MS&E 346 Assignment 3

Constantin Eulenstein

January 2022

1

For a deterministic policy, value function in terms of value function:

$$\begin{aligned} V^\pi(s) &= \mathcal{R}^\pi(s) + \gamma \cdot \sum_{s' \in \mathcal{N}} \mathcal{P}^\pi\left(s, s'\right) \cdot V^\pi\left(s'\right) \\ &= \pi(s) \cdot \mathcal{R}(s) + \gamma \cdot \pi(s) \sum_{s' \in \mathcal{N}} \mathcal{P}\left(s, s'\right) \cdot V^\pi\left(s'\right) \\ &= \pi(s) \cdot \left(\mathcal{R}(s) + \gamma \cdot \sum_{s' \in \mathcal{N}} \mathcal{P}\left(s, s'\right) \cdot V^\pi\left(s'\right)\right) \\ &= \left(\mathcal{R}(s) + \gamma \cdot \sum_{s' \in \mathcal{N}} \mathcal{P}\left(s, s'\right) \cdot V^\pi\left(s'\right)\right) \text{ for all } s \in \mathcal{N} \end{aligned} \tag{1}$$

For a deterministic policy, value function in terms of action-value function:

$$V^\pi(s) = Q^\pi(s, a) \text{ for all } s \in \mathcal{N} \tag{2}$$

Note: there is a deterministic a for each s.

For a deterministic policy, action-value function in terms of value function:

$$Q^\pi(s, a) = V^\pi(s) \text{ for all } s \in \mathcal{N} \tag{3}$$

Note: there is a deterministic a for each s.

For a deterministic policy, action-value function in terms of action-value function:

$$Q^\pi(s, a) = \left(\mathcal{R}(s, a) + \gamma \cdot \sum_{s' \in \mathcal{N}} \mathcal{P}\left(s, s'\right) \cdot Q^\pi\left(s', a\right)\right) \text{ for all } s \in \mathcal{N} \tag{4}$$

2

For every state s, both the probability of transitioning and the rewards stay the same. Therefore, $R(s, a) = R(a)$.

The expected reward per move is:

$$\mathbb{E}R(a) = Pr(s+1 \mid s) \cdot R(s+1 \mid s) + Pr(s \mid s) \cdot R(s \mid s) = a(1-a) + (1-a)(1+a) = -2a^2 + a + 1$$

This is maximized at $a = 0.25$.

Consequently, the optimal policy is to always choose a equal to 0.25. The expected reward will be 1.125

As a result, the optimal Value function $V(s)$, for all s, is equal to:

$$V^*(s) = 1.25 + 0.5 \cdot 1.25 + 0.5^2 \cdot 1.25 \cdots = 1.25 \frac{1}{1-0.5} = 2.5$$

3

See code.

4

In the myopic case, the expected discounted sum of costs $\mathbb{E}[G_t] = \mathbb{E}[R_{t+1}]$. Let's denote that $g(s') = \exp(as')$ which is a log normal distribution, which has mean $\exp(as + a^2\sigma^2/2)$. Therefore,

$$\mathbb{E}[G_t] = \mathbb{E}[g(s')] = \exp(as + a^2\sigma^2/2)$$

Taking the derivative gives us:

$$\exp(as + a^2\sigma^2/2)(s + a\sigma^2)$$

Setting this equal to 0, yeilds an optimal action of $a = -\frac{s}{\sigma^2}$