that V^π is a fixed point of B_π, and that B_π enjoys the *monotonicity* and *contraction* properties; using these, we showed for any vector $v \in \mathbf{R}^{|S|}$, we have

$$V^\pi = \lim_{k \to \infty} B_\pi^k(v)$$

We can also treat this as another definition of V^π.

0.3 Problem 3.

Problem Statement. Compute the value vector of policy π_2 in **Problem 1** (c) above using the recursive formula

$$V^{\pi}(s) = R(s, \pi(s)) + \gamma \sum_{s'} \mathbf{P}_{s,s'}(\pi(s)) V^{\pi}(s')$$

for all $s \in S$ with $\gamma = 0.9$.

Solution. Let $V^{\pi_2}(s_1) = x$, $V^{\pi_2}(s_2) = y$ and $V^{\pi_2}(s_3) = z$. We want to compute the vector (x, y, z). Using the above equation for V^{π_1}, we get the following three equations.

$$\begin{aligned}
x &= R(s_1, a_1) + \gamma P_{s_1,s_1}(a_1)x + \gamma P_{s_1,s_2}(a_1)y + \gamma P_{s_1,s_3}(a_1)z \\
y &= R(s_2, a_1) + \gamma P_{s_2,s_1}(a_1)x + \gamma P_{s_2,s_2}(a_1)y + \gamma P_{s_2,s_3}(a_1)z \\
z &= R(s_3, a_2) + \gamma P_{s_3,s_1}(a_2)x + \gamma P_{s_3,s_2}(a_2)y + \gamma P_{s_3,s_3}(a_2)z
\end{aligned}$$

Plugging in the values of the probabilities and the rewards, we get the following equations.

$$\begin{aligned}
x &= 1 + \gamma x \\
y &= 0 + \gamma y \\
z &= 4 + \gamma 0.1x + \gamma 0.9y
\end{aligned}$$

Solving the above, we get

$$\begin{aligned}
x &= \frac{1}{1-\gamma} = 10 \\
y &= 0 \\
z &= 4 + \frac{\gamma}{10} \cdot \frac{1}{1-\gamma} = 4.9
\end{aligned}$$

So, we get that $V^{\pi_2} = (10, 0, 4.9)$.

0.4 Problem 4.

Problem Statement. We have seen that $Q^V(s, a)$ is the one-step backup of action a at state s with respect to vector V, i.e

$$Q^V(s, a) = R(s, a) + \gamma \sum_{s'} \mathbf{P}_{s,s'}(a) V(s')$$

(a) Compute the Q-values of all state-action pairs of the MDP M with respect to V^{π_2} computed in **Problem 3** above.

(b) Find a policy π' that is better than policy π_2 for MDP M just by using the Q-values computed above. *Note*: Policy π' is better than policy π if the value vector of π' is greater than the value vector of π.

Solution. For (a), we need to compute the Q-values of all state-action pairs. This is straightforward. Recall that we computed $V^{\pi_2} = (10, 0, 4.9)$.

$$\begin{aligned}
Q^{V_2^{\pi}}(s_1, a_1) &= R(s_1, a_1) + \gamma \mathbf{P}_{s_1,s_1}(a_1)V^{\pi_2}(s_1) + \gamma \mathbf{P}_{s_1,s_2}(a_1)V^{\pi_2}(s_2) + \gamma \mathbf{P}_{s_1,s_3}(a_1)V^{\pi_2}(s_3) \\
Q^{V_2^{\pi}}(s_1, a_2) &= R(s_1, a_2) + \gamma \mathbf{P}_{s_1,s_1}(a_2)V^{\pi_2}(s_1) + \gamma \mathbf{P}_{s_1,s_2}(a_2)V^{\pi_2}(s_2) + \gamma \mathbf{P}_{s_1,s_3}(a_2)V^{\pi_2}(s_3) \\
Q^{V_2^{\pi}}(s_2, a_1) &= R(s_2, a_1) + \gamma \mathbf{P}_{s_2,s_1}(a_1)V^{\pi_2}(s_1) + \gamma \mathbf{P}_{s_2,s_2}(a_1)V^{\pi_2}(s_2) + \gamma \mathbf{P}_{s_2,s_3}(a_1)V^{\pi_2}(s_3) \\
Q^{V_2^{\pi}}(s_2, a_2) &= R(s_2, a_2) + \gamma \mathbf{P}_{s_2,s_1}(a_2)V^{\pi_2}(s_1) + \gamma \mathbf{P}_{s_2,s_2}(a_2)V^{\pi_2}(s_2) + \gamma \mathbf{P}_{s_2,s_3}(a_2)V^{\pi_2}(s_3) \\
Q^{V_2^{\pi}}(s_3, a_1) &= R(s_3, a_1) + \gamma \mathbf{P}_{s_3,s_1}(a_1)V^{\pi_2}(s_1) + \gamma \mathbf{P}_{s_3,s_2}(a_1)V^{\pi_2}(s_2) + \gamma \mathbf{P}_{s_3,s_3}(a_1)V^{\pi_2}(s_3) \\
Q^{V_2^{\pi}}(s_3, a_2) &= R(s_3, a_2) + \gamma \mathbf{P}_{s_3,s_1}(a_2)V^{\pi_2}(s_1) + \gamma \mathbf{P}_{s_3,s_2}(a_2)V^{\pi_2}(s_2) + \gamma \mathbf{P}_{s_3,s_3}(a_2)V^{\pi_2}(s_3)
\end{aligned}$$

