Let us write the 4 MDP Bellman Policy Equations for a deterministic policy. A deterministic policy $\pi_D : S \to A$ is such that:

$\pi(s, \pi_D(s)) = 1$ and $\pi(s, a) = 0$ if $a \neq \pi_D(s)$

By replacing in the formula of the Bellman Policy Equations, we get that under this deterministic policy:

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

As $\pi(s, \pi_D(s)) = 1$, we get:

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

Besides:

$$V^\pi(s) = Q^\pi(s, \pi_D(s))$$

And we also get:

$$Q^\pi(s, a) = R(s, a) + \gamma \sum_{s' \in N} P(s, a, s')V^\pi(s')$$

The last MDP Bellman Policy Equation is:

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

CME 241: Assignment 3: Problem 2

Using the MDP Bellman optimality Equation, let us find $V^*(s)$ for all $s \in S$.

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

And $R(s,a) = \mathbb{E}(R_{t+1} | (S_t = s, A_t = a))$. In this example $R(s,a) = a(1-a) + (1-a)(1+a)$

And $\sum_{s' \in N} P(s,a,s') V^*(s') = a V^*(s+1) + (1-a) V^*(s)$

Hence $V^*(s) = \max_{a \in A} \left\{ a(1-a) + (1-a)(1+a) + \gamma [a V^*(s+1) + (1-a) V^*(s)] \right\}$

$= \max_{a \in A} \left\{ -a^2 + a + 1 - a^2 + a[\gamma V^*(s+1) - \gamma V^*(s)] + \gamma V^*(s) \right\}$

This gives us: $V^*(s) = \max_{a \in A} \left\{ -2a^2 + a[1 + \gamma V^*(s+1) - \gamma V^*(s)] + 1 + \gamma V^*(s) \right\}$

The maximum of the function we try to maximize is reached when its derivative is equal to 0.

Hence when: $-4a + 1 + \gamma V^*(s+1) - \gamma V^*(s) = 0$

$$a = \frac{1 + \gamma V^*(s+1) - \gamma V^*(s)}{4}$$

We get: $V^*(s) = -2 \left(\frac{1 + \gamma V^*(s+1) - \gamma V^*(s)}{4} \right)^2 + \frac{(1 + \gamma V^*(s+1) - \gamma V^*(s))^2}{4} + 1 + \gamma V^*(s)$

This leads to: $\frac{(1 + \gamma V^*(s+1) - \gamma V^*(s))^2}{8} + 1 = (1-\gamma) V^*(s)$

For $\gamma = \frac{1}{2}$, we get: $\frac{1 + V^*(s+1) - V^*(s) + \frac{1}{4}(V^*(s+1) - V^*(s))^2}{8} + 1 = \frac{1}{2} V^*(s)$

This gives us: $8 + 1 + V^*(s+1) - V^*(s) + \frac{1}{4}(V^*(s+1) - V^*(s))^2 = 4 V^*(s)$

$$5 V^*(s) = 9 + V^*(s+1) + \frac{1}{4}(V^*(s+1) - V^*(s))^2$$

Note here that the structure of the rewards given are the same regardless of the start state. To maximize the expected sum of rewards, and get the optimal value function, we could simply find the deterministic policy that maximizes the expected rewards at a given step.

Or, like we will do here, we could see that: $\forall s \in S \; V^*(s) = V^*(s+1)$ since the rewards do not depend on the state we're in.

This gives us that $V^*(s) = \frac{9}{4}$, and the optimal deterministic policy π^* is such that $\forall s \in S$: $\pi^*(s) = \frac{1 + \gamma V^*(s+1) - \gamma V^*(s)}{4} = \frac{1}{4}$

Let us define the state space, action space, transitions function and rewards function to solve this problem using MDP theory.

