ECEN 689: RL: Fall 2019
Assignment 2

1. (5 points) Let $\mathcal{X}$ be a given set, and $g_1 : \mathcal{X} \to \mathbb{R}$ and $g_2 : \mathcal{X} \to \mathbb{R}$ be two real-valued functions on $\mathcal{X}$. Also assume that both function are bounded. Then, show that

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

Answer:

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

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.

2. (2 points) The reward function is given by $R(x, a)$ for a given state x and given action a. Suppose the policy $\pi(\cdot)$ is stochastic. If the state at time t is x_t, and actions are selected according to policy π, then what is the expected reward at time t?

Answer: $\sum_a R(x_t, a)\pi(x_t, a)$

3. (5 points) Consider the MDP shown in Fig 1

The only decision to be made is that in the top state, where two actions are available, left and right. The numbers show the rewards that are received deterministically after each action. There are exactly two deterministic policies, π_{left} and π_{right}. What policy is optimal if $\gamma = 0$? If $\gamma = 0.9$? $\gamma = 0.5$?

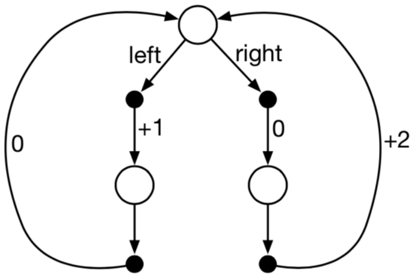

Figure 1: MDP

Answer: If $\gamma = 0$, the delayed rewards do not count at all. The left action has value 1 from the top state, and the right action has value 0; π_{left} is optimal.

If $\gamma = 0.9$, then delayed rewards are given substantial weight. Under π_{left}, the sequence of rewards from the top state is $1, 0, 1, 0, 1, 0, \ldots$, and the corresponding return and value is $1 + \gamma^2 + \gamma^4 + \ldots = 1/(1 - \gamma^2) \approx 5.26$. Under π_{right}, the sequence of rewards from the top state is $0, 2, 0, 2, 0, 2, \ldots$, and the corresponding return and value is $2\gamma + 2\gamma^3 + 2\gamma^5 + \ldots = 2\gamma/(1 - \gamma^2)$, which is clearley better as $2\gamma > 1$.

If $\gamma = 0.5$, we have a borderline case. The returns are exactly as computed above, but now we have $2\gamma = 1$, and thus the two actions have the same value. Both policies are optimal.

4. (8 points) Consider a finite MDP $(\mathcal{X}, \mathcal{A}, P, R, \gamma)$. Assume that $\max_x \max_a |R(x, a)| = C$. Let V^* be the optimal value function. Also, $|| \cdot ||$ indicates the sup norm.

   (a) (2 points) What is the maximum possible value for $||V^*||$?

   Answer: Maximum possible value is $C/(1 - \gamma)$.

   (b) (3 points) Consider the Value Iteration algorithm $V_{k+1} = TV_k$, where T is the Bellman operator. Show that for any given $\epsilon > 0$, there exists a finite integer k_0 such that $||V_k - V^*|| \leq \epsilon$ for all $k \geq k_0$. What is the minimum value of k_0?

   Answer: We have

$$\|V_k - V^*\| \leq \gamma\|V_{k-1} - V^*\| \leq \ldots \leq \gamma^k\|V_0 - V^*\|$$

   Let k_0 be the the smallest integer such that

$$\gamma^{k_0}\|V_0 - V^*\| \leq \epsilon$$

   Then,

$$k_0 \ \log \gamma \leq \log\left(\frac{\epsilon}{\|V_0 - V^*\|}\right)$$

   Assuming that $\frac{\epsilon}{\|V_0 - V^*\|} < 1$ (this is typically the case when ϵ is very small), we get

$$k_0 \geq \frac{1}{\log(1/\gamma)} \log\left(\frac{\|V_0 - V^*\|}{\epsilon}\right)$$

   Now, if we assume that $\max_x \max_a |R(x, a)| = C$, you can argue that

$$k_0 \geq \frac{1}{\log(1/\gamma)} \log\left(\frac{2C}{\epsilon(1 - \gamma)}\right)$$

   (c) (3 points) Consider the Value Iteration algorithm $V_{k+1} = TV_k$, where T is the Bellman operator. Let m be an integer such that

$$||V_{m+1} - V_m|| \leq \frac{\epsilon(1 - \gamma)}{2\gamma}$$

Then, show that

$$||V_{m+1} - V^*|| \leq \epsilon/2.$$

Discuss how this result can be used as a stopping criteria for the Value Iteration algorithm

