Reinforcement Learning: HW 2

Siddhant Chaudhary

Contents

0.1 Problem 1.

Problem Statement. Let $M = \langle S, A, P, R\rangle$ be the same MDP as in **Problem 1** of the first homework. Let π be the stochastic policy that takes action a_1 with probabiltiy 0.5 and action a_2 with probability 0.5 in all the three states. In our notation, we have

$$\begin{aligned} \pi(a_1|s_1) &= 0.5, \quad \pi(a_2|s_1) = 0.5 \\ \pi(a_1|s_2) &= 0.5, \quad \pi(a_2|s_2) = 0.5 \\ \pi(a_1|s_3) &= 0.5, \quad \pi(a_2|s_3) = 0.5 \end{aligned}$$

Give the state-transition matrix and the reward vector for policy π.

Solution. Just for clarity, we recall that the transition-probability matrix is given by

	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

	a_1	a_2
s_1	1	2
s_2	0	3
s_3	1	4

First, we compute the state-transition matrix P^π for the stochastic policy π. Let s be any state. As we have seen in class, the sth row of the matrix P^π will be the weighted sum of the rows of the transition matrix indexed by the state-action pairs $(s, a)_{a\in A}$, where the corresponding weights will be $\pi(a|s)_{a\in A}$; in other words, if M_a denotes the ath row of some matrix M, then we have

$$P_s^\pi = \sum_{a\in A} \pi(a|s) P_{(s,a)}$$

Using this, we can immediately compute the matrix P^π to get the following.

$$\begin{aligned} P^\pi_{s_1} &= 0.5\begin{bmatrix}1 & 0 & 0\end{bmatrix} + 0.5\begin{bmatrix}0 & 0.5 & 0.5\end{bmatrix} \\ P^\pi_{s_2} &= 0.5\begin{bmatrix}0 & 1 & 0\end{bmatrix} + 0.5\begin{bmatrix}0.3 & 0 & 0.7\end{bmatrix} \\ P^\pi_{s_3} &= 0.5\begin{bmatrix}0 & 0 & 1\end{bmatrix} + 0.5\begin{bmatrix}0.1 & 0.9 & 0\end{bmatrix} \end{aligned}$$

and hence we get

$$P^\pi = \begin{bmatrix} 0.5 & 0.25 & 0.25 \\ 0.15 & 0.5 & 0.35 \\ 0.05 & 0.45 & 0.5 \end{bmatrix}$$

Next, we compute the reward vector R^π for the policy. Again, let $s \in S$ be some state. The formula for sth coordinate of R^π is

$$R^\pi(s) = \sum_{a\in A} \pi(a|s)R(s,a)$$

and so we get the following.

$$\begin{aligned} R^\pi(s_1) &= 0.5\cdot 1 + 0.5\cdot 2 = 1.5 \\ R^\pi(s_2) &= 0.5\cdot 0 + 0.5\cdot 3 = 1.5 \\ R^\pi(s_3) &= 0.5\cdot 1 + 0.5\cdot 4 = 2.5 \end{aligned}$$

So, the reward vector is $R^\pi = (1.5, 1.5, 2.5)$.

0.2 Problem 2.

Problem Statement. Write down the primal and dual LPs for the MDP given in **Problem 1**. Let your LPs be in standard form with variables lined up on the left and constants on the right of the constraints. Please use v_1, v_2 and v_3 as primal variables and d_{ij} as the dual variable for the constraint corresponding to state s_i and action a_j.

Solution. We will have three primal variables v_1, v_2 and v_3; the optimal values for these variables will correspond to the coordinates $V^*(s_1)$, $V^*(s_2)$ and $V^*(s_3)$ of the optimal value vector. Since we have $3 \times 2 = 6$ state-action pairs, we will have a total of six constraints. From class, we know that the constraints are of the form

$$V^*(s) \geq R(s,a) + \gamma \sum_{s'\in S} \mathbf{P}_{s,s'}(a)V^*(s') \qquad \forall s, a$$

and the objective is to minimize the sum $\sum_{s\in S} V^*(s)$. So, in our case, the primal LP is the following.

$$\begin{array}{lcccccccccc} \text{Minimize:} & v_1+v_2+v_3 \\ \text{Subject to:} & v_1 & \geq & 1 & + & \gamma\cdot 1\cdot v_1 & + & \gamma\cdot 0\cdot v_2 & + & \gamma\cdot 0\cdot v_3 \\ & v_1 & \geq & 2 & + & \gamma\cdot 0\cdot v_1 & + & \gamma\cdot 0.5\cdot v_2 & + & \gamma\cdot 0.5\cdot v_3 \\ & v_2 & \geq & 0 & + & \gamma\cdot 0\cdot v_1 & + & \gamma\cdot 1\cdot v_2 & + & \gamma\cdot 0\cdot v_3 \\ & v_2 & \geq & 3 & + & \gamma\cdot 0.3\cdot v_1 & + & \gamma\cdot 0\cdot v_2 & + & \gamma\cdot 0.7\cdot v_3 \\ & v_3 & \geq & 1 & + & \gamma\cdot 0\cdot v_1 & + & \gamma\cdot 0\cdot v_2 & + & \gamma\cdot 1\cdot v_3 \\ & v_3 & \geq & 4 & + & \gamma\cdot 0.1\cdot v_1 & + & \gamma\cdot 0.9\cdot v_2 & + & \gamma\cdot 0\cdot v_3 \\ & v_1, v_2, v_3 & \in & \mathbf{R} \end{array}$$