Here the state space is $S = \{s|0 \leq s \leq n\}$, and each state represents the lilypad on which the frog is sitting. Note that the states $s = 0$ and $s = n$ are terminating states/

The frog has the choice between two different sounds when on a lilypad. Hence, the action space is $\{A, B\}$.

We can now compute the state transitions probabilities:

$\mathbb{P}(s-1|(s,A)) = \frac{s}{n}$, $\mathbb{P}(s+1|(s,A)) = \frac{n-s}{n}$ and $\mathbb{P}(s'|(s,A)) = 0$ if $s' \neq s-1$ and $s' \neq i+1$.

Besides:

$\mathbb{P}(s'|(s,B)) = \frac{1}{n}$ for all $s' \neq s$ and $0 \leq s' \leq n$.

Note that these equalities hold for $1 \leq s \leq n-1$.

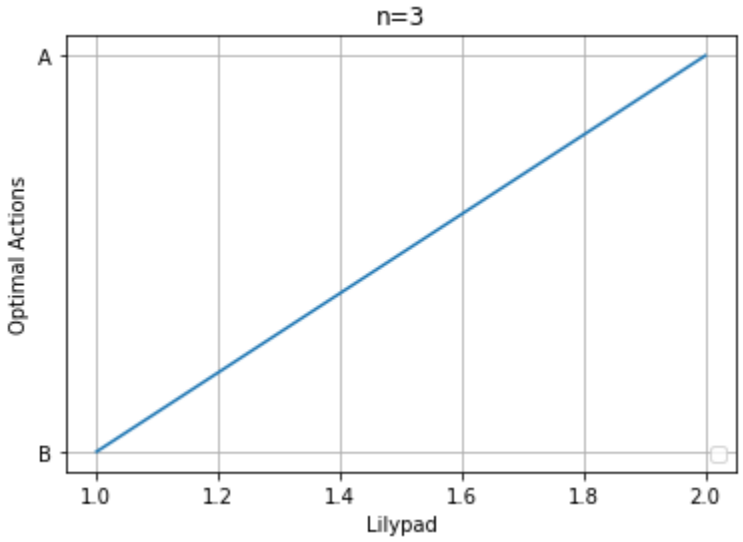
Let us now define the rewards function for this frog-escape problem. The goal of the frog is to escape the pond, it will only succeed when it lands on lilypad n.

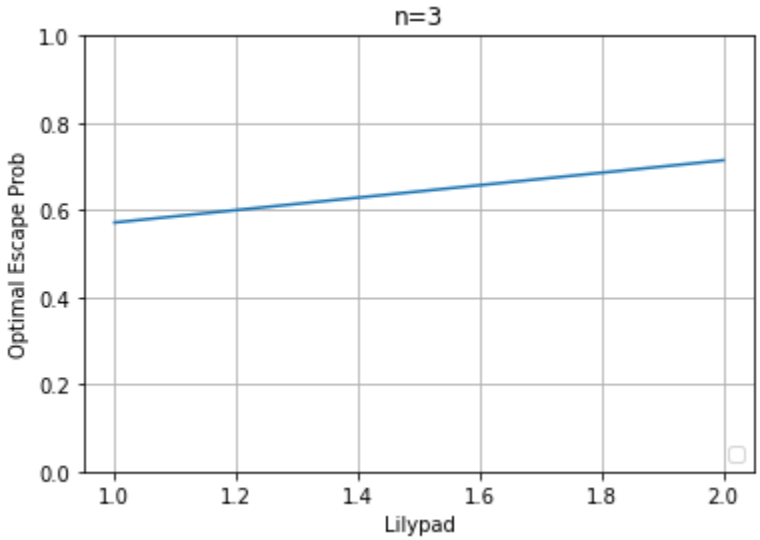
To this extent, the reward when transitioning from a state s to a state s' after taking an action a is for $1 \leq s \leq n-1$ and $a \in \{A, B\}$:

