Problem 1 (Q-Value Iteration) (15+15=30 points)

(a) Recall that in Lecture 3, we define $V^*(s) := \max_\pi V^\pi(s)$ and $Q^*(s,a) := \max_\pi Q^\pi(s,a)$. Suppose $\gamma \in (0,1)$. Prove the following Bellman optimality equations:

$$V^*(s) = \max_a Q^*(s,a) \tag{1}$$

$$Q^*(s,a) = R_{s,a} + \gamma \sum_{s'} P^a_{ss'} V^*(s'). \tag{2}$$

Please carefully justify every step of your proof. (Hint: For (1), you may first prove that $V^*(s) \le \max_a Q^*(s,a)$ and then show $V^*(s) < \max_a Q^*(s,a)$ cannot happen by contradiction. On the other hand, (2) can be shown by using the similar argument or by leveraging the fact that $Q^\pi(s,a) = R_{s,a} + \gamma \sum_{s'} P^a_{ss'} V^\pi(s')$)

Suppose:

① For any finite MDP, there always exists an π^*, s.t. $\pi^* \ge \pi \ \forall \pi \in \Pi$.

② A deterministic defined policy $\pi^*(a|s)$ can always be found.

$$\pi^*(a|s) = \begin{cases} 1, & \text{if } a = \arg\max_{a' \in A} Q^*(s,a') \\ 0, & \end{cases}$$

(1)

$V^*(s) = \max_{\pi \in \Pi} V^\pi(s)$ and already know that $V^\pi(s) = \sum_{a \in A} \pi(a|s) \cdot Q^\pi(s,a)$.

With a given optimal policy π^*, for each $s \in S$, take action $a \in \arg\max_{a \in A} Q^{\pi^*}(s,a)$.

we have $V^*(s) = V^{\pi^*}(s) = \sum_{a \in A} \pi^*(a|s) \cdot Q^{\pi^*}(s,a) \le \max_{a \in A} Q^*(s,a) \ldots (1)$

Now assume a state s', an action a', and a policy ϕ s.t. $V^*(s') < Q^\phi(s',a') = \max_{a \in A} Q^*(s',a)$

$\Rightarrow V^\phi(s') = \sum_{a \in A} \phi(a|s') \cdot Q^\phi(s,a) = \max_{a \in A} Q^*(s',a)$ $\boxed{V^\phi(s') > V^*(s')}$.

which contradicts $V^*(s) = \max_{\pi \in \Pi} V^\pi(s)$ #

(2)

Since we have $Q^\pi(s,a) = R_{s,a} + \gamma \sum_{s' \in S} P^a_{ss'} V^\pi(s')$ and $V^*(s) = \max_{a \in A} Q^*(s,a)$

With a given optimal policy π^*, for each $s \in S$, take action $a \in \arg\max_{a \in A} Q^{\pi^*}(s,a)$.

We can derive $Q^{\pi^*}(s,a) = Q^*(s,a) = R_{s,a} + \boxed{\gamma \sum_{s' \in S} P^a_{ss'} V^{\pi^*}(s')} \longrightarrow V^{\pi^*}(s') = V^*(s')$

$\Rightarrow Q^*(s,a) = R_{s,a} + \gamma \sum_{s' \in S} P^a_{ss'} \cdot V^*(s')$

(b) Based on (a), we thereby have the recursive Bellman optimality equation for the optimal action-value function Q_* as:

$$Q^*(s,a) = R_{s,a} + \gamma \sum_{s'} P^a_{ss'} \big(\max_{a'} Q^*(s',a') \big) \tag{3}$$

Similar to the standard Value Iteration, we can also study the Q-*Value Iteration* by defining the Bellman optimality operator $T^* : \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \to \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ for the action-value function: for every state-action pair (s,a)

$$[T^*(Q)](s,a) := R_{s,a} + \gamma \sum_{s'} P^a_{ss'} \max_{a'} Q(s',a') \tag{4}$$

Show that the operator T^* is a γ-contraction operator in terms of ∞-norm. Please carefully justify every step of your proof. (Hint: For any two action-value functions Q, Q', we have $\|T^*(Q) - T^*(Q')\|_\infty = \max_{(s,a)} \big|[T^*(Q)](s,a) - [T^*(Q')](s,a)\big|$)

By hint: $\| T^*(Q) - T^*(Q') \|_\infty = \max_{(s,a)} \big| [T^*(Q)](s,a) - [T^*(Q')](s,a) \big|$

