Assignment 3

Joseph Wakim

February 4, 2021

1 Bellman Policy Equations for Deterministic Policy

The *action-value function*, $Q^{\pi}(s, a)$, provides the expected return, G_t, from current state, s, and action, a, for a given policy, π. The function maps the sets of non-terminal states, $\mathcal{N}$, and actions, $\mathcal{A}$, to a real number, $\mathbb{R}$.

$$Q^{\pi}(s, a) \,:\, \mathcal{N} \times \mathcal{A} \longrightarrow \mathbb{R}$$

Meanwhile, the *value function*, $V^{\pi}(s)$, predicts expected return from current state alone, for a given policy. As such, the function maps non-terminal states to a real number.

$$V^{\pi}(s) \,:\, \mathcal{N} \longrightarrow \mathbb{R}$$

The action-value function and the value function are related to one-another for Markov decision processes (MDPs) with discrete time countable states; the value function equivalent to the probability-weighted average action-value function over all actions, as determined by policy:

$$V^{\pi}(s) = \sum_{a \in \mathcal{A}} \pi(s, a) \cdot Q^{\pi}(s, a) \qquad [1.1]$$

The *Bellman policy equation* provides a recursive definition of the value function for a discrete-time, countable state MDP:

$$V^{\pi}(s) = \sum_{a \in \mathcal{A}} \pi(s, a) \cdot \left(\mathcal{R}(s, a) + \gamma \cdot \sum_{s' \in \mathcal{N}} \mathcal{P}(s, a, s') \cdot V^{\pi}(s') \right) \qquad [1.2]$$

where $\mathcal{R}$ is the reward function evaluated for a state and action and $\mathcal{P}$ is the state transition probability function evaluated for the transition from state s to s' given action a. By comparison of equations 1.1 and 1.2, the definition of the value-action function is obtained in terms of the value function as:

$$Q^{\pi}(s, a) = \mathcal{R}(s, a) + \gamma \cdot \sum_{s' \in \mathcal{N}} \mathcal{P}(s, a, s') \cdot V^{\pi}(s') \qquad [1.3]$$

A recursive expression for the value action function is then obtained by substituting $V^{\pi}(s)$ from equation 1.1 into 1.3:

$$Q^{\pi}(s, a) = \mathcal{R}(s, a) + \gamma \cdot \sum_{s' \in \mathcal{N}} \mathcal{P}(s, a, s') \sum_{a' \in \mathcal{A}} \pi(s', a') \cdot Q^{\pi}(s', a') \qquad [1.4]$$

A **deterministic policy** is one which constantly selects a single action. If we denote that action with a'', then a deterministic policy suggests:

$$\pi(s, a) = \begin{cases} 1 & a = a'' \\ 0 & a \neq a'' \end{cases}$$

Substituting this definition into equations 1.1-1.4 gives the four MDP Bellman policy equations for a deterministic policy:

$$V^{\pi}(s) = Q^{\pi}(s, a'') \tag{1.5}$$

$$V^{\pi}(s) = \mathcal{R}(s, a'') + \gamma \cdot \sum_{s' \in \mathcal{N}} \mathcal{P}(s, a'', s') \cdot V^{\pi}(s') \tag{1.6}$$

$$Q^{\pi}(s, a'') = \mathcal{R}(s, a'') + \gamma \cdot \sum_{s' \in \mathcal{N}} \mathcal{P}(s, a'', s') \cdot V^{\pi}(s') \tag{1.7}$$

$$Q^{\pi}(s, a'') = \mathcal{R}(s, a'') + \gamma \cdot \sum_{s' \in \mathcal{N}} \mathcal{P}(s, a, s') \cdot Q^{\pi}(s', a'') \tag{1.8}$$

Because the policy is deterministic, meaning the actions taken are fixed by the policy, equations 1.6, 1.7, and 1.8 can be rewritten in terms of policy-implied reward functions and transition probabilities, as given by equations 1.9, 1.10, and 1.11, respectively:

$$V^{\pi}(s) = \mathcal{R}^{\pi}(s) + \gamma \cdot \sum_{s' \in \mathcal{N}} \mathcal{P}^{\pi}(s, s') \cdot V^{\pi}(s') \tag{1.9}$$

$$Q^{\pi}(s, a'') = \mathcal{R}^{\pi}(s) + \gamma \cdot \sum_{s' \in \mathcal{N}} \mathcal{P}^{\pi}(s, s') \cdot V^{\pi}(s') \tag{1.10}$$

$$Q^{\pi}(s, a'') = \mathcal{R}^{\pi}(s) + \gamma \cdot \sum_{s' \in \mathcal{N}} \mathcal{P}^{\pi}(s, s') \cdot Q^{\pi}(s', a'') \tag{1.11}$$

