National University of Singapore
School of Computing
CS3243 Introduction to AI

Tutorial 6: Reinforcement Learning

Issued: March 13, 2020 Due: Week 9 In Tutorial Class

Important Instructions:

- *Your solutions for this tutorial must be TYPE-WRITTEN.*
- *Make TWO copies of your solutions: one for you and one to be SUBMITTED TO THE TUTOR IN CLASS. Your submission in your respective tutorial class will be used to indicate your CLASS ATTENDANCE. Late submission will NOT be entertained.*
- *YOUR SOLUTION TO QUESTION* 1 *will be GRADED for this tutorial.*
- *You may discuss the content of the questions with your classmates. But everyone should work out and write up ALL the solutions by yourself.*

We say that a policy π is *deterministic* if, given the current state s_t and the sequence of actions/states/rewards from time 0 to $t-1$, the action chosen next, $\pi(s_t)$, is not randomized. In other words, an adversary observing what the agent does at times $0, \ldots, t-1$ can completely predict what their action will be in time t. A policy π is randomized if we allow an agent to follow a distribution over actions, rather than a single action.

1. We have argued in class that a deterministic policy may result in extremely suboptimal outcomes. This is especially true when the state transition model is *adversarial*, i.e. the next state is chosen by an adversary who wants to minimize your reward. Consider the game of scissors/paper/stone played repeatedly (infinitely many times). At each turn, Player 1 and Player 2 pick either "scissors" "paper" or "stone". The states and rewards are given as $\langle state \rangle \rightarrow reward$ below

$$\begin{aligned}
\langle scissors, paper \rangle &\rightarrow 1 \\
\langle scissors, stone \rangle &\rightarrow -1 \\
\langle paper, scissors \rangle &\rightarrow -1 \\
\langle paper, stone \rangle &\rightarrow 1 \\
\langle stone, paper \rangle &\rightarrow -1 \\
\langle stone, scissors \rangle &\rightarrow 1
\end{aligned}$$

When Player 2 picks the same action as Player 1, Player 1's reward is 0. Player 1 picks the lefthand action of the tuple, whereas Player 2 picks the righthand action.

(a) Suppose that Player 1 follows a deterministic policy π to pick their next move at time t. Assuming that Player 2 observes everything that Player 1 does and knows the policy π that Player 1 uses, what is the optimal (reward maximizing) policy for Player 2 to follow?

Solution: In that case Player 2 should just pick the action that beats the deterministic action chosen by player 1. This is similar to the idea we've seen in regret minimization.

(b) What is the optimal randomized policy for Player 1, assuming that Player 2 will choose their action adversarially? (i.e. to minimize Player 1's revenue, under the worst-case assumption that Player 2 knows exactly Player 1's policy)

Hint: While it is intuitively 'obvious' that the best thing to do is to pick one's action uniformly at random, you need to formally argue why: what happens if Player 1's policy is not uniform? Can you come up with a simple strategy for Player 2 to get better revenue?

Solution: The best policy is to play each action with prob. $\frac{1}{3}$. Suppose that this is not the case, and Player 1 plays some other distribution $(p_{scissors}, p_{paper}, p_{stone})$; then some action is played the highest probability, which is in particular $> \frac{1}{3}$ (say, $p_{scissors} > \frac{1}{3}$). Let's assume that there is a unique action with highest probability (if there are two such actions the proof is similar). The expected reward for Player 1, assuming that Player 2 aims to make them lose as much as possible, is negative. To see why - suppose that Player 2 plays $stone$ all the time. The expected reward for Player 1 is:

$$p_{scissors} \times (-1) + p_{paper} \times 1 + p_{stone} \times 0 = p_{paper} - p_{scissors} < 0.$$

On the other hand, it is easy to verify that Player 1's expected reward is 0 when they play $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, an improvement over a negative reward.