(by def.) $= \max_{(s,a)} \Big| \Big[\cancel{R_{s,a}} + r \sum_{s'} P^a_{ss'} \max_{a'} Q(s',a') \Big] - \Big[\cancel{R_{s,a}} + r \sum_{s'} P^a_{ss'} \max_{a'} Q'(s',a') \Big] \Big|$

$= \gamma \cdot \max_{(s,a)} \Big| \sum_{s'} P^a_{ss'} \Big(\max_{a'} Q(s',a') - \max_{a'} Q'(s',a') \Big) \Big|$

$(|a|+|b| > |a+b|)$ $\le \gamma \cdot \max_{(s,a)} \sum_{s'} P^a_{ss'} \Big| \max_{a'} Q(s',a') - \max_{a'} Q'(s',a') \Big|$

$\le \gamma \cdot \max_{(s,a)} \sum_{s'} P^a_{ss'} \max_{a'} \big| Q(s',a') - Q'(s',a') \big|$

$\max_{x \in X} | f(x) - g(x) | \ge | \max_x f(x) - \max_x g(x) |$

$\le r \| Q - Q' \|_\infty$

Problem 2 (Regularized MDPs) (10+10=20 points)

In Lecture 4, we formally describe the regularized MDP, which is a direct extension of the classic MDP with a regularizer Ω. In this problem, for simplicity, suppose we use the Shannon entropy as our regularizer, i.e., $\Omega(\pi(\cdot|s)) \equiv H(\pi(\cdot|s)) := -\sum_{a \in \mathcal{A}} \pi(a|s) \ln \pi(a|s)$. Let us verify a few important properties mentioned in Lecture 4 as follows.

(a) Recall that we introduce the "regularized Bellman expectation operator" T^π_Ω as

$$[T^\pi_\Omega V](s) := R^\pi_s + \Omega(\pi(\cdot|s)) + \gamma P^\pi_{ss'} V. \tag{5}$$

Please verify that T^π_Ω is a contraction operator in L_∞ norm. (Hint: Try to extend the proof procedure of the contraction property of T^π in Lecture 3)

$\| T^\pi_\Omega (V) - T^\pi_\Omega (V') \|_\infty = \max_s \big| [T^\pi_\Omega (V)](s) - [T^\pi_\Omega (V')](s) \big|$

$= \max_s \Big| \Big(\cancel{R^\pi_s} + \cancel{\Omega(\pi(\cdot|s))} + \gamma P^\pi_{ss'} V(s') \Big) - \Big(\cancel{R^\pi_s} + \cancel{\Omega(\pi(\cdot|s))} + \gamma P^\pi_{ss'} V'(s') \Big) \Big|$

$= \max_s \Big| r P^\pi_{ss'} \big(V(s') - V'(s') \big) \Big|$

$\le \gamma \max_s P^\pi_{ss'} \big| V(s') - V'(s') \big| \le \gamma \| V - V' \|_\infty$

(b) Moreover, under regularized MDPs, we study the optimal value functions V_Ω^* and optimal Q functions Q_Ω^* and learn the Bellman optimality equations as

$$V_\Omega^*(s) = \max_{\pi\in\Pi} R_s^\pi + \gamma P_s^\pi V_\Omega^* \tag{6}$$

$$Q_\Omega^*(s,a) = R_{s,a} + \gamma E_{s'\sim P(\cdot|s,a)}[V_\Omega^*(s')]. \tag{7}$$

Could you design an iterative algorithm that can obtain V_Ω^* and Q_Ω^*? Please clearly write down the complete pseudo code of your algorithm and provide comments on each line of your pseudo code. (Hint: Try to extend the Value Iteration for standard MDPs to the regularized MDPs based on Equation 6)

For $V_\Omega^*(s)$:

initialize $V_\Omega^*(s) = 0 \quad \forall s \in S$, $\Delta \leftarrow 0$, $v \leftarrow V_\Omega^*(s)$

Repeat until $V_\Omega^\pi(s)$ converge

for $s \in S$:

$V_\Omega^*(s) \longleftarrow \max_{\pi\in\Pi} R_s^\pi + \gamma \mathbb{P}_s^\pi V_\Omega^*(s')$

For $s \in S$:

$\pi^*(s) \longleftarrow \operatorname{argmax}_{a\in A} \left(\operatorname{argmax}_{\pi\in\Pi} R_s^\pi + \gamma \mathbb{P}_s^\pi \cdot V_\Omega^*(s') \right)$