Plugging in all the values, we get the following.

$$\begin{aligned}
Q^{V^{\pi_2}}(s_1, a_1) &= 1 + 0.9 \cdot 1 \cdot 10 && = 10 \\
Q^{V^{\pi_2}}(s_1, a_2) &= 2 + 0.9 \cdot 0.5 \cdot 0 + 0.9 \cdot 0.5 \cdot 4.9 && = 4.205 \\
Q^{V^{\pi_2}}(s_2, a_1) &= 0 + 0.9 \cdot 0 && = 0 \\
Q^{V^{\pi_2}}(s_2, a_2) &= 3 + 0.9 \cdot 0.3 \cdot 10 + 0.9 \cdot 0.7 \cdot 4.9 && = 8.787 \\
Q^{V^{\pi_2}}(s_3, a_1) &= 1 + 0.9 \cdot 4.9 && = 5.41 \\
Q^{V^{\pi_2}}(s_3, a_2) &= 4 + 0.9 \cdot 0.1 \cdot 10 + 0.9 \cdot 0.9 \cdot 0 && = 4.9
\end{aligned}$$

So, we've computed all the required Q-values.

Now we come to part (b). Using the Q-values we've just computed, we will try to find a policy π' that is better than policy π_2. Note that for state s_2, the Q-value $Q^{V_2^\pi}(s_2, a_2)$ is greater than the Q-value $Q^{V_2^\pi}(s_2, a_1)$ (which is 0). So, consider the policy π' which is equal to policy π_2 on states s_1 and s_3, but for state s_2 it takes action a_2; in other words, consider the policy π' defined by

$$\begin{aligned}
\pi'(s_1) &= a_1 \\
\pi'(s_2) &= a_2 \\
\pi'(s_3) &= a_2
\end{aligned}$$

As we have seen in class, this policy is better that π_2, because it attains a higher Q-value at state s_2; more precisely, we have $B_{\pi'}(V^{\pi_2}) \geq V^{\pi_2}$.

0.5 Problem 5.

Problem Statement. We have seen the Bellman backup operator B_π and its properties in class. Write down the definition of the operator and its two properties formally using the right notation.

Solution. Let π be any policy for an MDP. The Bellman backup operator B_π is defined as a mapping $\mathbf{R}^{|S|} \to \mathbf{R}^{|S|}$ using the formula

$$B_\pi(v) = R^\pi + \gamma P^\pi v$$

for $v \in \mathbf{R}^{|S|}$. Here, R^π is the vector whose sth coordinate is the reward $R(s, \pi(s))$; $\gamma \in (0, 1)$ and P^π is the $|S| \times |S|$ matrix whose sth row is the probability vector associated to the state-action pair $(s, \pi(s))$. In class, we have shown that B_π satisfies two important properties: *monotonicity* and *contraction*. We now state what these mean.

(1) (*Monotonicity*) Let $\leq$ be the partial order on $\mathbf{R}^{|S|}$ which compares two vectors component-wise. Then, if $u, v \in \mathbf{R}^{|S|}$ such that $u \leq v$, then it is true that

$$B_\pi(u) \leq B_\pi(v)$$

(2) (*Contraction*) Consider the $||\cdot||_\infty$ norm on $\mathbf{R}^{|S|}$. Then, for any $u, v \in \mathbf{R}^{|S|}$, it holds that

$$||B_\pi(u) - B_\pi(v)||_\infty \leq \gamma \, ||u - v||_\infty$$

0.6 Problem 6.

Problem Statement. We know that the visitation frequency matrix D_π of policy π is

$$D_\pi = I + \gamma P^\pi + \gamma^2 (P^\pi)^2 + \gamma^3 (P^\pi)^3 + \cdots \tag{1}$$

(a) Write down the state-transition matrix P^{π_2} of policy π_2 as in **Problem 1** (c) above. Label the rows and columns of P^{π_2}. Compute D_{π_2} using equation (1) with $\gamma = 0.9$.
(b) What is the sum of the entries along the three rows of D_{π_2}? What is the sum of the entries along any row of D_π in general?
(c) Compute V^{π_2} using D_{π_2} with $\gamma = 0.9$.

Solution. First consider part (a). The state-transition matrix P^{π_2} of policy π_2 is given by the following.

