ECEN 689: RL: Reinforcement Learning
Exam 1

---

1. (2 point) Consider an MDP $M_1 = (\mathcal{X}, \mathcal{A}, P, R_1, \gamma)$ and let π_1^* be the optimal policy of M_1. Let M_2 be another MDP, exactly the same as M_1 except in its reward function R_2, which is given as $R_2(x,a) = cR_1(x,a), \forall (x,a) \in \mathcal{X} \times \mathcal{A}, c > 0$. Show that the optimal policy of M_2 is also π_1^*.

   Solution:

$$\mathbb{E}_{\pi_1^*}\left[\sum_{t=0}^{\infty} \gamma^t R_1(x_t, a_t)\right] \geq \max_{\pi} \ \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t R_1(x_t, a_t)\right]$$

   where $E_\pi[\cdot]$ indicates the expectation with respect to the transition probability function P when policy π is used. Multiply both sides by c and we get

$$\mathbb{E}_{\pi_1^*}\left[\sum_{t=0}^{\infty} \gamma^t R_2(x_t, a_t)\right] \geq \max_{\pi} \ \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t R_2(x_t, a_t)\right].$$

   So, π_1^* is the optimal policy for the MDP M_2.

2. (2 points) Define the mapping $F : \mathbb{R}^{|\mathcal{X}||\mathcal{A}|} \to \mathbb{R}^{|\mathcal{X}||\mathcal{A}|}$ as

$$(FQ)(x,a) = R(x,a) + \gamma \sum_{y} P(y|x,a) \max_{b} Q(y,b)$$

   Show that F is contraction w.r.t. $\|\cdot\|_\infty$

   Solution: For any arbitrary (x,a),

$$\begin{aligned}
|(FQ_1)(x,a) - (FQ_2)(x,a)| &= |\gamma \sum_{y} P(y|x,a)(\max_{b} Q_1(y,b) - \max_{b} Q_2(y,b))| \\
&\leq \gamma \sum_{y} P(y|x,a)|(\max_{b} Q_1(y,b) - \max_{b} Q_2(y,b))| \\
&\leq \gamma \sum_{y} P(y|x,a) \max_{b} |Q_1(y,b) - Q_2(y,b)| \\
&\leq \gamma \sum_{y} P(y|x,a)\|Q_1 - Q_2\|_\infty = \gamma \|Q_1 - Q_2\|_\infty.
\end{aligned}$$

   Since (x,a) is arbitrary, we have

$$\max_{(x,a)} |(FQ_1)(x,a) - (FQ_2)(x,a)| \leq \gamma \|Q_1 - Q_2\|_\infty.$$

   This, by definition, is the desired result.

3. (4 points) Consider the value iteration algorithm, $V_{k+1} = TV_k$, with $V_0 = 0$. Let k_0 be a given integer. Show that, for any $\epsilon > 0$, we can find an integer n_0 (that will depend on ϵ) such that $\|V_{n_0+k_0} - V_{n_0}\|_\infty < \epsilon$. Give a sufficient condition for selecting such an n_0.

   Solution: Denote $V_k = TV_{k-1} = TTV_{k-2} = \ldots = T^{(k)}V_0$, where $T^{(k)}$ indicates k repeated application of T.

   We can show that $T^{(k)}$ is a contraction with contraction coefficient γ^k. To see this, for any U_1, U_2

   $$\|T^{(k)}U_1 - T^{(k)}U_2\|_\infty = \|TT^{(k-1)}U_1 - TT^{(k-1)}U_2\|_\infty \leq \gamma\|T^{(k-1)}U_1 - T^{(k-1)}U_2\|_\infty.$$

   Repeating this, we get, $\|T^{(k)}U_1 - T^{(k)}U_2\|_\infty \leq \gamma^k\|U_1 - U_2\|_\infty$.

   Now, for any $n \geq 1$,

   $$\begin{aligned}\|V_{n+k_0} - V_n\|_\infty &\leq \sum_{i=0}^{k_0-1} \|V_{n_0+i+1} - V_{n+i}\|_\infty = \sum_{i=0}^{k_0-1} \|T^{(n+i+1)}V_0 - T^{(n+i)}V_0\|_\infty \\ &\leq \sum_{i=0}^{k_0-1} \gamma^{n+i}\|TV_0 - V_0\|_\infty \leq R_{\max}\gamma^n \frac{1}{(1-\gamma)}.\end{aligned}$$

   So, for any $n \geq n_0 = \frac{1}{\log(1/\gamma)} \log \frac{R_{\max}}{\epsilon(1-\gamma)}$, we get $\|V_{n+k_0} - V_n\|_\infty \leq \epsilon$.