Rearranging the above equations to standard form, we get the following.

$$\begin{array}{llccccccc}
\text{Minimize:} & v_1 + v_2 + v_3 & & & & & & \\
\text{Subject to:} & (1-\gamma)v_1 & & & & & \geq & 1 \\
& v_1 & - & 0.5\gamma v_2 & - & 0.5\gamma v_3 & \geq & 2 \\
& & & (1-\gamma)v_2 & & & \geq & 0 \\
& -0.3\gamma v_1 & + & v_2 & - & 0.7\gamma v_3 & \geq & 3 \\
& & & & & (1-\gamma)v_3 & \geq & 1 \\
& -0.1\gamma v_1 & - & 0.9\gamma v_2 & + & v_3 & \geq & 4
\end{array}$$

Now, let us write down the dual of this LP. Our variables will be d_{ij} for $1 \leq i \leq 3$, $1 \leq j \leq 2$. The dual LP is the following.

$$\begin{array}{lccc}
\text{Maximize:} & d_{11} + 2d_{12} + 3d_{22} + d_{31} + 4d_{32} & & \\
\text{Subject to:} & (1-\gamma)d_{11} + d_{12} + 0 \cdot d_{21} - 0.3\gamma d_{22} + 0 \cdot d_{31} - 0.1\gamma d_{32} & = & 1 \\
& 0 \cdot d_{11} - 0.5\gamma d_{12} + (1-\gamma)d_{21} + d_{22} + 0 \cdot d_{31} - 0.9\gamma d_{32} & = & 1 \\
& 0 \cdot d_{11} - 0.5\gamma d_{12} + 0 \cdot d_{21} - 0.7\gamma d_{22} + (1-\gamma)d_{31} + d_{32} & = & 1 \\
& d_{11}, d_{12}, d_{21}, d_{22}, d_{31}, d_{32} & \geq & 0
\end{array}$$

0.3 Problem 3.

Problem Statement. We have seen the Bellman optimality operator B and its properties in class. Write down the definition of the operator and its two properties formally using the right notation. Prove the contraction property of B.

Solution. Let S be the set of states of the MDP in consideration. The Bellman optimality operator B is defined as a map $B : \mathbf{R}^{|S|} \to \mathbf{R}^{|S|}$ given by the following.

$$B[V](s) := \max_{a \in A} \left[R(s,a) + \gamma \sum_{s' \in S} \mathbf{P}_{s,s'}(a) V(s') \right] = \max_{a \in A} Q^V(s,a)$$

This operator satisfies the following two properties.

(1) (Monotonicity) If $u, v \in \mathbf{R}^{|S|}$ are such that $u \leq v$, then $B[u] \leq B[v]$.

(2) (Contraction) If $u, v \in \mathbf{R}^{|S|}$ are any vectors, then

$$||B[u] - B[v]||_\infty \leq \gamma \, ||u - v||_\infty$$

where $\gamma \in (0,1)$ is the discount factor.

Let's prove the contraction property of B. Let $s \in S$ be any state. We will show that

$$|B[u](s) - B[v](s)| \leq \gamma \, ||u - v||_\infty$$

Clearly, the contraction property will follow from this (since we are dealing with the $||\cdot||_\infty$ norm). Now, let a_u and a_v be actions such that $B[u](s) = Q^u(s, a_u)$ and $B[v](s) = Q^v(s, a_v)$. We have the following two cases.

(1) In the first case, suppose $Q^u(s, a_u) \geq Q^v(s, a_v)$. Since $Q^v(s, a_u) \leq Q^v(s, a_v)$, in this

case we see that

$$\begin{aligned}
|B[u](s) - B[v](s)| &= |Q^u(s, a_u) - Q^v(s, a_v)| \\
&\le |Q^u(s, a_u) - Q^v(s, a_u)| \\
&= \left| R(s, a_u) + \gamma \sum_{s' \in S} \mathbf{P}_{s,s'}(a_u) u(s') - R(s, a_u) - \gamma \sum_{s' \in S} \mathbf{P}_{s,s'}(a_u) v(s') \right| \\
&= \left| \gamma \sum_{s' \in S} \mathbf{P}_{s,s'}(a_u)[u(s') - v(s')] \right| \\
&\le \gamma \sum_{s' \in S} \mathbf{P}_{s,s'}(a_u) \, ||u - v||_\infty \\
&= \gamma \, ||u - v||_\infty
\end{aligned}$$

(2) In the second case, we have $Q^u(s, a_u) < Q^v(s, a_v)$. But this case is similar to the first case, with the roles of u and v reversed.

