ECEN 689: RL: Reinforcement Learning
Assignment 1

---

1. (2 points) Show that the Bellman operator is a monotone operator, i.e, for any $V_1, V_2 \in \mathbb{R}^{|\mathcal{X}|}$ with $V_1 \geq V_2$ (elementwise), $TV_1 \geq TV_2$.

   Solution: For any $x \in \mathcal{X}$,

   $$\begin{aligned}(TV_1)(x) &= \max_a (R(x,a) + \gamma \sum_y P(y|x,a) V_1(y)) \\ &\geq \max_a (R(x,a) + \gamma \sum_y P(y|x,a) V_2(y)) = (TV_2)(x),\end{aligned}$$

   where the inequality is from the assumption $V_1 \geq V_2$.

2. (3 points) Consider the function $f : \mathbb{R}^n \to \mathbb{R}^n$, $f(u) = Au$, where $A \in \mathbb{R}^n \times \mathbb{R}^n$. Assume that the row sums of A is strictly less than 1, i.e., $\sum_j |a_{ij}| \leq \alpha < 1$. Show that $f(\cdot)$ is a contraction mapping with respect to $\|\cdot\|_\infty$.

   Solution:

   $$\begin{aligned}\|Au - Av\|_\infty = \|A(u-v)\|_\infty &= \max_{i,1\leq i\leq n} \left| \sum_{j=1}^n a_{ij}(u_j - v_j) \right| \\ &\leq \max_{i,1\leq i\leq n} \sum_{j=1}^n |a_{ij}| \ |(u_j - v_j)| \leq \max_{i,1\leq i\leq n} \sum_{j=1}^n |a_{ij}| \ \|u-v\|_\infty \leq \alpha \|u-v\|_\infty\end{aligned}$$

3. (4 points) Let $\mathcal{U}$ be a given set, and $g_1 : \mathcal{U} \to \mathbb{R}$ and $g_2 : \mathcal{U} \to \mathbb{R}$ be two real-valued functions on $\mathcal{U}$. Also assume that both functions are bounded. Show that

   $$|\max_u g_1(u) - \max_u g_2(u)| \leq \max_u |g_1(u) - g_2(u)|$$

   Solution:

   $$\begin{aligned}\max_x g_1(x) &= \max_x \left(g_1(x) + g_2(x) - g_2(x)\right) \leq \max_x (g_2(x) + |g_1(x) - g_2(x)|) \\ &\leq \max_x (g_2(x) + \max_y |g_1(y) - g_2(y)|) = \max_x g_2(x) + \max_y |g_1(y) - g_2(y)|\end{aligned}$$

   From this, we can get

   $$\max_x g_1(x) - \max_x g_2(x) \leq \max_x |g_1(x) - g_2(x)|$$

   Similarly, we can get

   $$\max_x g_2(x) - \max_x g_1(x) \leq \max_x |g_1(x) - g_2(x)|$$

   Combining both, we will get the desired result.

4. (6 points) Consider a finite MDP $(\mathcal{X}, \mathcal{A}, P, R, \gamma)$. Assume that $\max_x \max_a |R(x,a)| = R_{\max}$. Let V^* be the optimal value function. Consider the value iteration algorithm $V_{k+1} = TV_k$, with $V_0 = 0$, where T is the Bellman operator.

   (a) (1 points) Show that $||V^*||_\infty \leq R_{\max}/(1-\gamma)$.
   Solution:

   $$|V^*(x)| = |\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t R(x_t, a_t)]| \leq \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t |R(x_t, a_t)|] \leq \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t R_{\max}] = \frac{R_{\max}}{(1-\gamma)}$$

   Since this is true for any x, we get $\max_x |V^*(x)| \leq R_{\max}/(1-\gamma)$, which is by definition $||V^*||_\infty \leq R_{\max}/(1-\gamma)$.

   (b) (2 points) Show that $\|V_k - V^*\|_\infty \leq \epsilon$, for $k \geq \frac{1}{\log(1/\gamma)} \log\left(\frac{R_{\max}}{\epsilon(1-\gamma)}\right)$.
   Solution: We have

   $$\|V_k - V^*\|_\infty \leq \gamma^k \|V_0 - V^*\|_\infty = \gamma^k \|V^*\|_\infty \leq \frac{\gamma^k R_{\max}}{(1-\gamma)}$$

   The RHS is less than or equal to ϵ for $k \geq \frac{1}{\log(1/\gamma)} \log\left(\frac{R_{\max}}{\epsilon(1-\gamma)}\right)$.

