Homework Set 5

Problem 1: We would like to apply the conservative policy iteration algorithm. We have started with an initial policy π_0 that selects actions over the action space $\mathcal{A} = \{-1, 0, 1\}$ uniformly at random for each state in the state space $\mathcal{S} = \{-3, -2, -1, 0, 1, 2, 3\}$. The class of state-action value function Q we are considering is

$$\mathcal{Q} = \{Q_\theta : Q_\theta(s,a) = \theta_1 s^2 + \theta_2 a^2 + \theta_3 sa + \theta_4,\ \theta \in \mathbb{R}^4\}.$$

1. Write the objective function of the empirical risk minimization with square loss if we want to fit a Q function to the following data set obtained by the roll-in and roll-out process from policy π_0:

$$s^1 = 1, a^1 = 0, y^1 = 1$$
$$s^2 = -2, a^2 = 1, y^2 = 0$$
$$s^3 = 1, a^3 = -1, y^3 = 3$$

   The only variables in the objective function should be θ_1, θ_2, θ_3, and θ_4. There is no need to solve the minimization problem.

2. After collecting a larger data set from the current policy π_0 and solving the empirical risk minimization problem over θ, we have arrived at

$$\hat{Q}^{\pi_0}(s,a) = 2s^2 + a^2 - sa + 0.5.$$

   Compute the updated policy π_1 for state $s = 1$ and state $s = 2$ if the parameter α is set to 0.25.

Problem 2: We aim to find an optimal policy using Q-learning with function approximation, i.e., restricting the state-action value function Q to

$$\mathcal{Q} = \{Q_\theta : \theta \in \mathbb{R}^d\}.$$

In the course of the algorithm applied over an MDP with state space $\mathcal{S} = \{b, c, d, e\}$, action space $\mathcal{A} = \{x, y\}$, and discount factor $\gamma = 0.9$, we have obtained

$$Q(b,x) = -1.5\ ,\ Q(b,y) = -2.5,$$
$$Q(c,x) = -0.5\ ,\ Q(c,y) = -1.0,$$
$$Q(d,x) = 0.0\ ,\ Q(d,y) = -0.25,$$
$$Q(e,x) = -0.5\ ,\ Q(e,y) = 0.75.$$

1. Form an ϵ-greedy policy according to this Q function, setting $\epsilon = 0.1$.

2. Using this ϵ-greedy policy, suppose we have collected the following sample sub-trajectory

$$(s_t = c, a_t = x, r_t = 1, s_{t+1} = e, a_{t+1} = x, r_{t+1} = -2, s_{t+2} = b, a_{t+2} = y, r_{t+2} = -0.5, s_{t+3} = b)\,.$$

   Create a data set of three data points that can be used to update the parameter θ of the Q function using supervised learning.

3. Let $\mathcal{Q}$ represent a class of linear functions

$$Q_\theta(s,a) = \theta^\top \phi(s,a) = \sum_{l=1}^{2} \theta_l \phi_l(s,a)$$

based on feature $\phi = [\phi_1(s,a) \quad \phi_2(s,a)]^\top$ defined as

$$\begin{aligned} \phi_1(s,a) &= -2\mathbb{1}[s=b] - \mathbb{1}[s=c] - 0.5\mathbb{1}[s=d], \\ \phi_2(s,a) &= \mathbb{1}[a=x] - \mathbb{1}[a=y], \end{aligned}$$

and with parameter $\theta = [\theta_1 \quad \theta_2]^\top \in \mathbb{R}^2$. Write the objective function of the empirical risk minimization with square loss when fitting a new Q function to the data set in Part 2. The only variables in the objective function should be θ_1 and θ_2.

4. Solve the optimization problem in Part 3. Report the values the solution yields for parameter θ and the objective function.

Problem 3: Recall that in the policy gradient algorithms like REINFORCE, actor-critic, and advantage actor-critic where we search over a parameterized class of policies

$$\Pi = \{\pi_\theta : \theta \in \mathbb{R}^d\},$$

the gradient of the objective function (value function) relies on computing $\nabla_\theta \log \pi_\theta(a_t|s_t)$ for sampled state-action pair (s_t, a_t). Notice that $\pi_\theta(a|s)$ is a function over state-action pairs, i.e., $\pi_\theta : \mathcal{S} \times \mathcal{A} \to (0,1)$ that is parameterized by $\theta \in \mathbb{R}^d$.