This completes the proof.

0.4 Problem 4.

Problem Statement. We have seen the definitions of $V^\pi(s)$, $Q^V(s, a)$, $B_\pi[V](s)$ and $B[V](s)$ several times in class. Review the definitions and answer the following questions.

(a) Give the equivalent Q-value and Bellman backup value for $V^\pi(s)$, i.e

$$V^\pi(s) = Q^?(s, ?) = B_?[?](s)$$

(b) Give the equivalent Q-value and Bellman backup value for $V^*(s)$, i.e

$$V^*(s) = Q^?(s, ?) = B_?[?](s)$$

(c) Let π and π' be two policies of an MDP. We know that if π' is such that $Q^\pi(s, \pi'(s)) \ge V^\pi(s)$ for all s, then $V^{\pi'} \ge V^\pi$. What can we say about $V^{\pi'}$ and V^π if $Q^\pi(s, \pi'(s)) \le V^\pi(s)$ for all s?

Solution. For part (a), we have

$$\begin{aligned}
V^\pi(s) &= R(s, \pi(s)) + \gamma \sum_{s' \in S} \mathbf{P}_{s,s'}(\pi(s)) V^\pi(s') \\
&= Q^{V^\pi}(s, \pi(s)) \\
&= B_\pi[V^\pi](s)
\end{aligned}$$

For part (b), let π^* be an optimal policy. Then we have

$$\begin{aligned}
V^*(s) &= R(s, \pi^*(s)) + \gamma \sum_{s' \in S} \mathbf{P}_{s,s'}(\pi^*(s)) V^*(s') \\
&= Q^{V^*}(s, \pi^*(s)) \\
&= B_{\pi^*}[V^*](s)
\end{aligned}$$

Finally, we come to part (c). Note that the condition $Q^\pi(s, \pi'(s)) \le V^\pi(s)$ for all s implies that

$$B_{\pi'}[V^\pi](s) \le V^\pi(s) \qquad \forall s \in S$$

which is just saying that

$$B_{\pi'}(V^{\pi}) \leq V^{\pi}$$

Now, by the monotonicity of the operator $B^{\pi'}$, this implies that for all $k \geq 1$ we have

$$B^{k}_{\pi'}(V^{\pi}) \leq V^{\pi}$$

But, we also know that $V^{\pi'} = \lim_{k\to\infty} B^{k}_{\pi'}(V^{\pi})$, and hence this clearly implies that $V^{\pi'} \leq V^{\pi}$.

0.5 Problem 5.

Problem Statement. Let π be a greedy policy with respect to vector $V \in \mathbf{R}^n$. Show that if $||B[V] - V||_{\infty} \leq \epsilon$ then $||V - V^{\pi}||_{\infty} \leq \frac{\epsilon}{1-\gamma}$, where $\gamma \in (0, 1)$.

Solution. Suppose $||B[V] - V||_{\infty} \leq \epsilon$. Since π is a greedy policy with respect to V, we know that

$$\pi(s) = \text{argmax}_{a\in A} Q^{V}(s, a)$$

for each $s \in S$. Now, let $s \in S$ be any state. Then, note that

$$\begin{aligned} B_{\pi}[V](s) &= Q^{V}(s, \pi(s)) && \text{(By definition of } B_{\pi}\text{)} \\ &= \max_{a\in A} Q^{V}(s, a) && \text{(Since } \pi \text{ is greedy w.r.t } V\text{)} \\ &= B[V](s) \end{aligned}$$

where B_{π} is the Bellman backup operator of the policy π. Clearly, this means that

$$B_{\pi}[V] = B[V]$$

and hence we see that

$$||B_{\pi}[V] - V||_{\infty} \leq \epsilon \tag{1}$$

Now, we know that V^{π} is a fixed point of B_{π}, i.e $B_{\pi}[V^{\pi}] = V^{\pi}$. So, by the contraction property of B_{π}, we see that

$$||B_{\pi}[V] - V^{\pi}||_{\infty} = ||B_{\pi}[V] - B_{\pi}[V^{\pi}]||_{\infty} \leq \gamma\, ||V - V^{\pi}||_{\infty} \tag{2}$$

Also, by the triangle inequality, we know that

$$\begin{aligned} \gamma\, ||V - V^{\pi}||_{\infty} &\leq \gamma\, ||B_{\pi}[V] - V^{\pi}||_{\infty} + \gamma\, ||B^{\pi}[V] - V||_{\infty} && (3) \\ &\leq \gamma\, ||B_{\pi}[V] - V^{\pi}||_{\infty} + \epsilon\gamma && \text{(By (1))} \quad (4) \end{aligned}$$

Combining the above equation with (3), we obtain

$$\gamma\, ||V - V^{\pi}||_{\infty} \leq \gamma^2\, ||V - V^{\pi}||_{\infty} + \epsilon\gamma$$

Cancelling γ out from both sides, the claim follows.