$P^{\pi_2} =$

	s_1	s_2	s_3
$(s_1, \pi_2(s_1) = a_1)$	1	0	0
$(s_2, \pi_2(s_2) = a_1)$	0	1	0
$(s_3, \pi_2(s_3) = a_2)$	0.1	0.9	0

Next, we will compute D_{π_2} using equation (1). To use this equation, we will have to compute the entries of the matrix $(P^{\pi_2})^k$ for any $k \geq 1$. But, it is easy to see that in our case

$$(P^{\pi_2})^2 = P^{\pi_2}$$

by just computing the square of the matrix. This implies that $(P^{\pi_2})^k = P^{\pi_2}$ for all $k \geq 1$. So, we get that

$$\begin{aligned} D_{\pi_2} &= I + \sum_{t=1}^{\infty} \gamma^t P^{\pi_2} \\ &= I + \frac{\gamma}{1-\gamma} P^{\pi_2} \\ &= \begin{bmatrix} 10 & 0 & 0 \\ 0 & 10 & 0 \\ 0.9 & 8.1 & 1 \end{bmatrix} \end{aligned}$$

Let us now focus on part (b), i.e we will compute the sum of the entries along any row of D_π. Let $s \in S$, and we will compute the sum of all entries in the sth row of D_π. We want to compute the sum of all entries in the sth row of the matrix

$$\sum_{t=0}^{\infty} \gamma^t (P^\pi)^t$$

Let $t \geq 0$ be any integer, and consider the matrix $\gamma^t (P^\pi)^t$. Note that the sth row of the matrix $(P^\pi)^t$ is a probability vector; the s'th entry in this row is just the probability of reaching state s' under the policy π in t steps starting from the state s. So, we see that the sum of all entries in the sth row of $(P^\pi)^t$ is 1, and hence, the sum of all entries in the sth row of $\gamma^t (P^\pi)^t$ is γ^t. So, we see that the sum of all entries in the sth row of D_π is just the sum

$$\sum_{t=0}^{\infty} \gamma^t = \frac{1}{1-\gamma}$$

Clearly, this agrees with the sum of the rows of the matrix D^{π_2} we've computed above.

We finally come to part (c). We know that

$$\begin{aligned} V^{\pi_2} &= D_{\pi_2} R^{\pi_2} \\ &= \begin{bmatrix} 10 & 0 & 0 \\ 0 & 10 & 0 \\ 0.9 & 8.1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 4 \end{bmatrix} \\ &= \begin{bmatrix} 10 \\ 0 \\ 4.9 \end{bmatrix} \end{aligned}$$

This matches with the value of V^{π_2} we computed before.

0.7 Problem 7.

Problem Statement. Suppose we added a constant vector $C = \begin{bmatrix} c & c & \cdots & c \end{bmatrix}^T$ with mn components to the reward vector R of an MDP with n states and m actions. How does this affect the value vectors of policies? Does it change the optimal policy?

Solution. For any policy π, let V^{π}_{new} denote the new value vector (i.e the value vector for policy π under the modified MDP). Also, let $\mathbf{1}_{|S|}$ be the $|S| \times 1$ column vector, all of whose entries are 1. We will show the following two claims.

(1) For all policies π

$$V^{\pi}_{\text{new}} = V^{\pi} + \frac{c}{1-\gamma}\mathbf{1}_{|S|}$$

(2) If π, π' are policies such that $V^{\pi} \leq V^{\pi'}$, then

$$V^{\pi}_{\text{new}} \leq V^{\pi'}_{\text{new}}$$

First let us show (1). Let π be any policy. Let R^{π}_{new} be the new reward vector, i.e $R^{\pi}_{\text{new}} = R^{\pi} + c\mathbf{1}_{|S|}$. We know that V^{π}_{new} satisfies the equation

$$V^{\pi}_{\text{new}} = D^{\pi} R^{\pi}_{\text{new}}$$

where D^{π} is the visitation frequency matrix. Now, observe that

$$\begin{aligned} D^{\pi} R^{\pi}_{\text{new}} &= D^{\pi}(R^{\pi} + c\mathbf{1}_{|S|}) \\ &= D^{\pi} R^{\pi} + cD^{\pi}\mathbf{1}_{|S|} \\ &= V^{\pi} + cD^{\pi}\mathbf{1}_{|S|} \end{aligned}$$

From **Problem 6** (b), we know that the sum of all entries along any row of D^{π} is $\frac{1}{1-\gamma}$; this implies that $cD^{\pi}\mathbf{1}_{|S|} = \frac{c}{1-\gamma}\mathbf{1}_{|S|}$, and hence we obtain that

$$V^{\pi}_{\text{new}} = V^{\pi} + \frac{c}{1-\gamma}\mathbf{1}_{|S|}$$

In other words, each coordinate of V^{π} will be increased by $\frac{c}{1-\gamma}$ in the modified MDP.

Next, let's focus on (2). So, let π, π' be policies such that $V^{\pi} \leq V^{\pi'}$. Combining this with (1), we clearly see that

$$\begin{aligned} V^{\pi}_{\text{new}} &= V^{\pi} + \frac{c}{1-\gamma}\mathbf{1}_{|S|} \\ &\leq V^{\pi'} + \frac{c}{1-\gamma}\mathbf{1}_{|S|} \\ &= V^{\pi'}_{\text{new}} \end{aligned}$$

and this proves (2). Clearly, (2) implies that the set of optimal policies doesn't change in the new MDP.