1. Compute $\nabla_\theta \log \pi_\theta(a_t|s_t)$ for the case of using the class of softmax policies

$$\forall s \in \mathcal{S}, \forall a \in \mathcal{A} : \quad \pi_\theta(a|s) = \frac{\exp(\theta_{s,a})}{\sum_{a' \in \mathcal{A}} \exp(\theta_{s,a'})},$$

where $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$.

2. Compute $\nabla_\theta \log \pi_\theta(a_t|s_t)$ for the case of using the class of softmax linear policies

$$\forall s \in \mathcal{S}, \forall a \in \mathcal{A} : \quad \pi_\theta(a|s) = \frac{\exp(\theta^\top \phi(s,a))}{\sum_{a' \in \mathcal{A}} \exp(\theta^\top \phi(s,a'))},$$

where $\theta \in \mathbb{R}^d$ and $\phi(s,a) \in \mathbb{R}^d$.

3. Compute $\nabla_\theta \log \pi_\theta(a_t|s_t)$ for the case of using the class of softmax neural policies

$$\forall s \in \mathcal{S}, \forall a \in \mathcal{A} : \quad \pi_\theta(a|s) = \frac{\exp(f_\theta(s,a))}{\sum_{a' \in \mathcal{A}} \exp(f_\theta(s,a'))},$$

where $\theta \in \mathbb{R}^d$ and $f_\theta(s,a) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. Assume that $f_\theta(s,a)$ is a differentiable function and keep your solution in terms of its partial derivatives.

4. Consider searching for an optimal policy over a parameterized class of softmax policies

$$\Pi = \{\pi_\theta : \pi_\theta(a|s) = \frac{\exp(\theta_{s,a})}{\sum_{a' \in \mathcal{A}} \exp(\theta_{s,a'})} \; for\ all\ s \in \mathcal{S}, a \in \mathcal{A}\ and\ \theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}\},$$

where $\mathcal{S} = \{1, 2, \ldots, 100\}$ and $\mathcal{A} = \{b, c, d, e\}$, by applying the advantage actor-critic algorithm in the infinite-horizon discounted setting. At iteration t of the algorithm, we arrive at a state-action pair $(s_t, a_t) = (3, b)$ for which the advantage function $A^{\pi_{\theta^t}}(s_t = 3, a_t = b) < 0$. If we update the parameters by applying stochastic gradient ascent with the gradient estimate formed by this sample, what will happen (decrease, stay the same, or increase) to each the following parameters? [Explain your reasoning.]

(a) parameter $\theta_{1,d}$

(b) parameter $\theta_{2,b}$

(c) parameter $\theta_{3,b}$

(d) parameter $\theta_{3,c}$

Problem 4: In maximum entropy reinforcement learning (and inverse reinforcement learning), the reward is augmented with the entropy of the policy, i.e., the goal is to find a stochastic policy that maximizes

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

where $\lambda \geq 0$. Soft policy iteration algorithm is a variant of policy iteration algorithm that finds such an optimal stochastic policy through iterative policy evaluation and policy improvement. The policy evaluation step computes the state-action value function $Q^{\pi_t}(s, a)$ of the current policy π_t for all $s \in \mathcal{S}, a \in \mathcal{A}$ and the policy improvement step computes the next policy π_{t+1} using

$$\pi_{t+1}(.|s) = \arg\max_{\pi \in \Pi} \quad \mathbb{E}_{A \sim \pi(.|s)} \left[Q^{\pi_t}(s, A) \right] + \lambda \mathcal{H}(\pi(.|s)) \qquad \forall s \in \mathcal{S}.$$