   (c) (3 points) Let m be an integer such that $||V_{m+1} - V_m||_\infty \leq \frac{\epsilon(1-\gamma)}{2\gamma}$. Then, show that $||V_{m+1} - V^*||_\infty \leq \epsilon/2$. Discuss how this result can be used as a stopping criteria for the value iteration algorithm.
   Solution:

   $$\|V_m - V^*\|_\infty \leq \|V_m - V_{m+1}\|_\infty + \|V_{m+1} - V^*\|_\infty \leq \|V_m - V_{m+1}\|_\infty + \gamma \, \|V_m - V^*\|_\infty$$

   From this, we get $\|V_m - V^*\|_\infty \leq \frac{1}{(1-\gamma)} \|V_m - V_{m+1}\|_\infty \leq \epsilon/2\gamma$.
   Now, $\|V_{m+1} - V^*\| \leq \gamma \, \|V_m - V^*\| \leq \epsilon/2$.

5. (4 points) Consider two finite MDPs $M_1 = (\mathcal{X}, \mathcal{A}, P_1, R, \gamma)$ and $M_1 = (\mathcal{X}, \mathcal{A}, P_2, R, \gamma)$ that differ only in the transition probability functions. Let V_1^* and V_2^* be the optimal value function of M_1 and M_2, respectively. Assume that $\max_{x,a} \|P_1(\cdot|x,a) - P_2(\cdot|x,a)\|_1 = \epsilon$. Show that $\|V_1^* - V_2^*\|_\infty \leq \frac{\gamma \epsilon R_{\max}}{(1-\gamma)^2}$, where $R_{\max} = \max_{x,a} |R(x,a)|$.
   Solution:

   $$|V_1^*(x) - V_1^*(x)| = |\max_a (R(x,a) + \gamma \sum_y P_1(y|x,a) V_1^*(y)) - \max_a (R(x,a) + \gamma \sum_y P_2(y|x,a) V_2^*(y))|$$

   $$\leq \gamma \max_a |\sum_y P_1(y|x,a) V_1^*(y) - \sum_y P_2(y|x,a) V_2^*(y)|$$

   $$\leq \gamma \max_a |\sum_y P_1(y|x,a)(V_1^*(y) - V_2^*(y)) + \sum_y (P_1(y|x,a) - P_2(y|x,a)) V_2^*(y)|$$

   $$\leq \gamma \max_a (\sum_y P_1(y|x,a) \|V_1^* - V_2^*\|_\infty + \sum_y |(P_1(y|x,a) - P_2(y|x,a))| \|V_2^*\|_\infty$$

   $$\leq \gamma \|V_1^* - V_2^*\| + \gamma \epsilon \frac{R_{\max}}{(1-\gamma)}$$

This implies $\|V_1^* - V_2^*\| \leq \gamma \|V_1^* - V_2^*\| + \gamma\epsilon \frac{R_{\max}}{(1-\gamma)}$. The desired result follows from this.

6. (5 points) Let $\bar{Q}$ be such that $\|\bar{Q} - Q^*\|_\infty \leq \epsilon$. Let $\bar{\pi}$ be the greedy policy with respect to $\bar{Q}$, i.e.,$\bar{\pi}(x) = \arg\max_a \bar{Q}(x,a)$. Show that $\|V^* - V_{\bar{\pi}}\|_\infty \leq \frac{2\epsilon}{(1-\gamma)}$.

   Solution:

$$
\begin{aligned}
&|V^*(x) - V_{\bar{\pi}}(x)| = |Q^*(x,\pi^*(x)) - Q_{\bar{\pi}}(x,\bar{\pi}(x))| \\
&\leq |Q^*(x,\pi^*(x)) - Q^*(x,\bar{\pi}(x))| + |Q^*(x,\bar{\pi}(x)) - Q_{\bar{\pi}}(x,\bar{\pi}(x))| \\
&\leq |Q^*(x,\pi^*(x)) - Q^*(x,\bar{\pi}(x))| + \gamma \sum_y P(y|x,\bar{\pi}(x))|V^*(y) - V_{\bar{\pi}}(y)| \\
&\leq |Q^*(x,\pi^*(x)) - Q^*(x,\bar{\pi}(x))| + \gamma \|V^* - V_{\bar{\pi}}\|_\infty \\
&\leq |Q^*(x,\pi^*(x)) - \bar{Q}(x,\pi^*(x))| + |\bar{Q}(x,\pi^*(x)) - Q^*(x,\bar{\pi}(x))| + \gamma \|V^* - V_{\bar{\pi}}\|_\infty \\
&\overset{(i)}{\leq} |Q^*(x,\pi^*(x)) - \bar{Q}(x,\pi^*(x))| + |\bar{Q}(x,\bar{\pi}(x)) - Q^*(x,\bar{\pi}(x))| + \gamma \|V^* - V_{\bar{\pi}}\|_\infty \\
&\leq 2\|\bar{Q} - Q^*\|_\infty + \gamma \|V^* - V_{\bar{\pi}}\|_\infty,
\end{aligned}
$$

   where (i) is obtained by the fact that $\bar{Q}(x,\bar{\pi}(x)) \geq \bar{Q}(x,\pi^*(x))$. We can now get the desired bound.

