Homework Set 3

Problem 1: Consider an MDP with three states $\mathcal{S} = \{1, 2, 3\}$, initial state 1, two actions $\mathcal{A} = \{g, h\}$, transition function $P(s'|s, a)$ defined as

$$\begin{aligned}
P(1|1,g) &= 0.1\,, \quad P(2|1,g) = 0.8\,, \quad P(3|1,g) = 0.1\,,\\
P(1|1,h) &= 0.8\,, \quad P(2|1,h) = 0.1\,, \quad P(3|1,h) = 0.1\,,\\
P(1|2,g) &= 0.1\,, \quad P(2|2,g) = 0.1\,, \quad P(3|2,g) = 0.8\,,\\
P(1|2,h) &= 0.1\,, \quad P(2|2,h) = 0.8\,, \quad P(3|2,h) = 0.1\,,\\
P(1|3,g) &= 0.8\,, \quad P(2|3,g) = 0.1\,, \quad P(3|3,g) = 0.1\,,\\
P(1|3,h) &= 0.1\,, \quad P(2|3,h) = 0.1\,, \quad P(3|3,h) = 0.8\,,
\end{aligned}$$

reward function $R(s, a)$ outputting 1 for $(3, h)$ and 0 otherwise, and discount factor 0.95 in the infinite-horizon, discounted setting. Consider a deterministic policy π

$$\pi(1) = g\,, \quad \pi(2) = g\,, \quad \pi(3) = h.$$

1. Compute $\mathbf{P}^\pi \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ which is the probability transition matrix for the induced Markov chain under policy π.

2. Recall the definition of the state occupancy measure $\rho^\pi_{\mu_0}(s)$ given initial state distribution μ_0 and under policy π:

$$\rho^\pi_{\mu_0}(s) = \sum_{t=0}^{\infty} \gamma^t \mathbb{P}(S_t = s|\mu_0, P, \pi).$$

   It can be shown that

$$\rho^\pi_{\boldsymbol{\mu_0}} = \boldsymbol{\mu_0} + \gamma \mathbf{P}^{\pi\top} \rho^\pi_{\boldsymbol{\mu_0}},$$

   where $\rho^\pi_{\boldsymbol{\mu_0}} \in \mathbb{R}^{|\mathcal{S}|}$ is a column vector concatenating $\rho^\pi_{\mu_0}(s)$ 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. Hence, $\rho^\pi_{\mu_0}(s)$ can be analytically computed by

$$\rho^\pi_{\boldsymbol{\mu_0}} = (I - \gamma \mathbf{P}^{\pi\top})^{-1} \boldsymbol{\mu_0}.$$

   Use this formula to compute $\rho^\pi_{\mu_0}(s)$ and report its value in all states for policy π.

3. Compute the normalized state occupancy measure $\bar{\rho}^\pi_{\mu_0}$ and report its value in all states for the same policy.

4. Now, consider another MDP that is identical to the original MDP except that its transition function $P'(s'|s, a)$ is different in the following entries:

$$\begin{aligned}
P'(1|1,g) &= 0.15\,, \quad P'(2|1,g) = 0.7\,, \quad P'(3|1,g) = 0.15\,,\\
P'(1|2,g) &= 0.05\,, \quad P'(2|2,g) = 0.05\,, \quad P'(3|2,g) = 0.9\,,\\
P'(1|2,h) &= 0.15\,, \quad P'(2|2,h) = 0.8\,, \quad P'(3|2,h) = 0.05\,.
\end{aligned}$$

   Compute the L1-difference between the transition probabilities of these two MDPs at every state under policy π. In particular, report

$$\left\| P'(.|s, \pi(s)) - P(.|s, \pi(s)) \right\|_1$$

   for all $s \in \mathcal{S}$.

5. Using the results of the previous parts, apply the simulation lemma to upper-bound the difference between the value functions of the given policy over these two MDPs in the initial state $s = 1$, i.e., $|V'^{\pi}(1) - V^{\pi}(1)|$.

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

Problem 2: Consider the same MDP as in Problem 1 with three states $\mathcal{S} = \{1, 2, 3\}$, initial state 1, two actions $\mathcal{A} = \{g, h\}$, transition function $P(s'|s, a)$ defined as

$$\begin{aligned}
P(1|1,g) &= 0.1\,, \quad & P(2|1,g) &= 0.8\,, \quad & P(3|1,g) &= 0.1\,,\\
P(1|1,h) &= 0.8\,, \quad & P(2|1,h) &= 0.1\,, \quad & P(3|1,h) &= 0.1\,,\\
P(1|2,g) &= 0.1\,, \quad & P(2|2,g) &= 0.1\,, \quad & P(3|2,g) &= 0.8\,,\\
P(1|2,h) &= 0.1\,, \quad & P(2|2,h) &= 0.8\,, \quad & P(3|2,h) &= 0.1\,,\\
P(1|3,g) &= 0.8\,, \quad & P(2|3,g) &= 0.1\,, \quad & P(3|3,g) &= 0.1\,,\\
P(1|3,h) &= 0.1\,, \quad & P(2|3,h) &= 0.1\,, \quad & P(3|3,h) &= 0.8\,,
\end{aligned}$$

reward function $R(s, a)$ outputting 1 for $(3, h)$ and 0 otherwise, and discount factor 0.95 in the infinite-horizon, discounted setting. Consider a deterministic policy π

$$\pi(1) = g\,, \quad \pi(2) = g\,, \quad \pi(3) = h.$$

1. Evaluate policy π using the analytical solution for policy evaluation. Report the value function V^{π} at all states.

2. Assume oracle access to $P(s'|s, a)$ where by passing a state-action pair (s, a), a sample $s' \sim P(.|s, a)$ can be generated. Estimate the true transition function P by sampling each state-action pair 100 times. Report the estimated transition function $\hat{P}$.

