CME241 Assignment3

Quinn Hollister

January 2022

1 Problem 1

For a Deterministic Policy, write with precise mathematical notation the 4 MDP Bellman Policy Equations

$$V^{\pi_D}(s) = \mathcal{R}(s, \pi_D(s)) + \gamma \cdot \sum_{s' \in \mathcal{N}} (\mathcal{P}(s, \pi_D(s), s') \cdot V^{\pi_D}(s') \tag{1}$$

$$V^{\pi_D}(s) = Q^{\pi_D}(s, \pi_D(s)) \tag{2}$$

$$Q^{\pi_D}(s, \pi_D(s)) = \mathcal{R}(s, \pi_D(s)) + \gamma \cdot \sum_{s' \in \mathcal{N}} \mathcal{P}(s, \pi_D(s), s') \cdot V^{\pi_D}(s') \tag{3}$$

$$Q^{\pi_D}(s, \pi_D(s)) = \mathcal{R}(s, \pi_D(s)) + \gamma \cdot \sum_{s' \in \mathcal{N}} P(s, \pi_D(s), s') \cdot Q^{\pi_D}(s', \pi_D(s')) \tag{4}$$

2 Problem 2

Consider an MDP with infinite states. Use the MDP Bellman Optimality Equation to calculate $V^*(s)$

Setup: S = 1,2,3,..., with our start state as s = 1, and s allows a continuous set of actions $a \in [0, 1]$. The transition probabilities are given by :

$$P[s+1|s, a] = a, P[s|s, a] = 1 - a$$

The reward for the first transition is 1-a, and 1+a for the second transition. We'll also use a gamma of 0.5. To calculate the optimal value function, we just need to use the MDP bellman optimality equation, stated here:

$$V^*(s) = \max_{a \in \mathcal{A}} \{R(s, a) + \gamma \cdot \sum_{s' \in \mathcal{N}} P(s, a, s') \cdot V^*(s')\} \tag{5}$$

$$V^*(s) = \max_{a \in \mathcal{A}} \{R(s, a) + \gamma (P(s, a, s+1) \cdot V^*(s+1) + P(s, a, s) * V^*(s)\} \tag{6}$$

But, we can notice that the value function itself doesn't depend on the state at all. We can see this by remembering that the value function is defined as the future expected rewards given we are in our current state. But, each individual reward is only a function of the action we take between time steps, so $V^*(s) = V^*(s'), \forall s, s' \in \mathcal{S}$. Replacing $V^*(s+1) = V^*(s)$ into equation 2, and putting in our values for each quantity gives us:

$$V^*(s) = \max_{a \in \mathcal{A}} \{(1-a)*(a) + (1+a)(1-a) + \gamma \cdot (a * V^*(s) + (1-a) * V^*(s))\} \quad (7)$$

$$V^*(s) = \max_{a \in \mathcal{A}} \{(1-a)*(a) + (1+a)(1-a) + \gamma * V^*(s) \cdot (a+1-a)\} \quad (8)$$

And we can now take out of the max function any expression that's become decoupled from the choice of a:

$$V^*(s) = \max_{a \in \mathcal{A}} \{(1-a)*(1+2a)\} + \gamma * V^*(s) \quad (9)$$

The maximum of $(1-a)*(1+2a)$ over the range $[0,1]$ occurs at $a = \frac{1}{4}$ and so the value is $\frac{9}{8}$. Thus, our equation becomes

$$V^*(s) = \frac{1}{2} \cdot V^*(s) + \frac{9}{8} \quad (10)$$

$$V^*(s) = \frac{9}{4} \quad (11)$$

In order to get the optimal policy, we remember that two equations:

$$V^*(s) = Q^*(s, \pi_D^*(s)) \quad (12)$$

$$\pi_D^*(s) = arg\max\{Q^*(s,a)\} \forall s \in \mathcal{N} \quad (13)$$

Thus, we remember that we attain our optimal value function when $a = \frac{1}{4}$, so that must be the argmax, and obviously the policy doesn't depend on the state as the value function itself doesn't depend on the state, so our optimal policy is just to take our action a, such that $a = \frac{1}{4}$

$$\pi_D^*(s) = \frac{1}{4}$$

3 Problem 3

Express the state space, action space, transition function, and rewards function of the MDP for the frog-escape problem

$$S = \{0, 1, 2, ..., n\}, = \{A, B\}, \mathcal{T} = \{0, n\}$$

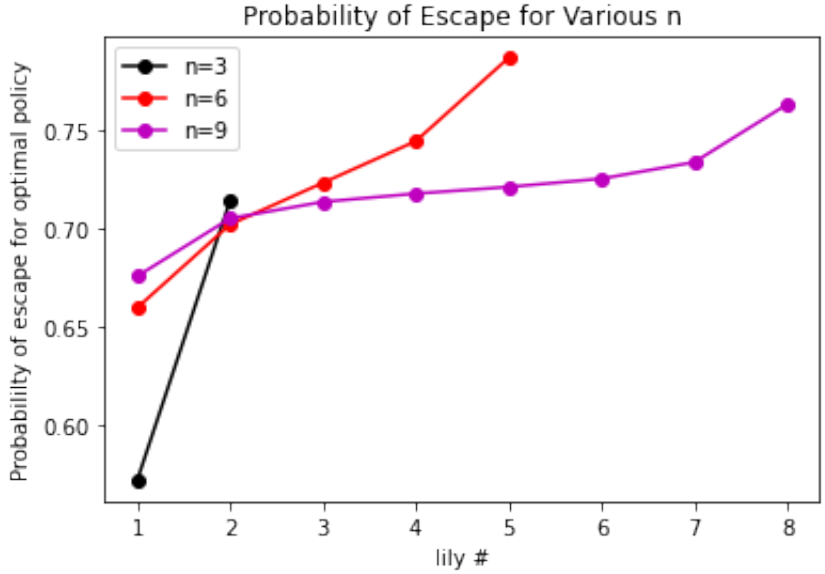
$$P(i, A, i+1) = \frac{n-i}{n}, P(i, A, i-1) = \frac{i}{n}$$