2. Consider the MDP described in Figure 1

The agent has two actions, a_0 and a_1, whose effects in each state $\sigma_0, \ldots, \sigma_3$ are described in Figure 1. The edges from actions are labeled with the probability that this transition occurs. For example, $\Pr[s_{t+1} = \sigma_2 \mid s_t = \sigma_0, a_t = a_1] = 1$; similarly, $\Pr[s_{t+1} = \sigma_0 \mid s_t = \sigma_1, a_t = a_0] = 1 - p$. If there is no edge from a state to an action, that action results in you remaining in the same state. Thus, choosing either a_0 or a_1 in σ_3 results in the agent remaining in state σ_3 (it is a sink state); action a_1 cannot be taken in state σ_1. The rewards in each state are independent of the actions, and are $r(\sigma_0) = r(\sigma_2) = 0; r(\sigma_1) = 1; r(\sigma_3) = 10$.

(a) What are the possible (deterministic) policies for this MDP?

Solution: There are four policies in total, determined by one's actions at σ_0 and σ_2:

$$\pi_1(\sigma_0) = a_0; \pi_1(\sigma_2) = a_0$$
$$\pi_2(\sigma_0) = a_0; \pi_2(\sigma_2) = a_1$$
$$\pi_3(\sigma_0) = a_1; \pi_3(\sigma_2) = a_0$$
$$\pi_4(\sigma_0) = a_1; \pi_4(\sigma_2) = a_1$$

(b) What is the value function for each of these policies, when the discount factor is $0 < \gamma < 1$ and we start from s_0? (in other words compute $V^\pi(\sigma_0)$ for each π you found in the previous point)

Solution: For all policies we have that $V(\sigma_3) = 10 + \gamma V(\sigma_3) \Rightarrow V(\sigma_3) = \frac{10}{1-\gamma}$.
For π_1, we have

$$\begin{aligned} V_1(\sigma_0) &= 0 + \gamma V_1(\sigma_1) \\ V_1(\sigma_1) &= 1 + \gamma(p \times V_1(\sigma_2) + (1-p) \times V_1(\sigma_0)) \\ V_1(\sigma_2) &= 0 + \gamma V_1(\sigma_1) \end{aligned}$$

Plugging in the values of $V_1(\sigma_0)$ and $V_1(\sigma_2)$ into $V_1(\sigma_1)$ we get that $V_1(\sigma_1) = 1 + \gamma^2 V_1(\sigma_1)$, i.e. $V_1(\sigma_1) = \frac{1}{1-\gamma^2}$, and therefore $V_1(\sigma_0) = V_1(\sigma_2) = \frac{\gamma}{1-\gamma^2}$.
For π_2 we have

$$\begin{aligned} V_2(\sigma_0) &= 0 + \gamma V_2(\sigma_1) \\ V_2(\sigma_1) &= 1 + \gamma(p \times V_2(\sigma_2) + (1-p) \times V_2(\sigma_0)) \\ V_2(\sigma_2) &= 0 + \gamma(q \times V_2(\sigma_2) + (1-q) \times V_2(\sigma_3)) \end{aligned}$$

Plugging in the value of $V_2(\sigma_3)$ we have

$$\begin{aligned} V_2(\sigma_2) &= \gamma\left(q \times V_2(\sigma_2) + (1-q) \times \frac{10}{1-\gamma}\right) && \iff \\ (1-\gamma q)V_2(\sigma_2) &= \frac{10\gamma(1-q)}{1-\gamma} && \iff \\ V_2(\sigma_2) &= \frac{10\gamma(1-q)}{(1-\gamma)(1-\gamma q)} \end{aligned}$$