3. Compute the L1-difference (error) between the true and estimated transition probabilities at every state under policy π. In particular, report

$$\left\| \hat{P}(.|s, \pi(s)) - P(.|s, \pi(s)) \right\|_1$$

for all $s \in \mathcal{S}$.

4. **[Bonus]** Apply the simulation lemma to upper-bound the difference between the value functions of the given policy over the true and estimated transition functions in the initial state $s = 1$, i.e., $|\hat{V}^{\pi}(1) - V^{\pi}(1)|$.
[Hint: You can reuse the normalized state occupancy measure computed in Problem 1.]

5. Evaluate policy π over the estimated transition function $\hat{P}$ using the analytical solution for policy evaluation. Report the value function $\hat{V}^{\pi}$ at all states.

6. Measure and report the exact difference between the value functions computed in Part 1 and Part 5 in the initial state $s = 1$, i.e., $|\hat{V}^{\pi}(1) - V^{\pi}(1)|$.

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

Problem 3: Consider linear function approximation for representing the state-action value function in the infinite-horizon, discounted setting. In this case, one may define a set of k features $\phi_l(s,a)$ for $l \in \{1, 2, \ldots, k\}$ and approximate the state-action value function as a linear combination of these features

$$Q_\theta(s,a) = \theta^\top \phi(s,a) = \sum_{l=1}^{k} \theta_l \phi_l(s,a)$$

weighted by parameters θ_l. Here θ and $\phi(s,a)$ are column vectors concatenating θ_l and $\phi_l(s,a)$, respectively, for all $l \in \{1, 2, \ldots, k\}$.

1. For an MDP with a finite state space $\mathcal{S}$ and a finite action space $\mathcal{A}$, one can use a tabular representation of the state-action value function $Q(s,a)$ for all states and actions. Show that the tabular representation is a special class of linear functions by creating a class of linear functions
$$\mathcal{Q} = \{Q_\theta : Q_\theta(s,a) = \theta^\top \phi(s,a) \textit{ for all } s \in \mathcal{S} \textit{ and } a \in \mathcal{A},\ \theta \in \mathbb{R}^k\},$$
parametrized by θ that is equivalent to the tabular representation. Determine how the number of features k and the vector-valued function $\phi(s,a) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^k$ that maps each state-action pair to a k-dimensional feature vector should be defined.

2. Consider an MDP with a finite state space $\mathcal{S}$ and a finite action space $\mathcal{A}$. Suppose we have identified that the state space and the action space can be partitioned based on the similarity between the states and actions:
$$\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2 \cup \ldots \cup \mathcal{S}_n \textit{ such that } \mathcal{S}_{j_1} \cap \mathcal{S}_{j_2} = \emptyset \textit{ for all } j_1 \neq j_2,$$
$$\mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2 \cup \ldots \cup \mathcal{A}_m \textit{ such that } \mathcal{A}_{j_1} \cap \mathcal{A}_{j_2} = \emptyset \textit{ for all } j_1 \neq j_2,$$
i.e., the subsets $\mathcal{S}_j$ and $\mathcal{A}_j$ group the similar states and actions (in terms of value functions), respectively. Create a class of linear functions
$$\mathcal{Q} = \{Q_\theta : Q_\theta(s,a) = \theta^\top \phi(s,a) \textit{ for all } s \in \mathcal{S} \textit{ and } a \in \mathcal{A},\ \theta \in \mathbb{R}^k\},$$
parametrized by θ that uses this partitioning to efficiently represent the state-action value function $Q(s,a)$ for all states and actions. Determine how the number of features k and the vector-valued function $\phi(s,a) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^k$ that maps each state-action pair to a k-dimensional feature vector should be defined.

3. **[Bonus]** Suppose we want to use supervised learning to approximate the state-action value function $Q^\pi(s,a)$ of an MDP with a large state space $\mathcal{S}$ and large action space $\mathcal{A}$ under policy π. We have collected a data set $\{x^i = (s^i, a^i), y^i\}_{i=1}^N$ of size N, where x^i represents the i$^{\text{th}}$ sampled state-action pair and y^i represents a sampled return for x^i. The goal is to learn the best fit to the data by empirical risk minimization with square loss, i.e., solving
$$\hat{Q}^\pi(s,a) = \underset{Q \in \mathcal{Q}}{\arg\min} \sum_{i=1}^{N} \left(Q(s^i, a^i) - y^i\right)^2,$$
where
$$\mathcal{Q} = \{Q_\theta : Q_\theta(s,a) = \theta \phi(s,a) \textit{ for all } s \in \mathcal{S} \textit{ and } a \in \mathcal{A},\ \theta \in \mathbb{R}\}$$

represents a class of linear functions based on a single (known) feature $\phi(s, a)$ and parametrized by a single parameter θ. Assume that

$$\sum_{i=1}^{N} \phi^2(s^i, a^i) \neq 0.$$

Derive a closed-form solution for $\hat{Q}^\pi(s, a)$.
[Hint: You can use the fact that this linear parametrization makes the objective function of the minimization problem a convex function in θ such that the solution is a stationary point of the objective function, i.e., a point where the derivative is zero.]