2 Bellman Optimality for Infinite States

Consider a Markov decision process with infinite states $\mathcal{S} \in \{1, 2, 3, ...\}$, starting state $s = 1$, continuous actions $a \in [0, 1]$, and transition probabilities given below:

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

Assume a discount factor $\gamma = 0.5$. If transitions from s to s are assigned a reward of $1 + a$ and transitions from s to $s + 1$ are assigned a reward of $1 - a$, then the expected reward function is given by:

$$\mathbb{E}\big[\mathcal{R}(s, a)\big] = a(1 - a) + (1 - a)(1 + a)$$

The optimal value function $V^*(s)$ for all $s \in \mathcal{S}$ is given by the Bellman Optimality Equation:

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

Substituting the expected reward function and transition probability function yields:

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

Find a critical point by taking the derivative of the left hand side with respect to a and setting that derivative equal to zero:

$$\begin{aligned} 0 &= \frac{\partial}{\partial a} \Big\{ a(1 - a) + (1 - a)(1 + a) + 0.5\big(a \cdot V^*(s + 1) + (1 - a) \cdot V^*(s)\big) \Big\} \\ &= (1 - a) - a + (1 - a) - (1 + a) + 0.5\big(V^*(s + 1) - V^*(s)\big) \\ &= 1 - 4a + 0.5\big(V^*(s + 1) - V^*(s)\big) \end{aligned}$$

Since there are infinite states with the same transition probabilities and reward functions, $V^*(s)$ and $V^*(s + 1)$ are indistinguishable. Therefore, $V^*(s + 1) - V^*(s) = 0$ and the critical point exists at:

$$0 = 1 - 4a \implies a = 0.25$$

To verify that this critical point is a maximum, verify that the second derivative of the Bellman Optimality Equation is negative at $a = 0.25$:

$$\frac{\partial}{\partial a}\big(1 - 4a\big)\Big|_{a=0.25} = -4 < 0$$

Since the second derivative is negative, the critical point is in fact a maximum, and $\boxed{a = 0.25}$ yields the Optimal Value Function:

$$\begin{aligned} V^{*}(s) &= 0.25(1 - 0.25) + (1 - 0.25)(1 + 0.25) + 0.5\big(0.25 \cdot V^{*}(s + 1) + (1 - 0.25) \cdot V^{*}(s)\big) \\ &= 0.25(1 - 0.25) + (1 - 0.25)(1 + 0.25) + 0.5\big(0.25 \cdot V^{*}(s) + (1 - 0.25) \cdot V^{*}(s)\big) \\ &= 0.25(1 - 0.25) + (1 - 0.25)(1 + 0.25) + 0.5\big(V^{*}(s)\big) \end{aligned}$$

$$\therefore 0.5V^{*}(s) = 1.125$$

$$\therefore \boxed{V^{*}(s) = 2.25}$$

An optimal deterministic policy is a policy which yields the optimal value function. Therefore, the optimal deterministic policy is given by:

$$\boxed{{\pi^{*}}_{D}(s) = 0.25} \qquad [2.1]$$

3 Frog Escape Problem

The "frog escape problem" is specified by the conditions below:

- A frog is placed in a river containing lily pads at positions 0 to N.
- At position 0 is a snake, which eats the frog, and the frog successfully crosses the river if it reaches position N.
- During each move, the frog croaks either "A" of "B."
- If a frog at position i croaks "A," then the frog moves to position $i + 1$ with probability $(N - i)/N$ and to position $i - 1$ with probability i/N.
- If the frog croaks "B," then it jumps to any other position with uniform probability.
- The solution to the problem answers the following question: **what croak type at each position maximizes the probability of successfully crossing the river?**

The problem can be framed as a Markov decision process, where the state space, $\mathcal{S}$, represents the frog's position:

$$\mathcal{S} = \mathbb{Z} \, : \, \mathcal{S} \in [0, N]$$

The action space, $\mathcal{A}$, represents the croak types:

