Stanford CME 241 (Winter 2021) - Assignment 3

Assignments:

1. For a *Deterministic* Policy $\pi_D : \mathcal{S} \rightarrow \mathcal{A}$, write with precise mathematical notation the 4 MDP Bellman Policy Equations, i.e., V^{π_D} in terms of Q^{π_D}, Q^{π_D} in terms of V^{π_D}, V^{π_D} in terms of V^{π_D}, Q^{π_D} in terms of Q^{π_D}. Note that in the book, we have written the 4 MDP Policy Equations in terms of the notation for a stochastic policy $\pi : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$.

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

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

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

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

2. Consider an MDP with an infinite set of states $\mathcal{S} = \{1, 2, 3, \ldots\}$. The start state is $s = 1$. Each state s allows a continuous set of actions $a \in [0, 1]$. The transition probabilities are given by:

$$\mathbb{P}[s+1 \mid s, a] = a, \mathbb{P}[s \mid s, a] = 1 - a \text{ for all } s \in \mathcal{S} \text{ for all } a \in [0, 1]$$

For all states $s \in \mathcal{S}$ and actions $a \in [0, 1]$, transitioning from s to $s+1$ results in a reward of $1 - a$ and transitioning from s to s results in a reward of $1 + a$. The discount factor $\gamma = 0.5$.

- Using the MDP Bellman Optimality Equation, calculate the Optimal Value Function $V^*(s)$ for all $s \in \mathcal{S}$
- Calculate an Optimal Deterministic Policy $\pi^*(s)$ for all $s \in \mathcal{S}$

$$P = a \Rightarrow R = 1 - a$$
$$P = 1 - a \Rightarrow R = 1 + a$$

$$V^*(s) = a(1-a) + (1-a)(1+a) + \gamma a V(s) + \gamma(1-a)V(s)$$

$\gamma = 0.5$, then optimizing over V^* yields

$$\frac{dV^*}{da} = 1 - 2a - 2a = 0$$

$\nearrow \pi(s^*)$

$a = \frac{1}{4}$, this is the optimal policy for all states.

$V^*(s) = \frac{9}{4}$ for all states.

3. Consider an array of $n+1$ lilypads on a pond, numbered 0 to n. A frog sits on a lilypad other than the lilypads numbered 0 or n. When on lilypad i ($1 \le i \le n-1$), the frog can croak one of two sounds A or B. If it croaks A when on lilypad i ($1 \le i \le n-1$), it is thrown to lilypad $i-1$ with probability $\frac{i}{n}$ and is thrown to lilypad $i+1$ with probability $\frac{n-i}{n}$. If it croaks B when on lilypad i ($1 \le i \le n-1$), it is thrown to one of the lilypads $0, \ldots, i-1, i+1, \ldots n$ with uniform probability $\frac{1}{n}$. A snake, perched on lilypad 0, will eat the frog if the frog lands on lilypad 0. The frog can escape the pond (and hence, escape the snake!) if it lands on lilypad n.

   What should the frog croak when on each of the lilypads $1, 2, \ldots, n-1$, in order to maximize the probability of escaping the pond (i.e., reaching lilypad n before reaching lilypad 0)? Although there are more than one ways of solving this problem, we'd like to solve it by modeling it as an MDP and identifying the Optimal Policy.

   - Express with clear mathematical notation the state space, action space, transitions function and rewards function of an MDP so that the above *frog-escape* problem would be solved by arriving at the Optimal Value Function (and hence, the Optimal Policy) of this MDP.

$$S = \{0, 1, 2, \ldots, n\}$$

$$\mathcal{N} = 1, 2, \ldots, n-1$$

$$\mathcal{A} = \{A, B\}$$

$$P(s,a,s') = \begin{cases} \frac{s}{n} & a = A \quad s \in \mathcal{N} \quad s' = s-1 \\ \frac{n-s}{n} & a = A \quad s \in \mathcal{N} \quad s' = s+1 \\ \frac{1}{n} & a = B \quad s \in \mathcal{N} \\ 0 & \text{otherwise} \end{cases}$$

$$R(s,a) = \begin{cases} \frac{n-s}{n} - \frac{s}{n} = \frac{n-2s}{n} & a = A \\ \frac{1}{n}\left(\sum_{i=0}^{n} i\right) = \frac{n+1}{2} & a = B \end{cases}$$

4. Consider a continuous-states, continuous-actions, discrete-time, non-terminating MDP with state space as $\mathbb{R}$ and action space as $\mathbb{R}$. When in state $s \in \mathbb{R}$, upon taking action $a \in \mathbb{R}$, one transitions to next state $s' \in \mathbb{R}$ according to a normal distribution $s' \sim \mathcal{N}(s, \sigma^2)$ for a fixed variance $\sigma^2 \in \mathbb{R}^+$. The corresponding cost associated with this transition is $e^{as'}$, i.e., the cost depends on the action a and the state s' one transitions to. The problem is to minimize the infinite-horizon *Expected Discounted-Sum of Costs* (with discount factor $\gamma < 1$). For this assignment, solve this problem just for the special case of $\gamma = 0$ (i.e., the myopic case) using elementary calculus. Derive an analytic expression for the optimal action in any state and the corresponding optimal cost.

Optimization Problem :

Find optimal a to minimize $\mathbb{E}[e^{as'} \mid s] = V^*$

$$V^*(a,s) = -e^{as + \frac{\sigma^2 a^2}{2}}$$

$$\frac{d}{da} \mathbb{E}[e^{as'} \mid s]$$

$$= -\left[e^{as + \frac{\sigma^2 a^2}{2}} (s + a\sigma^2)\right] \overset{set}{=} 0$$

$$a^* = -\frac{s}{\sigma^2} \quad \text{(optimal action)}$$

$$c^* = -e^{-\frac{s^2}{2\sigma^2}} \quad \text{(optimal cost)}$$