A3Q2: Infinite MDP

Bodhi Nguyen

January 21st, 2021

1 Problem

Consider an MDP with an infinite set of states $S = \{1, 2, 3, ...\}$. The start state is $s = 1$. Each state s allows a continuous set of actions $a \in [0, 1]$. The transition probabilities are given by:

$$P(s+1|s,a) = a, P(s|s,a) = 1 - a, \ \forall s \in S, \ \forall a \in [0,1]$$

For all state and actions, transitioning from s to $s+1$ results in a reward of $1-a$ and transitioning from s to s results in a reward of $1+a$. The discount factor $\gamma = 0.5$

2 Optimal Value Function

The optimal value function is described as follows (a little bit different notation from the book):

$$V^*(s) = \max_{a \in A} E[r_{t+1} + \gamma * V^*(s_{t+1})|s_t = s, a_t = a] \tag{1}$$

Filling in the information from the problem, we have two cases which simplifies the expected value: first that the process stays in place, and second that the process moves ahead by one. So the optimal value equation becomes

$$V^*(s) = \max_{a \in A}(a * (1-a) + (1-a)(1+a) + 0.5((1-a)V^*(s) + aV^*(s+1))) \tag{2}$$

We can solve this directly by noting since there are an infinite number of identical states, and the reward only depends on the transition itself (not state number), the optimal value function should be the same for any state, that is, $V^*(s) = V^*(s')$ for any s, s'. Thus we get the following:

$$V^*(s) = \max_{a \in A}(-2a^2 + a + 1 + 0.5((1-a)V^*(s) + aV^*(s))) \tag{3}$$

which is equivalent to

$$V^*(s) = \max_{a \in A}(\frac{-2a^2 + a + 1}{0.5}) \tag{4}$$

3 Optimal Deterministic Policy

An optimal deterministic policy is a deterministic policy whose value function is greater than or equal to any other deterministic policy's value function for all states.

We can take our previous analytical expression for the optimal value function and maximize it directly with respect to a:

$$\frac{dV^(s)}{da} = -8a + 2 \tag{5}$$

which implies that $a = 0.25$ is a maximum, since the second derivative is less than 0. Thus the deterministic optimal policy is to choose $a = 0.25$ each time.