Plugging these into $V_2(\sigma_1)$ we get

$$\begin{aligned} V_2(\sigma_1) =& 1 + \gamma \left(p \times \frac{10\gamma(1-q)}{(1-\gamma)(1-\gamma q)} + (1-p) \times \gamma V_2(\sigma_1) \right) \quad \Longleftrightarrow \\ \left(1-(1-p)\gamma^2\right) V_2(\sigma_1) =& 1 + \gamma \left(p \times \frac{10\gamma(1-q)}{(1-\gamma)(1-\gamma q)} \right) \quad \Longleftrightarrow \\ V_2(\sigma_1) =& \frac{1 + \frac{10\gamma^2 p(1-q)}{(1-\gamma)(1-\gamma q)}}{(1-(1-p)\gamma^2)} = \frac{(1-\gamma)(1-\gamma q) + 10\gamma^2 p(1-q)}{(1-\gamma)(1-\gamma q)\left(1-(1-p)\gamma^2\right)} \end{aligned}$$

from which $V_2(\sigma_0) = \frac{\gamma\left((1-\gamma)(1-\gamma q)+10\gamma^2 p(1-q)\right)}{(1-\gamma)(1-\gamma q)(1-(1-p)\gamma^2)}$.
For π_3 we have

$$\begin{aligned} V_3(\sigma_0) =& 0 + \gamma V_3(\sigma_2) \\ V_3(\sigma_1) =& 1 + \gamma \left(p \times V_3(\sigma_2) + (1-p) \times V_3(\sigma_0) \right) \\ V_3(\sigma_2) =& 0 + \gamma V_3(\sigma_1) \end{aligned}$$

Therefore $V_3(\sigma_0) = \gamma^2 V_3(\sigma_1)$. Plugging these into $V_3(\sigma_1)$ we get

$$\begin{aligned} V_3(\sigma_1) =& 1 + \gamma \left(p \times \gamma V_3(\sigma_1) + (1-p) \times \gamma^2 V_3(\sigma_1) \right) \quad \Longleftrightarrow \\ V_3(\sigma_1) =& \frac{1}{(1-\gamma)(1+\gamma+\gamma^2-\gamma^2 p)} \end{aligned}$$

which automatically yields $V_3(\sigma_0) = \frac{\gamma^2}{(1-\gamma)(1+\gamma+\gamma^2-\gamma^2 p)}$
For π_4 we have

$$\begin{aligned} V_4(\sigma_0) =& 0 + \gamma V_4(\sigma_2) \\ V_4(\sigma_1) =& 1 + \gamma \left(p \times V_4(\sigma_2) + (1-p) \times V_4(\sigma_0) \right) \\ V_4(\sigma_2) =& 0 + \gamma \left(q V_4(\sigma_3) + (1-q) V_4(\sigma_2) \right) \end{aligned}$$

We first compute $V_4(\sigma_2) = \frac{10\gamma q}{1-\gamma} + \gamma(1-q)V_4(\sigma_2)$, i.e. $V_4(\sigma_2) = \frac{10\gamma q}{(1-\gamma)(1-\gamma(1-q))}$. Thus, $V_4(\sigma_0) = \frac{10\gamma^2 q}{(1-\gamma)(1-\gamma(1-q))}$

(c) Suppose that $p = q = 0.5$. Decide what is the optimal policy when we start from σ_0, as a function of γ (you may still assume that $0 < \gamma < 1$).

Solution: First let us write the values of $V_i(\sigma_0)$ when $p = q = 0.5$.

$$\begin{aligned}
V_1(\sigma_0) =& \frac{\gamma}{1-\gamma^2} \\
V_2(\sigma_0) =& \frac{\gamma\left((1-\gamma)(1-0.5\gamma)+2.5\gamma^2\right)}{(1-\gamma)(1-0.5\gamma)\left(1-0.5\gamma^2\right)} \\
V_3(\sigma_0) =& \frac{\gamma^2}{(1-\gamma)(1+\gamma+0.5\gamma^2)} \\
V_4(\sigma_0) =& \frac{5\gamma^2}{(1-\gamma)(1-0.5\gamma)}
\end{aligned}$$