7. (6 points) A spider and fly move along a straight line at times $t = 0, 1, \ldots$. The initial position of the fly and the spider are integers. At each time period, the fly moves one unit to the left with a probability p, one unit to the right with a probability p, and stays where it is with a probability $1 - 2p$. The spider, knows the position of the fly at the beginning of each period, and will always move one unit towards the fly if its distance from the fly is more than one unit. If the spider is one unit away from the fly, it will either move one unit towards the fly, or stay where it is. If the spider and the fly land in the same position at the end of a period, then the spider captures the fly, and the process terminates. The spider's objective is to capture the fly in minimum expected number of steps.

   (a) Give a closed-form expression for the expected number of steps for capture when the spider is one unit away from the fly.

   (b) Give a closed-form expression for the expected number of steps for capture when the spider is two units away from the fly.

   Solution:

   Define the state as the distance between the spider and the fly. State 0 is the termination state where the spider catches the fly. The actions are move (`m`) or not move (`nm`). The control action for states $x \geq 2$ is already given as `m`. The transition probability function for states $x \geq 2$ is given by

$$p(x|x,\mathtt{m}) = p, \quad p(x-1|x,\mathtt{m}) = 1-2p, \quad p(x-2|x,\mathtt{m}) = p$$

and zero for any other transitions.

For $x = 1$,

$$p(1|1, \mathtt{m}) = 2p,\ p(0|1, \mathtt{m}) = 1 - 2p,\ p(1|1, \mathtt{nm}) = 1 - 2p,\ p(0|1, \mathtt{nm}) = p,\ p(2|1, \mathtt{nm}) = p$$

Instead of a reward maximization, we will treat this as a cost minimization problem. Assume that the cost $c(x, a) = 1$ for all (x, a) until the termination. Also assume that $\gamma = 1$. Then, the value function for any policy is the expected number of steps for capture under that policy. The optimal value function will give the minimum expected number of steps for capture. We will find the optimal value function by solving Bellman equation.

By definition. $V^*(0) = 0$. We can write the Bellman equation for $x \geq 2$ as follows:

$$V^*(x) = 1 + pV^*(x) + (1 - 2p)V^*(x - 1) + pV^*(x - 2), x \geq 2. \tag{1}$$

The only state where the spider has a choice is when it is one unit away from the fly, and for that state Bellman's equation is given by

$$V^*(1) = 1 + \min\{2pV^*(1),\ (1 - 2p)V^*(1) + pV^*(2)\}, \tag{2}$$

where the second expression with in the bracket is associated with the spider moving and not moving, respectively. By (1) for $x = 2$, we obtained

$$V^*(2) = 1 + pV^*(2) + (1 - 2p)V^*(1),$$

from which we get

$$V^*(2) = \frac{1}{1 - p} + \frac{(1 - 2p)V^*(1)}{1 - p}. \tag{3}$$

Substituting this expression in (2), we obtain

$$V^*(1) = 1 + \min\left\{2pV^*(1),\quad \frac{p}{1 - p} + \frac{p(1 - 2p)V^*(1)}{1 - p} + (1 - 2p)V^*(1)\right\},$$

or equivalently,

$$V^*(1) = 1 + \min\left\{2pV^*(1),\quad \frac{p}{1 - p} + \frac{(1 - 2p)V^*(1)}{1 - p}\right\}.$$

To solve the above equation, we consider the two cases where the first expression within the bracket is larger and is smaller than the second expression. Thus we solve for $V^*(1)$ in the two cases where,

$$V^*(1) = 1 + 2pV^*(1),\quad 2pV^*(1) \leq \frac{p}{1 - p} + \frac{(1 - 2p)V^*(1)}{1 - p}, \tag{4}$$

and

$$V^*(1) = 1 + \frac{p}{1-p} + \frac{(1-2p)V^*(1)}{1-p}, \quad 2pV^*(1) \geq \frac{p}{1-p} + \frac{(1-2p)V^*(1)}{1-p}. \qquad (5)$$

From (4),

$$V^*(1) = \frac{1}{1-2p}$$

and this solution is valid when

$$\frac{2p}{1-2p} \leq \frac{p}{1-p} + \frac{1}{1-p}$$

or equivalently, $p \leq 1/3$.

Thus, for $p \leq 1/3$, it is optimal for the spider to move when it is one unit away from the fly.

Similarly, from (5),

$$V^*(1) = \frac{1}{p},$$

and this solution is valid when

$$2 \geq \frac{p}{1-p} + \frac{(1-2p)V^*(1)}{1-p},$$

or equivalently, $p \geq 1/3$.

Thus, for $p \geq 1/3$, it is optimal for the spider not to move when it is one unit away from the fly.

The minimum expected number of steps for capture when the spirder is one unit away from the fly was calculated earlier to be

$$V^*(1) = \begin{cases} 1/(1-2p) & \text{if } p \leq 1/3 \\ 1/p & \text{if } p \geq 1/3 \end{cases}$$

Given the value of $V^*(1)$, we can now calculate $V^*(2)$ using (3).