Assignment 3

Miao-Chin Yen

March 7, 2022

Problem 1

For a deterministic Policy, $\pi_D : \mathcal{S} \rightarrow \mathcal{A}$, i.e., $\pi_D(s) = a$, where $s \in \mathcal{S}, a \in \mathcal{A}$.
MDP(State-Value Function) Bellman Policy Equation $V^{\pi_D} : \mathcal{N} \rightarrow \mathbb{R}$:

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

Action-Value Function (for policy π_D) $Q^{\pi_D} : \mathcal{N} \times \mathcal{A} \rightarrow \mathbb{R}$:

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

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

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

Problem 2

MDP State-Value Function Bellman Optimality Equation:

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

In this problem, $\mathcal{A} = [0, 1]$ and $\gamma = 0.5$:

$$V^*(s) = \max_{a \in [0,1]} \left\{ \mathcal{R}(s, a) + 0.5 \cdot \sum_{s' \in \mathcal{N}} \mathcal{P}\left(s, a, s'\right) \cdot V^*\left(s'\right) \right\}$$

and we write explicitly:

$$V^*(s) = \max_{a \in [0,1]} \left\{ a(1-a) + (1-a)(1+a) + 0.5 \cdot [V^*(s+1) \cdot a + V^*(s) \cdot (1-a)] \right\}$$

Notice that $\mathcal{R}(s, a)$ does not depend on s. Hence, $V^*(s) = V^*(s+1)$. Therefore,

$$V^*(s) = \max_{a \in [0,1]} \left\{ a(1-a) + (1-a)(1+a) + 0.5 \cdot V^*(s+1) \right\} \Longrightarrow a = 0.25$$

$$V^*(s) = 1.125 + 0.5 \cdot V^*(s+1)$$

and the optimal deterministic policy $\pi_D(s) = 0.25 \; \forall s \in \mathcal{S}$

Problem 3

Let the state space $\mathcal{S} = \{s \mid 0 \leq s \leq n\}$. State s means that the frog is sitting on lilypad numbered s. Terminal state space: $\mathcal{T} = \{0, n\}$. The action space $\mathcal{A} = \{A, B\}$ which stands for the two choices of croak sounds. The state transitions are as follows:

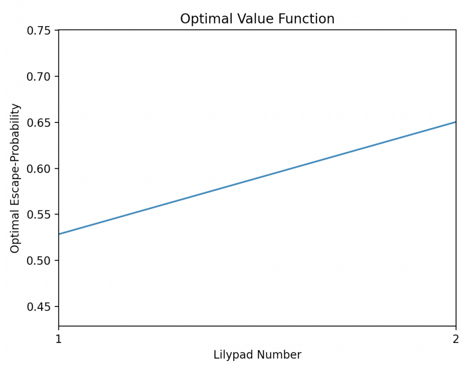
$$\mathcal{P}(s, A, s') = \mathbb{P}[S_{t+1} = s' | S_t = s, A_t = A] \text{ for } 1 \leq s \leq n-1 = \begin{cases} \frac{s}{n} & \text{for } s' = s-1 \\ \frac{n-s}{n} & \text{for } s' = s+1 \\ 0 & \text{otherwise} \end{cases}$$

$$\mathcal{P}(s, B, s') = \mathbb{P}[S_{t+1} = s' | S_t = s, A_t = B] \text{ for } 1 \leq s \leq n-1 = \begin{cases} \frac{1}{n} & \text{for all } 0 \leq s' \leq n \text{ and } s' \neq s \\ 0 & \text{for } s' = s \end{cases}$$

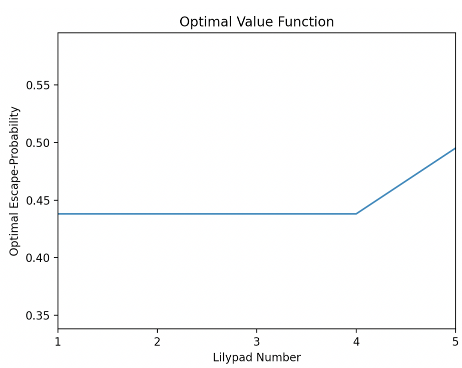
The reward function $R(s, a, s')$ is as follows:

$$R(s, a, s') \text{ for } 1 \leq s \leq n-1, a \in \{A, B\} = \begin{cases} 1 & \text{for } s' = n \\ 0 & \text{otherwise} \end{cases}$$

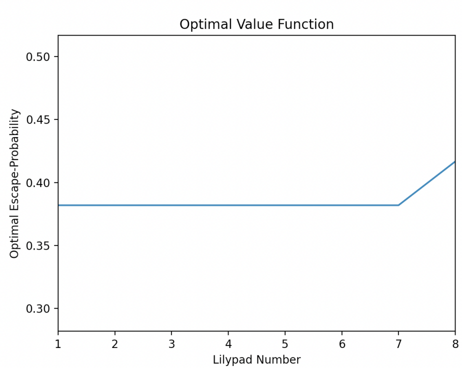
Let discount factor to be 0.9. We have the following graphs.
n=3:



n=6:



n=9:



We found that for $1 \leq s \leq n-2$, the frog should croak B. For $s = n-1$, it should croak A. Reference croaking_on_lilypads_mdp.py.

Problem 4

MDP State-Value Function Bellman Optimality Equation:

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

Consider the myopic case ($\gamma = 0$) and $S' \sim \mathcal{N}(s, \sigma^2)$:

$$V^*(s) = \max_{a \in \mathcal{A}} \mathcal{R}(s,a) = \max_{a \in \mathcal{A}} \sum_{s' \in \mathbb{R}} \mathcal{P}_{S'}\left(s'\right) \cdot [-e^{as'}] = \max_{a \in \mathcal{A}} \mathbb{E}[-e^{as'}] = \min_{a \in \mathcal{A}} M_{S'}(a)$$

$$M_{S'}(a) = e^{sa + \frac{\sigma^2 a^2}{2}}$$

To find a which maximizes the moment generating function, we take derivative w.r.t. a.

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

Hence, the optimal action a^* for state s is $\frac{-s}{\sigma^2}$ and the corresponding optimal cost is $V^*(s) = e^{\frac{-s^2}{\sigma^2} + \frac{s^2}{2\sigma^2}}$