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

The transition function, dependent on the croak type, is given by:

$$\mathcal{P}(s, a, s') = \begin{cases} \frac{s}{N} & a = A, \; s' = s + 1 \\ \frac{N-s}{N} & a = A, \; s' = s + 1 \\ 0 & a = A, \; s' \notin \{s + 1, s - 1\} \\ \frac{1}{N} & a = B, \; s' \neq s \\ 0 & a = B, \; s' = s \end{cases}$$

where s is the current state, a is the croak action, and s' is the next state. When evaluating this problem, we are optimizing for the probability of successfully crossing the river, and we are unconcerned with the number of transitions required to do so. Therefore, set the reward transition function, $\mathcal{R}_T$, such that all states have a reward of zero, except the state at the winning position N:

$$\mathcal{R}_T(s, a, s') = \begin{cases} 1 & s' = N \\ 0 & s' \neq N \end{cases}$$

With this setup, the value function evaluated at any position in the river gives the probability of successfully crossing the river at that position. To solve the frog escape problem using this

MDP framework, identify all possible deterministic policies and calculate the value function at each state for each policy. Then identify the optimal policy, which maximizes the value function at each position. Using this approach, the optimal croak at each position is specified by the policy.

The frog problem solved using this MDP framework is implemented in `Frog_Escape.py`. Figures 1, 2, and 3 plot the probability of successfully crossing the river at each state, color-coded by the optimal croak action that the frog should take.

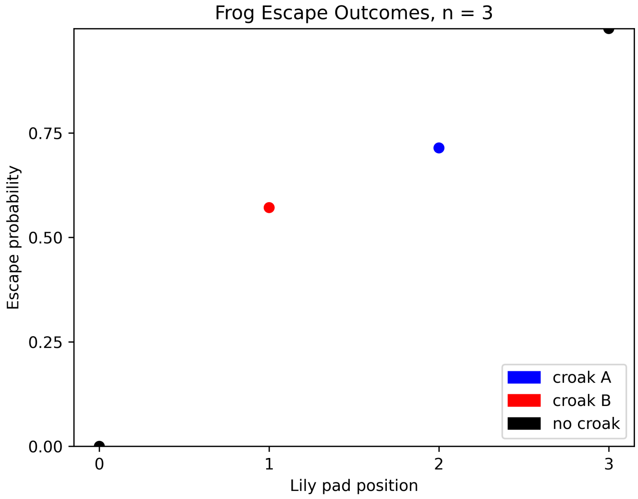

Figure 1: Probability of successfully escaping a river containing **3 lily pads** when applying optimal croak selection.

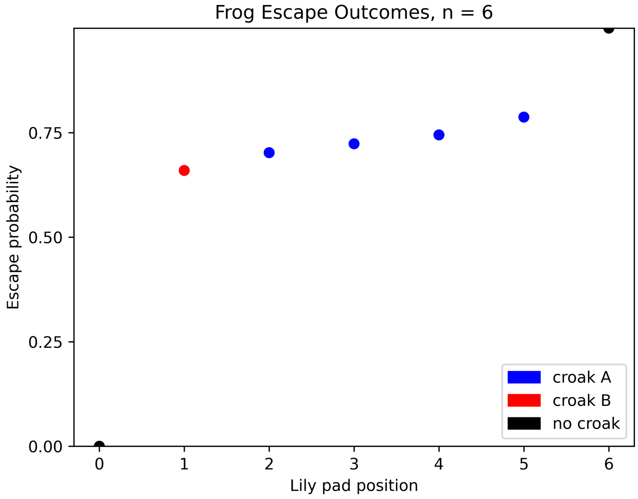

Figure 2: Probability of successfully escaping a river containing **6 lily pads** when applying optimal croak selection.

