MS&E 346 Assignment 3

Junting Duan

January 21, 2022

1

For a deterministic policy π_D, the value function for an MDP evaluated with π_D can be calculated as

$$\begin{aligned}V^{\pi_D}(s) &= \mathbb{E}_{\pi_D,P_R}[G_t|S_t = s]\\ &= R^{\pi_D}(s) + \gamma \cdot \sum_{s' \in \mathcal{N}} P^{\pi_D}(s, s') \cdot V^{\pi_D}(s')\\ &= R(s, \pi_D(s)) + \gamma \cdot \sum_{s' \in \mathcal{N}} P(s, \pi_D(s), s') \cdot V^{\pi_D}(s'),\end{aligned}$$

and the action-value function of an MDP evaluated with π_D is

$$\begin{aligned}Q^{\pi_D}(s, \pi_D(s)) &= \mathbb{E}_{\pi_D,P_R}[G_t|(S_t = s, A_t = \pi_D(s))]\\ &= R(s, \pi_D(s)) + \gamma \cdot \sum_{s' \in \mathcal{N}} P(s, \pi_D(s), s') \cdot V^{\pi_D}(s').\end{aligned}$$

Thus, we can see that

$$V^{\pi_D}(s) = Q^{\pi_D}(s, \pi_D(s)),$$

which leads to the equation

$$Q^{\pi_D}(s, \pi_D(s)) = R(s, \pi_D(s)) + \gamma \cdot \sum_{s' \in \mathcal{N}} P(s, \pi_D(s), s') \cdot Q^{\pi_D}(s', \pi_D(s')).$$

2

MDP Bellman Optimality Equation can be calculated as

$$\begin{aligned}V^*(s) &= \max_{a \in \mathcal{A}} \{R(s, a) + \gamma \cdot \sum_{s' \in \mathcal{N}} P(s, a, s') \cdot V^*(s')\}\\ &= \max_{a \in [0,1]} \{a(1 - a) + (1 - a)(1 + a) + \gamma \cdot (a \cdot V^*(s + 1) + (1 - a) \cdot V^*(s))\}.\end{aligned}$$

Since for all states s, we have the same transition mechanism, there should be $V^*(s) = V^*(s + 1)$. As a result, we have

$$V^*(s) = \max_{a \in [0,1]} \{a(1 - a) + (1 - a)(1 + a) + \gamma \cdot V^*(s)\} = \max_{a \in [0,1]} \{-2a^2 + a + 1\} + \gamma \cdot V^*(s) = \frac{9}{8} + \frac{1}{2}V^*(s),$$

which implies that $V^*(s) = 9/4$. Additionally, the optimal deterministic policy is

$$\pi^*(s) = \arg\max_{a \in [0,1]} \{a(1 - a) + (1 - a)(1 + a) + \gamma \cdot \sum_{s' \in \mathcal{N}} P(s, a, s') \cdot V^*(s')\} = \frac{1}{4}.$$

3

The state space is $\mathcal{S} = \{0, 1, \cdots, n\}$ with non-terminal state space $\mathcal{N} = \{1, \cdots, n-1\}$ and terminal state space $\mathcal{T} = \{0, n\}$. The action space of this problem is $\mathcal{A} = \{A, B\}$. The transition function for any $s \in \mathcal{N}$ is

$$P(s, A, s') = \begin{cases} \frac{n-i}{n} & \text{if } s' = s+1, \\ \frac{i}{n} & \text{if } s' = s-1, \\ 0 & \quad \text{otherwise} \end{cases}$$

and

$$P(s, B, s') = \frac{1}{n} \quad \text{for } s' \neq s.$$

The reward function of any implied MRP is $R^\pi(0) = -1$, $R^\pi(n) = 1$, and $R^\pi(s) = 0$ for $s \in \mathcal{N}$.

We model this MDP as an instance of the FiniteMarkovDecisionProcess class. The code and graphs are see in the code "Assignment 3.ipynb". From the results on the graph with $n = 3, 6, 9$, we observe that the optimal policy is always "croak B on lilypad 1 and croak A on the other lilypads".

4

Let $\pi(s, a)$ denotes the continuous policy function, then we should have $\int_{\mathbb{R}} \pi(s, a) da = 1$ for any $s \in \mathbb{R}$. Furthermore, the value function V^π can be calculated as

$$V^\pi(s) = \int_{\mathbb{R}} \pi(s, a) R(s, a) da + \gamma \cdot \int_{\mathbb{R}} \int_{\mathbb{R}} p(s, a, s') \cdot V^\pi(s') da ds'.$$

Based on this question, we have

$$R(s, a) = \int_{\mathbb{R}} p(s, a, s') \cdot \exp(as') ds',$$

in which

$$p(s, a, s') = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{(s'-s)^2}{2\sigma^2} \Big).$$

Therefore, for the special case of $\gamma = 0$, the optimal value function can be written as

$$\begin{aligned} V^*(s) &= \min_{a \in \mathbb{R}} \int_{\mathbb{R}} \pi(s, a) \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\Big(as' - \frac{(s'-s)^2}{2\sigma^2} \Big) ds' da \\ &= \min_{a \in \mathbb{R}} \int_{\mathbb{R}} \pi(s, a) \cdot \exp\Big(\frac{2sa + \sigma^2 a^2}{2} \Big) da. \end{aligned}$$

Since $\frac{1}{2}\sigma^2 a^2 + sa$ achieves minimum when $a = -s/\sigma^2$, we deduce that the optimal action is

$$\pi^*(s) = -\frac{s}{\sigma^2},$$

and the corresponding optimal cost is

$$V^*(s) = \exp\big(- \frac{s^2}{2\sigma^2} \big).$$