For $Q_\Omega^*(s,a)$:

initialize $V_\Omega^*(s) = 0 \quad \forall s \in S$, initialize $\pi(s)$ randomly

Repeat until $\pi(s)$ converge:

Repeat until $V_\Omega^\pi(s)$ converge

for $s \in S$

$V_\Omega^*(s) \longleftarrow \max_{\pi\in\Pi} R_s^\pi + \gamma \mathbb{P}_s^\pi V_\Omega^*(s')$

For all (s,a) pair

$Q_\Omega^*(s,a) = R_s^a + \gamma \cdot \mathbb{E}_{s'\sim P(\cdot|s,a)}[V_\Omega^*(s')]$

$\pi^*(s) = \operatorname{argmax}_{a\in A} \left(Q_\Omega^*(s,a) \right)$

Problem 3 (A Property Used in Policy Gradient) (10 points)

Show the following useful property discussed in Lectures 5-6: for any function $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$,

$$\mathbb{E}_{\tau \sim P_\mu^{\pi_\theta}}\Big[\sum_{t=0}^{\infty} \gamma^t f(s_t, a_t)\Big] = \frac{1}{1-\gamma}\mathbb{E}_{s \sim d_\mu^{\pi_\theta}} \mathbb{E}_{a \sim \pi_\theta(\cdot|s)}\big[f(s,a)\big] \quad (8)$$

(Hint: It might be slightly easier to go from the RHS to LHS. Specifically, you may first expand the RHS of (8) into a sum of $f(s,a)$ over s and a and then apply the definition of $d_\mu^{\pi_\theta}$, which involves a sum of probability over t. Next, try to reorganize the triple summation into the form of the LHS of (8))

$$\mathbb{E}_{\tau \sim \mathbb{P}_M^{\pi_\theta}}\Big[\sum_{t=0}^{\infty} r^t f(s_t, a_t)\Big] = \sum_{\tau} \sum_{t=0}^{\infty} r^t f(s_t, a_t)\, \mathbb{P}_M^{\pi_\theta}(\tau)$$

$$\frac{1}{1-r}\mathbb{E}_{s \sim d_M^{\pi_\theta}} \mathbb{E}_{a \sim \pi_\theta(\cdot|s)}[f(s,a)] = \frac{1}{1-r}\sum_{s \in S} d_M^{\pi_\theta}(s) \cdot \mathbb{E}_{a \sim \pi_\theta(\cdot|s)}[f(s,a)]$$

$$\because d_M^{\pi_\theta}(s) := \mathbb{E}_{s_0 \sim M}[d_{s_0}^{\pi_\theta}(s)] \qquad = \cancel{\frac{1}{1-r}} \sum_{s \in S} \mathbb{E}_{s_0 \sim M}[d_{s_0}^{\pi_\theta}(s)] \cdot \mathbb{E}_{a \sim \pi_\theta(\cdot|s)}[f(s,a)]$$

$$\because d_{s_0}^{\pi_\theta}(s) = \cancel{(1-r)} \sum_{t=0}^{\infty} r^t \cdot \mathbb{P}(s_t = s | s_0) \qquad = \sum_{s \in S} \mathbb{E}_{s_0 \sim M}\Big[\sum_{t=0}^{\infty} r^t \cdot \mathbb{P}^{\pi_\theta}(s_t = s | s_0)\Big] \cdot \mathbb{E}_{a \sim \pi_\theta(\cdot|s)}[f(s,a)]$$

Triple Summation :

$$= \sum_{s \in S} \sum_{t=0}^{\infty} r^t \cdot \mathbb{P}_M^{\pi_\theta} \cdot \sum_{a \in A} \pi_\theta(a|s) \cdot [f(s_t, a_t)]$$

$$= \sum_{s \in S} \sum_{a \in A} \cancel{\pi_\theta(a|s)} \sum_{t=0}^{\infty} r^t \cdot \mathbb{P}_M^{\pi_\theta} f(s_t, a_t)$$

$$= \sum_{\tau} \mathbb{P}_M^{\pi_\theta}\Big[\sum_{t=0}^{\infty} r^t f(s_t, a_t)\Big] = \mathbb{E}_{\tau \sim \mathbb{P}_M^{\pi_\theta}}\Big[\sum_{t=0}^{\infty} r^t f(s_t, a_t)\Big]$$

△ The observation and output of problem 5. are shown in the zip. file

"dataset" is a dictionary, observation is nd array.