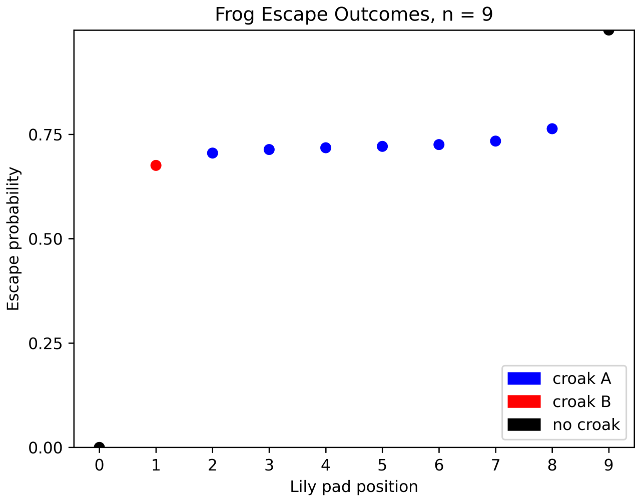

Figure 3: Probability of successfully escaping a river containing **9 lily pads** when applying optimal croak selection.

4 Continuous State, Continuous Action, Discrete Time, Non-terminating MDP

Consider a Markov decision process such that continuous states $s \in \mathbb{R}$ take continuous actions $a \in \mathbb{R}$ and transitions to a continuous next state $s' \in \mathbb{R}$ given by a normal distribution with mean s with constant variance σ^2.

- Denote reward (negative of the cost) at time t with R_t.
- Let the action-value function $Q(s, a)$ represent the infinite horizon expected discounted sum of rewards.
- Let the discounted sum of rewards at time t be denoted with G_t.

If the transition cost is given by $\exp(as')$, then $Q(s, a)$ can be obtained in the myopic case (where $\gamma = 0$) by the following:

$$G_t = \sum_{i=t+1}^{\infty} \gamma^{i-t-1} R_i = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots = R_{t+1} = -\exp(as')$$

$$\begin{aligned} Q(s, a) &= \mathbb{E}\left[G_t | S_t = s\right] \\ &= \mathbb{E}\left[R_{t+1} | S_t = s\right] \\ &= \mathbb{E}\left[-e^{as'} | S_t = s\right] \\ &= \mathbb{E}\left[-e^{a\mathcal{N}(s,\, \sigma^2)} | S_t = s\right] \\ &= \mathbb{E}\left[-e^{\mathcal{N}(as,\, a^2\sigma^2)} | S_t = s\right] \} \end{aligned}$$

The argument of the expectation is a multivariate log normal, for which, there is a known solution:

$$Q(s, a) = \mathbb{E}\left[-e^{\mathcal{N}(as,\, a^2\sigma^2)} | S_t = s\right] = -\exp\left(as + \frac{1}{2}a^2\sigma^2\right) \qquad [4.1]$$

To maximize $Q(s, a)$ with respect to a (equivalent to minimizing cost), find where the derivative of $Q(s, a)$ with respect to a equals zero. Denote the action which optimizes $Q(s, a)$ with a^*.

$$\frac{d}{da}\left\{-\exp\left(a^* s + \frac{1}{2}a^{*2}\sigma^2\right)\right\} = 0$$

$$\implies \frac{d}{da}\left\{-\exp(a^* s)\exp\left(\frac{1}{2}a^{*2}\sigma^2\right)\right\} = 0$$

$$\implies \frac{d}{da}\left\{\exp\left(a^* s\right)\exp\left(\frac{1}{2}a^{*2}\sigma^2\right)\right\} = 0$$

$$\implies s e^{a^* s} e^{\left(\frac{1}{2}a^{*2}\sigma^2\right)} + a^*\sigma^2 e^{a^* s} e^{\left(\frac{1}{2}a^{*2}\sigma^2\right)} = 0$$

$$\left(s + a^*\sigma^2\right)\left(e^{a^* s} e^{\left(\frac{1}{2}a^{*2}\sigma^2\right)}\right) = 0$$

$$s + a^*\sigma^2 = 0$$

$$\therefore \boxed{a^* = -\frac{s}{\sigma^2}} \qquad [4.2]$$

By substituting 4.2 into equations 4.1, an expression is formed for the expected optimal reward:

$$Q(s,a) = -\exp\left(\left[-\frac{s}{\sigma^2}\right]s + \frac{1}{2}\left[-\frac{s}{\sigma^2}\right]^2\sigma^2\right) = -\exp\left(-\frac{3s^2}{2\sigma^2}\right)$$

Taking the negative of $Q(s,a)$ gives the infinite horizon expected optimal cost:

$$\boxed{\text{Optimal Cost} = \exp\left(-\frac{3s^2}{2\sigma^2}\right)} \qquad [4.3]$$

Acknowledgements

LaTeX template comes from: `https://github.com/gijs-pennings/latex-homework`.