Homework Set 3

Problem 1: In the proof of the theorem on Bellman optimality, we used two inequalities that will be proved in this problem (the first one also reappeared in the proofs related to the value iteration and the policy iteration algorithms).

1. Let $g_1 : X \to \mathbb{R}$ and $g_2 : X \to \mathbb{R}$ denote two scalar-valued functions defined over the same domain. Show that

$$\left| \max_{x \in X} g_1(x) - \max_{x \in X} g_2(x) \right| \le \max_{x \in X} \left| g_1(x) - g_2(x) \right| .$$

   [Hint: Visualizing two such functions, e.g., defined over a finite, discrete domain may help.]

2. Show that the result of joint maximization is higher than or equal to that of sequential maximization. In particular, show that

$$\max_{x \in X, y \in Y} f(x, g(y)) \ge \max_{x \in X} f(x, \max_{y \in Y} g(y)),$$

   for two scalar-valued functions $f : X \times Z \to \mathbb{R}$ and $g : Y \to Z \subseteq \mathbb{R}$.
   [Hint: Use the fact that for a function $h : W \to \mathbb{R}$, $\max_{w \in W} h(w) \ge h(w')$ for any $w' \in W$.]

Problem 2: Consider a Markov decision process (MDP) in the infinite-horizon discounted setting with state space $\mathcal{S} = \{b, c\}$, action space $\mathcal{A} = \{x, y\}$, transition function $P(s'|s, a)$ with

$$\begin{aligned} P(b|b,x) &= 1.0\,, & P(c|b,x) &= 0.0\,, \\ P(b|b,y) &= 0.2\,, & P(c|b,y) &= 0.8\,, \\ P(b|c,x) &= 0.0\,, & P(c|c,x) &= 1.0\,, \\ P(b|c,y) &= 0.6\,, & P(c|c,y) &= 0.4\,, \end{aligned}$$

reward function $R(s, a)$ with

$$\begin{aligned} R(b,x) &= 0\,, & R(b,y) &= 0\,, \\ R(c,x) &= 1\,, & R(c,y) &= 1\,, \end{aligned}$$

and discount factor γ.

1. Suppose we are running a value iteration algorithm over the known MDP model to find the optimal value function. If at iteration $t = 6$, the value function is

$$V_6(b) = 10\,, \quad V_6(c) = 5\,,$$

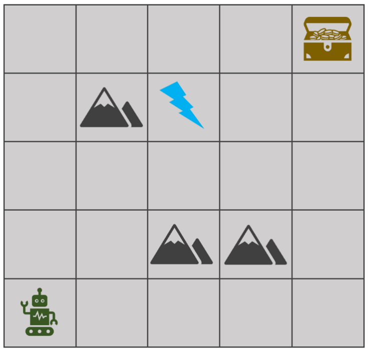
   what will the value function be at iteration $t = 7$? The answer will be in terms of γ.

2. Suppose we are running a policy iteration algorithm over the known MDP model to find an optimal policy. If at iteration $t = 8$, the value function for policy π_8 is

$$V^{\pi_8}(b) = 5\,, \quad V^{\pi_8}(c) = 15\,,$$

   what will the policy be at iteration $t = 9$?

Problem 3: Consider an MDP in the infinite-horizon discounted setting over the grid-world illustrated below, where the goal is to take the agent to the treasure chest while avoiding the lightning.



- The state space contains the cells of the grid-world.
- The agent starts at the bottom left corner.
- The action space is $\{\text{up}, \text{down}, \text{left}, \text{right}\}$.
- The agent can only move to its adjacent cells, i.e., the cells that are above, below, to the left, or to the right of its current cell. If the agent is not at the boundary cells, each action will take the agent to the expected cell with probability 0.85 and to one of the remaining three cells, each with probability 0.05. If the agent is at the boundary, it will remain at its current cell with the sum of probabilities that would have taken it outside the grid-world. The cells with a mountain cannot be accessed, i.e., if adjacent to the mountain, the agent will remain at its current cell with the probability that would have taken it to the mountain. All actions in the cell with the lightning bolt, the one with the treasure chest, and the ones with a mountain will keep the agent in its current cell.
- The agent receives a reward of 0 in every cell for all actions except two cells. If at the cell with a lightning bolt, it will receive a reward of -1 for all actions, and if at the cell with the treasure chest, it will receive a reward of $+1$ for all actions.
- The discount factor is 0.95.

1. Evaluate the deterministic, stationary policy shown in the figure below, where each arrow represents the prescribed action in each cell, over this MDP.
   (a) Apply the analytical solution for policy evaluation. Report the value function at all states. Also, report the expected discounted cumulative reward (the RL objective) $J(\pi)$ for this policy.
   (b) Suppose that instead of using the result in Part a, we would like to estimate $J(\pi)$ for this policy using a sampling-based technique. Generate $n = 100$ trajectories of length $T = 50$ according to the MDP and the policy. Compute and report the sample mean of the returns of these trajectories as the estimated RL objective $\hat{J}(\pi)$. Also, compute and report the error in this estimation, that is $|J(\pi) - \hat{J}(\pi)|$.

→	→	→	→	↑
←	↑	←	←	↑
↑	↑	→	→	→
↑	↓	↓	↓	↑
↑	→	→	↑	↑

(c) In the sampling-based technique in Part b, how do you expect increasing the length of the sampled trajectories to affect the estimation error? How do you expect increasing the number of the sampled trajectories to affect the estimation error? Explain your reasoning.