4. (8 points) Prof. K has an umbrella that he takes from his home to office and back. If it rains, and if the umbrella is in the place where he is, Prof. K takes the umbrella and goes to the other place, and this involves no cost. However, if he doesn't have the umbrella and it rains, there is a cost C_w for getting wet. If he takes the umbrella with him when it is not raining, he suffers an inconvenience cost C_i. If he does not take the umbrella with him when it is not raining, that incurs no additional cost. Assume that the probability of rain is p and costs are discounted at a factor γ. What is the optimal policy that will minimize the expected cumulative discounted cost?

   (a) (1 point) Formulate this as an MDP with three states.

   Solution: Define three states:
   (i) (s, r) for the case when the umbrella is in the same location as the person and it is raining, (ii) (s, n) for the case when the umbrella is in the same location as the person and it is not raining, (iii) o for the case when the umbrella is in the other location.

   (b) (1 point) How many control policies should we consider?

   Solution: In state (s, n), the person makes the decision whether or not to take the umbrella. In state (s, r), the person has no choice and takes the umbrella. In state o, the person also has no choice and does not take the umbrella.

(c) (3 points) Write down the Bellman optimality equation for all states.

Solution:

$$V(o) = pC_w + \gamma p V(s,r) + \gamma(1-p)V(s,n) \tag{1}$$
$$V(s,r) = \gamma p V(s,r) + \gamma(1-p)V(s,n) \tag{2}$$
$$V(s,n) = \min\{C_i + \ \gamma p V(s,r) + \gamma(1-p)V(s,n),\ \gamma V(o)\} \tag{3}$$

(d) (3 points) What is the optimal policy? Note that this will depend on the value of p (similar to the Homework problem)

Solution: From Eq. (2)

$$V(s,n) = \frac{(1-\gamma p)}{\gamma(1-p)}V(s,r) \tag{4}$$

using this in Eq. (1),

$$V(o) = pC_w + V(s,r), \tag{5}$$

and using this in Eq. (3),

$$V(s,n) = \min\{C_i + V(s,r),\ \gamma p C_w + \gamma V(s,r)\} \tag{6}$$

Now, with Eq. (4),

$$\frac{(1-\gamma p)}{\gamma(1-p)}V(s,r) = \min\{C_i + V(s,r),\ \gamma p C_w + \gamma V(s,r)\} \tag{7}$$

The optimal action at (s,n) is "take umbrella" if

$$C_i + V(s,r) < \gamma p C_w + \gamma V(s,r), \text{ or equivalently, } p > \frac{C_i + (1-\gamma)V(s,r)}{\gamma C_w}.$$

For this scenario, from Eq. (7), $V(s,r) = \frac{C_i\gamma(1-p)}{(1-\gamma)}$. Substituting, this in the above condition for p, we get

$$p > \frac{C_i(1+\gamma)}{\gamma(C_i + C_w)}. \tag{8}$$

5. (4 points) Let $(V_k^i)_{k\geq 1}$ be the sequence of value functions generated by value iteration. Also, let $(V_k^p)_{k\geq 1}$ be the sequence of value functions generated by policy iteration, where $V_k^p = V_{\pi_k}$, and π_k is the policy at iterate k. Assume that $V_0^p = V_0^i$. Then, show that $V_k^i \leq V_k^p \leq V^*$, for all $k \geq 0$, where V^* is the optimal value function.

Solution: We will prove this by induction. For $k = 0$, the hypothesis is true.

Assume that $V_m^i \leq V_m^p \leq V^*$. By the monotoniticy of T, we get

$$TV_m^i = V_{m+1}^i \leq TV_m^p \leq V^* \tag{9}$$

Now, by definition, $TV_m^p = TV_{\pi_m} = T_{\pi_{m+1}}V_{\pi_m}$. Also, the monotone property of $T_{\pi_{m+1}}$,

$$V_{\pi_m} \leq T_{\pi_{m+1}}V_{\pi_m} \leq T_{\pi_{m+1}}^{(2)}V_{\pi_m} \leq \cdots \leq T_{\pi_{m+1}}^{(m)}V_{\pi_m} \leq \cdots \leq \lim_{n\to\infty} T_{\pi_{m+1}}^{(n)}V_{\pi_m} = V_{\pi_{m+1}}.$$

So, we get $T_{\pi_{m+1}}V_{\pi_m} \leq V_{\pi_{m+1}}$. Also, by definition, $V_{\pi_{m+1}} \leq V^*$