Answer:

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

From this, we get

$$\|V_m - V^*\| \leq \frac{1}{(1-\gamma)} \|V_m - V_{m+1}\| \leq \frac{\epsilon}{2\gamma}$$

Now,

$$\|V_{m+1} - V^*\| \leq \gamma \; \|V_m - V^*\| \leq \frac{\epsilon}{2}$$

5. (5 points) Suppose you terminate the Value Iteration algorithm after n steps and let $\hat{V} = V_n = T^n V_0$ be the corresponding approximate value function. Assume that $||\hat{V} - V^*|| < \epsilon$. Now, compute the approximate policy $\hat{\pi}$ as

$$\hat{\pi}(x) = \arg\max_a (R(x,a) + \gamma \sum_{y \in \mathcal{X}} P(y|x,a)\hat{V}(y))$$

Let $V_{\hat{\pi}}$ be the value corresponding to the policy $\hat{\pi}$. Show that

$$||V_{\hat{\pi}} - V^*|| \leq \frac{2\gamma\epsilon}{(1-\gamma)}$$

Answer:

$$\begin{aligned}
V_{\hat{\pi}}(x) &= R(x,\hat{\pi}(x)) + \gamma \sum_y P(y|x,\hat{\pi}(x)) V_{\hat{\pi}}(y) \\
&= R(x,\hat{\pi}(x)) + \gamma \sum_y P(y|x,\hat{\pi}(x))(V_{\hat{\pi}}(y) - \hat{V}(y) + \hat{V}(y)) \\
&= R(x,\hat{\pi}(x)) + \gamma \sum_y P(y|x,\hat{\pi}(x))\hat{V}(y) + \gamma \sum_y P(y|x,\hat{\pi}(x))(V_{\hat{\pi}}(y) - \hat{V}(y)) \\
&= \max_a (R(x,a) + \gamma \sum_y P(y|x,a)\hat{V}(y)) + \gamma \sum_y P(y|x,\hat{\pi}(x))(V_{\hat{\pi}}(y) - \hat{V}(y)) \\
&\geq \max_a (R(x,a) + \gamma \sum_y P(y|x,a)(V^*(y) - \epsilon)) + \gamma \sum_y P(y|x,\hat{\pi}(x))(V_{\hat{\pi}}(y) - (V^*(y) + \epsilon)) \\
&= V^*(x) + \gamma \sum_y P(y|x,\hat{\pi}(x))(V_{\hat{\pi}}(y) - V^*(y)) - 2\gamma\epsilon
\end{aligned}$$

From this, we get

$$V^*(x) - V_{\hat{\pi}}(x) \leq 2\gamma\epsilon + \gamma \sum_y P(y|x,\hat{\pi}(x))(V_{\hat{\pi}}(y) - V^*(y)) \leq 2\gamma\epsilon + \gamma\|V^* - V_{\hat{\pi}}\|$$

This implies

$$\|V^* - V_{\hat{\pi}}\| \leq 2\gamma\epsilon + \gamma\|V^* - V_{\hat{\pi}}\|$$

And the result follows.

6. (15 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 integer. 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.

Answer:

Define the state as the distance between the spider and the fly. State 0 is the termination state where the spider catches the fly. Let $p_{i,j}$ e the transition probability from state i to state j when $i \geq 2$. Then,

$$p_{i,i} = p, \qquad p_{i,i-1} = 1 - 2p, \qquad p_{i,i-2} = 1 - 2p, \qquad i \geq 2.$$

Denote $p_{1,j}(m)$ and $p_{1,j}(\bar{m})$ the transition probabilities from state 1 to state j if the spider moves and does not move, respectively. Then,

$$p_{1,1}(m) = 2p, p_{1,0}(m) = 1 - 2p, p_{1,2}(\bar{m}) = p, p_{1,1}(\bar{m}) = 1 - 2p, p_{1,0}(\bar{m}) = p$$

For states $i \geq 2$, Bellman's equation is written as

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

where $V^*(0) = 0$ by definition. 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

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

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

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

from which

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

Substituting this expression in (2), we obtain

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

or equivalently,

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

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,

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

and

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

From (4),

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

and this solution is valid when

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

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

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).