Suppose the action space $\mathcal{A}$ is discrete and finite. The policy improvement step can be written as a constrained optimization problem

$$\begin{aligned} \max_{\pi(.|s) \geq \mathbf{0}} \quad & \mathbb{E}_{A \sim \pi(.|s)} \left[Q^{\pi_t}(s, A) \right] + \lambda \mathcal{H}(\pi(.|s)) \\ \textit{subject to} \quad & \sum_{a \in \mathcal{A}} \pi(a|s) = 1, \end{aligned}$$

and solved individually for each $s \in \mathcal{S}$. This constrained optimization problem is equivalent to

$$\max_{\pi(.|s) \geq \mathbf{0}} \min_{\beta \in \mathbb{R}} \quad \mathcal{L}(\pi, \beta),$$

where $\mathcal{L}(\pi, \beta)$ is called the Lagrangian function and is defined as

$$\mathcal{L}(\pi, \beta) = \mathbb{E}_{A \sim \pi(.|s)} \left[Q^{\pi_t}(s, A) \right] + \lambda \mathcal{H}(\pi(.|s)) - \beta (\sum_{a \in \mathcal{A}} \pi(a|s) - 1).$$

1. Find the derivative of the Lagrangian function with respect to $\pi(a|s)$ for each $a \in \mathcal{A}$.
2. Find the derivative of the Lagrangian function with respect to β.
3. Find the stationary point of the Lagrangian function, i.e., $\pi(a|s)$ and β where the derivatives (computed in Part 1 and Part 2) are zero. The answer should be based on the known values.
4. **[Bonus]** The resulting policy in Part 3 will be the optimal solution $\pi_{t+1}(.|s)$ to the original constrained optimization problem of the policy improvement step. Compute and write in simplified form its value function $V^{\pi_{t+1}}(s)$. Notice that in this setting, the state value function relates to the state-action value function through

$$V^{\pi}(s) = \mathbb{E}_{a\sim\mathcal{A}}[Q^{\pi}(s,a)] + \lambda\mathcal{H}(\pi(.|s)) \qquad \forall s \in \mathcal{S}.$$

Problem 5: Consider the setting of reinforcement learning from human feedback. We would like to learn a reward function from preferences provided by a human over pairs of sub-trajectories. We follow the Bradley-Terry probabilistic model of human preference using discounted return, with discount factor $\gamma = 0.95$, as quality of a sub-trajectory. In the course of learning, we have arrived at a reward function $\hat{R}(s,a)$ with partial values listed below:

$$\begin{aligned}
\hat{R}(empty, stay) &= -1, \quad \hat{R}(empty, explore) = 1,\\
\hat{R}(monster, stay) &= -2, \quad \hat{R}(monster, evade) = 3, \quad \hat{R}(monster, befriend) = -5,\\
\hat{R}(food, stay) &= -2, \quad \hat{R}(food, explore) = -2, \quad \hat{R}(food, collect) = 5,\\
\hat{R}(resource, stay) &= -2, \quad \hat{R}(resource, explore) = -2, \quad \hat{R}(resource, collect) = 4.
\end{aligned}$$

Now, consider the following sub-trajectories:

$$\begin{aligned}
\tau_1 &= (s_0 = empty, a_0 = stay, s_1 = monster, a_1 = evade, s_2 = resource, a_2 = collect)\\
\tau_2 &= (s_0 = empty, a_0 = explore, s_1 = food, a_1 = collect, s_2 = monster, a_2 = befriend)\\
\tau_3 &= (s_0 = empty, a_0 = explore, s_1 = empty, a_1 = explore, s_2 = resource, a_2 = collect)
\end{aligned}$$

1. Compute the discounted return for each sub-trajectory τ_1, τ_2, and τ_3.
2. If we query the human with trajectory pair (τ_1, τ_2), with what probability will the human prefer τ_1 over τ_2 according to $\hat{R}$?
3. If we query the human with trajectory pair (τ_1, τ_3), with what probability will the human prefer τ_1 over τ_3 according to $\hat{R}$?
4. Suppose when queried about (τ_1, τ_2), the human picks τ_1 and when queried about (τ_1, τ_3), the human picks τ_3. Compute the (maximum likelihood estimation) loss function over these two data points according to $\hat{R}$.
5. If the human picks τ_2 over τ_1 and τ_3 over τ_1, how do you expect the loss function to change (decrease, stay the same, or increase) compared to Part 4? [Explain your reasoning without additional computation.]