One thing that is immediately obvious is that $5\gamma^2 > \gamma^2$ and $(1-\gamma)(1+\gamma+0.5\gamma^2) > (1-0.5\gamma)$; therefore $V_4(\sigma_0) > V_3(\sigma_0)$ for any value of γ. Next,

$$\begin{aligned}
V_4(\sigma_0) \geq& V_1(\sigma_0) & \iff \\
\frac{5\gamma^2}{(1-\gamma)(1-0.5\gamma)} \geq& \frac{\gamma}{1-\gamma^2} & \iff \\
\frac{10\gamma}{(2-\gamma)} \geq& \frac{1}{1+\gamma} & \iff \\
10\gamma(1+\gamma) \geq& 2-\gamma & \iff \\
10\gamma^2+11\gamma-2 \geq& 0
\end{aligned}$$

Solving the quadratic equation we obtain that this holds true when $\frac{\sqrt{201}-11}{20} \leq \gamma < 1$ (so $\gamma \geq 0.151$ or so).

Next,

$$\begin{aligned}
V_4(\sigma_0) \geq& V_2(\sigma_0) & \iff \\
\frac{5\gamma^2}{(1-\gamma)(1-0.5\gamma)} \geq& \frac{\gamma\left((1-\gamma)(1-0.5\gamma)+2.5\gamma^2\right)}{(1-\gamma)(1-0.5\gamma)\left(1-0.5\gamma^2\right)} & \iff \\
5\gamma \geq& \frac{(1-\gamma)(1-0.5\gamma)+2.5\gamma^2}{1-0.5\gamma^2} & \iff \\
5\gamma \geq& \frac{6\gamma^2-3\gamma+2}{2-\gamma^2} & \iff \\
10\gamma-5\gamma^3 \geq& 6\gamma^2-3\gamma+2
\end{aligned}$$

We solve this analytically to obtain $\gamma \geq \frac{\sqrt{161}-11}{10} \simeq 0.168$. In other words, when

$\gamma \geq 0.168$ π_4 is the optimal policy. Finally, let us compare $V_1(\sigma_0)$ and $V_2(\sigma_0)$:

$$\begin{aligned}
V_1(\sigma_0) \geq & V_2(\sigma_0) & \iff \\
\frac{\gamma}{1-\gamma^2} \geq & \frac{\gamma\left((1-\gamma)(1-0.5\gamma)+2.5\gamma^2\right)}{(1-\gamma)(1-0.5\gamma)\left(1-0.5\gamma^2\right)} & \iff \\
\frac{1}{1+\gamma} \geq & \frac{(1-\gamma)(1-0.5\gamma)+2.5\gamma^2}{(1-0.5\gamma)\left(1-0.5\gamma^2\right)} = \frac{3\gamma^2-1.5\gamma+1}{(1-0.5\gamma)\left(1-0.5\gamma^2\right)} & \iff \\
0.25\gamma^3-0.5\gamma^2-0.5\gamma+1 \geq & (1+\gamma)(3\gamma^2-1.5\gamma+1)
\end{aligned}$$

which when solved analytically yields no solutions for γ in $(0, 1)$. In other words, $V_2(\sigma_0)$ dominates $V_1(\sigma_0)$ always.

(d) Try to understand what happens to this MDP as we vary the values of p and q, as well as the relative rewards in σ_1 and σ_3, and the relation of these to the discount factor.

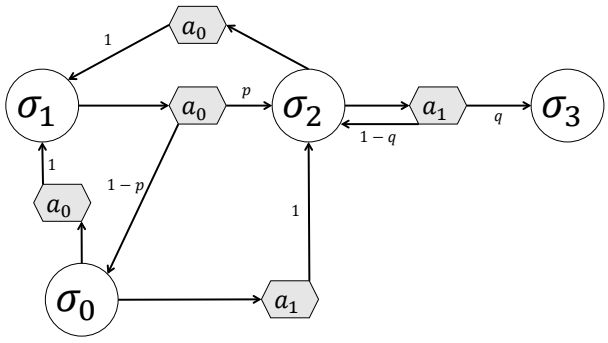

Figure 1: An MDP.