(d) Implement and run the iterative solution for approximate policy evaluation with a zero initialization for the value function. Pick the number of iterations in a way to ensure 0.01 accuracy in the final computed value function, i.e., $\|V_T - V^\pi\|_\infty \leq 0.01$ without using the results of the previous parts. Justify how you picked the number of iterations. Report the value function at all states.

(e) Plot the sequence of errors in the value function from the approximate policy evaluation with respect to the iterations. In particular, plot $\|V_t - V^\pi\|_\infty$ against $t \in \mathbb{N}_0$, where V_t is the value function at iteration t and V^π is the value function computed from the analytical solution.

2. Implement and run a value iteration algorithm to compute an optimal policy in this MDP. Initialize the value function at zero. Pick the number of iterations in a way to ensure 0.01 accuracy in the final computed value function. Demonstrate the learned policy and report the value function at all states.

3. Implement and run a policy iteration algorithm to compute an optimal policy in this MDP. Initialize the policy randomly, by using a uniform distribution over the actions. Pick the number of iterations in a way to ensure 0.01 accuracy in the final computed value function. Demonstrate the learned policy and report the value function at all states.

[**Note:** For this problem, attach all your code in a programming language of your choice to the end of your submission.]

Problem 4: Going through the steps of this problem, we aim to understand the dual linear program formulation for finding an optimal policy for an MDP $(\mathcal{S}, \mu_0, \mathcal{A}, P, R, \gamma)$ in the infinite-horizon discounted setting.

1. Recall the definition of the state-action occupancy measure $\nu_{\mu_0}^\pi(s,a)$ under a stationary, stochastic policy π:
$$\nu_{\mu_0}^\pi(s,a) = \sum_{t=0}^{\infty} \gamma^t \mathbb{P}(S_t = s, A_t = a | \mu_0, \pi, P).$$
Prove that for all $s \in \mathcal{S}$,
$$\sum_{a \in \mathcal{A}} \nu_{\mu_0}^\pi(s,a) = \mu_0(s) + \gamma \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} P(s|s',a') \nu_{\mu_0}^\pi(s',a').$$

2. Recall the definition of the state occupancy measure $\rho_{\mu_0}^\pi(s)$ under a stationary, stochastic policy π as
$$\rho_{\mu_0}^\pi(s) = \sum_{t=0}^{\infty} \gamma^t \mathbb{P}(S_t = s | \mu_0, \pi, P).$$
Prove that one can obtain the state occupancy measure and the state-action occupancy measure from one another according to:
$$\rho_{\mu_0}^\pi(s) = \sum_{a \in \mathcal{A}} \nu_{\mu_0}^\pi(s,a),$$
$$\nu_{\mu_0}^\pi(s,a) = \rho_{\mu_0}^\pi(s) \pi(a|s).$$

3. Represent the formula in Part 1 in vector form based on the state occupancy measure, i.e.,
$$\rho_{\boldsymbol{\mu_0}}^\pi = \boldsymbol{\mu_0} + \gamma \mathbf{P}^{\pi\top} \rho_{\boldsymbol{\mu_0}}^\pi$$
where $\rho_{\boldsymbol{\mu_0}}^\pi \in \mathbb{R}^{|\mathcal{S}|}$ is a column vector concatenating $\rho_{\mu_0}^\pi(s)$ for all $s \in \mathcal{S}$, $\boldsymbol{\mu_0} \in \mathbb{R}^{|\mathcal{S}|}$ is a column vector concatenating $\mu_0(s)$ for all $s \in \mathcal{S}$, and $\mathbf{P}^\pi \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$ is a probability transition matrix for the induced Markov chain under policy π.
[Hint: Use the properties proved in Part 2 and Problem 5 in Homework 2.]

4. Suppose we want to compute the state occupancy $\rho_{\mu_0}^\pi$ for a known policy π using the formula in Part 3. Show that the solution is unique and derive it in an analytical form.
[Hint: Reviewing Problem 5 in Homework 2 will be helpful.]

5. Recall that after solving the dual linear program and obtaining its optimal solution ν_{μ_0}, we construct a corresponding optimal policy as
$$\pi(a|s) = \begin{cases} \frac{\nu_{\mu_0}(s,a)}{\sum_{a' \in \mathcal{A}} \nu_{\mu_0}(s,a')} & \text{if } \sum_{a' \in \mathcal{A}} \nu_{\mu_0}(s,a') \neq 0, \\ \\ \text{arbitrary policy} & \text{otherwise.} \end{cases}$$
Based on the properties in Part 2 and the results in Part 3 and Part 4, explain why an optimal, stationary, stochastic policy should comply with this construction.

6. Consider the definition of the value function $V^\pi(\mu_0)$ under a stationary, stochastic policy π

$$V^\pi(\mu_0) = \mathbb{E}_{\substack{S_0 \sim \mu_0 \\ A_t \sim \pi(.|S_t) \\ S_{t+1} \sim P(.|S_t, A_t)}} \left[\sum_{t=0}^{\infty} \gamma^t R(S_t, A_t) \right].$$

Notice that here, the value function is defined more generally for a state distribution, particularly the initial distribution. This distribution-based value function $V^\pi(\mu)$ relates to the state-based value function $V^\pi(s)$, for any state distribution $\mu \in \Delta(\mathcal{S})$ as follows:

$$V^\pi(\mu) = \mathbb{E}_{S \sim \mu} \left[V^\pi(S) \right] = \sum_{s \in \mathcal{S}} \mu(s) V^\pi(s).$$

Prove that the following holds:

$$J(\pi) = V^\pi(\mu_0) = \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \nu^\pi_{\mu_0}(s, a) R(s, a).$$

7. Based on the previous parts of this problem, explain in words the role of the objective function and the constraints of the dual linear program formulated for finding an optimal policy.