National University of Singapore
School of Computing
CS3243 Introduction to AI

Tutorial 6: Reinforcement Learning

Issued: February 27, 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 $0, \dots, t-1$ 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 $\langle state \rangle \to reward$ below

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

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?

(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?

2. Consider the MDP described in Figure 1

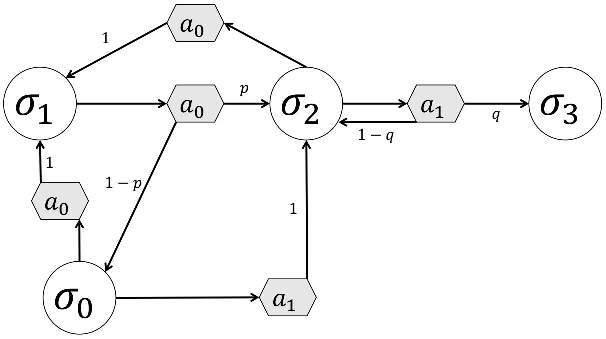

Figure 1: An MDP.

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?

(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)

(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 $0 < \gamma < 1$).

(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.