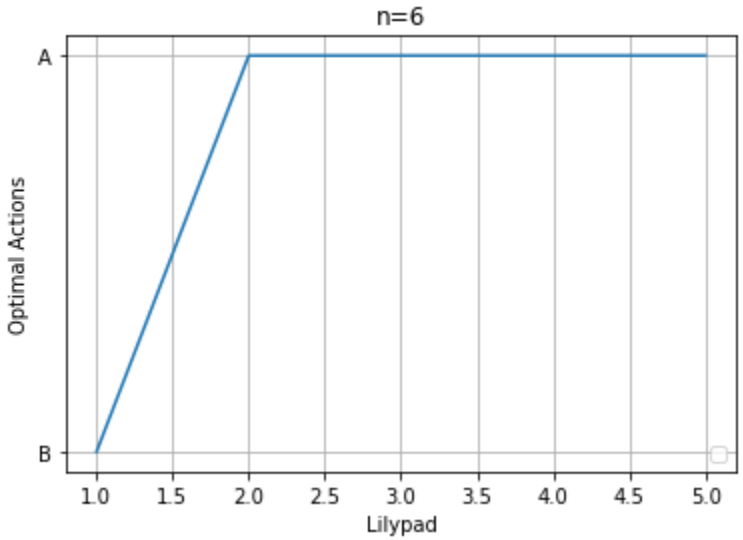
$R(s,a,s') = 1$ if $s' = n$ and $R(s,a,s') = 0$ otherwise.

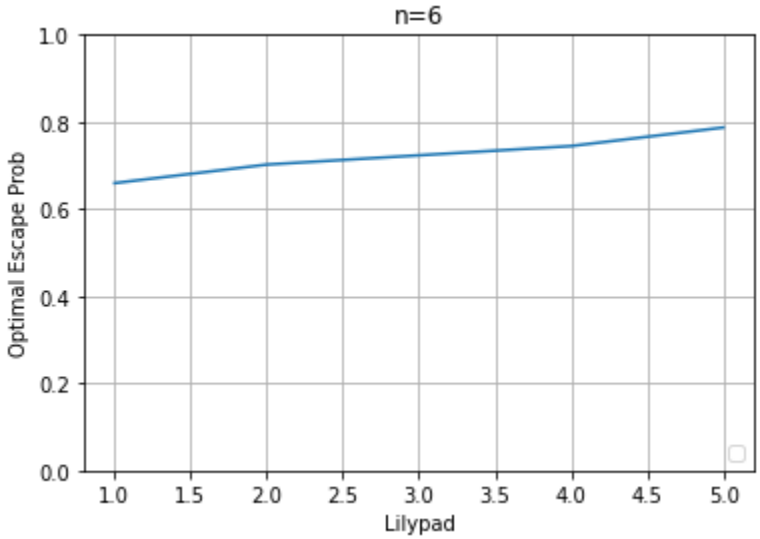
We then coded everything in the `assignment3_code.py` file.

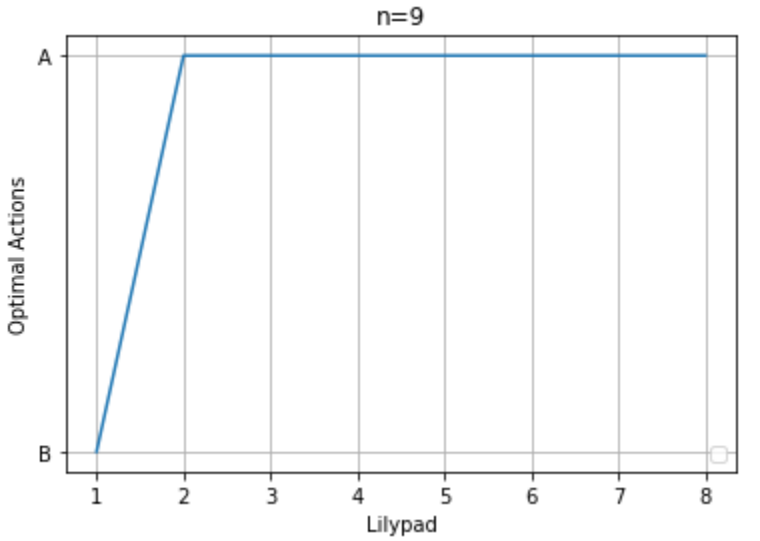
This gave us the following plots:

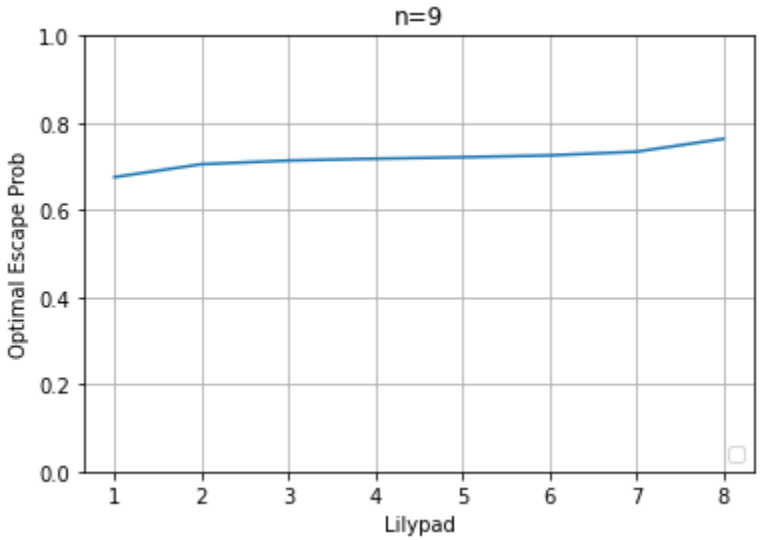



n=3
Optimal Escape Prob
1.0
0.8
0.6
0.4
0.2
0.0
1.0
1.2
1.4
1.6
1.8
2.0
Lilypad


n=6
Optimal Actions
A
B
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Lilypad


n=6
Optimal Escape Prob
1.0
0.8
0.6
0.4
0.2
0.0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Lilypad





Note that the results vary depending on the value of γ chosen. Here not discounting and choosing $\gamma = 1$ like we did was making sense.

The trend we observe when we vary n is the same. For $k = 1$, the optimal action is action B and for $2 \leq k \leq n - 1$ the optimal action is A.

CME 241 : Assignment 3 : Problem 4:

Let us derive an analytic expression for the optimal action in any state.

Let us solve the MDP State-Value Function Bellman Optimality Equation.

$$V^*(s) = \max_{a \in \mathbb{R}} \left\{ R(s,a) + \gamma \int_{s' \in \mathbb{R}} P(s,a,s') \, V^*(s') \, ds' \right\}$$

In the case where $\gamma = 0$, we are just trying to solve:

$$V^*(s) = \max_{a \in \mathbb{R}} \{R(s,a)\} \text{ and } R(s,a) = E(R_{t+1} \mid S_t = s, A_t = a)$$

Let us define the rewards as equal to minus the cost.

Hence $R(s,a) = -E(e^{as'} \mid S_t = s, A_t = a)$: as $s' \sim N(s, \sigma^2)$ we recognize the moment generating function of s'.

Hence $R(s,a) = -\exp\left(sa + \frac{\sigma^2 a^2}{2}\right)$

$\frac{\partial R(s,a)}{\partial a} = -\left(s + \frac{2\sigma^2}{2}\right) \exp\left(sa + \frac{\sigma^2 a^2}{2}\right)$: this is equal to 0 when $a = -\frac{s}{\sigma^2}$

This is our expression for the optimal action in any state $s \in \mathbb{R}$

The corresponding optimal cost is $e^{-\frac{s^2}{\sigma^2}}$