$$P(i, B, s') = \frac{1}{n}, \forall s' \in S \setminus \{i\}$$

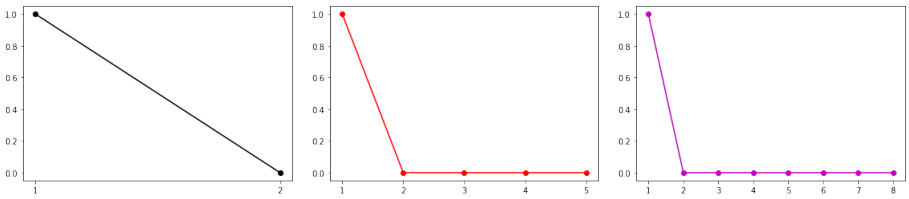
$$R(s) = I_n(s)$$

where $I_n(s)$ is an indicator function for s, i.e. it's zero for all states unless that state is n, the terminal state.

The optimal value function for our problem for each of the 3 required lily sizes are shown below:



The optimal policy for each of the sizes are also shown below for n =3, 6, 9:



The optimal policy for each of these different sizes is roughly equivalent: if you're at lily number 1, choose option B, otherwise choose option A on any other lilypad number. This makes some intuitive sense, since the odds that you land on lily number 0–and are thus eaten by the snake–are equivalent to between option A and B if you're at lily number 1. This is the only time where the benefits of option B outweigh the consequences since option B does allow you the chance to jump over multiple lily's, but usually the chance that you jump straight to the snake outweighs these benefits. But, like I said, now this

very negative outcome has an equivalent outcome when at lily 1, so we would like to have this upside.

4 Problem 4

The problem is to minimize the infinite-horizon Expected Discounted-Sum of Costs, where $\gamma = 0$

We start as always with the MDP Bellman Optimality Equation:

$$V^*(s) = \max_{a \in \mathcal{A}} \{R(s,a,s') + \gamma \cdot \int_{-\infty}^{\infty} P(s,a,s') \cdot V^*(s') ds'\} \tag{14}$$

But, we know that $\gamma = 0$, and the future rewards can be factored to:

$$V^*(s) = \max_{a \in \mathcal{A}} \{- \int_{-\infty}^{\infty} P(s,a,s') \cdot e^{\alpha * s'} ds'\} \tag{15}$$

$$V^*(s) = \max_{a \in \mathcal{A}} \{- \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{s_i - s}{\sigma})^2} e^{\alpha * s'} ds'\} \tag{16}$$

Which is just a simple optimization problem, i.e. find the maximum value of the function by taking a first derivative with respect to a. But, we should first simplify the expression by evaluating the integral over s'. It's easy to use an integral table for this definite integral, and we can find an analytical expression for the integral over the real space of the exponential function with a quadratic function as its argument.

$$\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2*\sigma^2}(s'^2 - s' \cdot (2s + 2 \cdot \sigma^2 \cdot a) + s^2} ds' \tag{17}$$

$$= -\frac{1}{\sigma\sqrt{2 \cdot \pi}} \cdot \sqrt{2\sigma^2\pi} e^{\frac{(s+\sigma^2 a)^2}{\sigma^4} \cdot \frac{\sigma^2}{2} - \frac{s^2}{2\sigma^2}} \tag{18}$$

$$= -e^{\frac{(s+\sigma^2 a)^2}{2\sigma^2} - \frac{s^2}{2\sigma^2}} \tag{19}$$

$$= -e^{\frac{2sa + \sigma^2 a^2}{2}} \tag{20}$$

So, now our expression for V^* becomes:

$$V^*(s) = \max_{a \in \mathcal{A}} \{-e^{sa + \frac{1}{2}\sigma^2 a^2}\} \tag{21}$$

And taking the derivative of the expression with respect to a, and setting it equal to zero to find the optimal a gives us:

$$\frac{\partial}{\partial a}(-e^{sa + \frac{1}{2}(\sigma a)^2}) = 0 \tag{22}$$

$$-(s + \sigma^2 a) \cdot e^{sa + \frac{1}{2}(\sigma \cdot a)^2} = 0 \tag{23}$$

So, either $a = \frac{-s}{\sigma^2}$ or the exponential must be zero, which is not true for any action a. Thus, the optimal action $\forall$ s $\in \mathcal{S}$ is $a = \frac{-s}{\sigma^2}$, and the cost associated with the action is $-e^{\frac{s \cdot s'}{\sigma^2}}$