Naser Kazemi 99102059

Q1

A) The purpose of the first expression $E_{\mu}[Q] - E_{\hat{\pi}_\beta}[Q]$ is to find the μ such that maximizes the $E_{\mu}[Q]$, and the minimizing it, preventing having (s,a) pairs which cause a pick in Q-value. i.e pushing down the Q-values. The $-E_{\hat{\pi}_\beta}[Q]$ is for the opposite, maximizing the Q-value for observed data, i.e pushing up on (s,a) samples. so we just minimize the Q-value for unobserved data.

B) We have

$$\hat{Q}^{k+1} = \min_{Q} \max_{\mu} \alpha \left(E_{s\sim D, a\sim\mu(a|s)}[Q(s,a)] - E_{s\sim D, a\sim\hat{\pi}_\beta(a|s)}[Q(s,a)] \right) + \frac{1}{2} E_{s,a,s'\sim D}\left[\left(Q(s,a) - \hat{\mathcal{B}}^{\pi_k} \hat{Q}^k(s,a) \right)^2 \right] + \mathcal{R}(\mu)$$

So at first we have to find μ such that

$$\max_{\mu} E_{s\sim D, a\sim\mu(a|s)}[Q(s,a) - \log \mu(a|s)]$$

$$\text{s.t.} \quad \sum_a \mu(a|s) = 1 \qquad \text{for each } s.$$

we can solve the above optimization problem for each $s\sim D$ separatly. since each $\mu(\cdot|s)$ is only conditioned on s. then for a fixed s we have the dual problem is:

$$E_{a\sim\mu(a|s)}[Q(s,a) - \log \mu(a|s)] - \lambda \sum_a \mu(a|s) - \lambda$$

so by the KKT conditions we have:

$$\nabla_\mu \sum_a \mu^*(a|s)(Q(s,a) - \log \mu^*(a|s)) - \nabla_\mu \lambda \sum_a \mu(a|s) = 0$$

$$\Rightarrow Q(s,a) - \log \mu^*(a|s) - 1 - \lambda = 0$$

$$\Rightarrow \log \mu^*(a|s) = Q(s,a) - 1 - \lambda$$

$$\Rightarrow \mu(a|s)^* = \frac{\exp(Q,a)}{e^{1+\lambda}}$$

and $\sum_a \mu^*(a|s) = 1 \Rightarrow e^{1+\lambda} = \sum_{a'} \exp(Q,a')$

$$\Rightarrow \mu^*(a|s) = \frac{\exp(Q(s,a))}{\sum_{a'} \exp Q(s,a')}$$

then we will have:

$$CQL = \min_Q \alpha E_{s\sim D}[E_{a\sim\mu^*}[Q(s,a) + \log \sum_{a'} \exp(Q(s,a') - Q(s,a)] - E_{a\sim\hat{\pi}_\beta}[Q(s,a)]] + \dots$$

$$= \min_Q \alpha E_{s\sim D}[\log \sum_a \exp(Q(s,a)) - E_{a\sim\hat{\pi}_\beta}[Q(s,a)]]$$

$$+ \tfrac{1}{2} E_{s,a,s'\sim D}[(Q - \hat{B}^{\pi_k}\hat{Q}^k)^2]$$

Adding this regularizing term, other than simplifing the maximization part, introduces the exploration to our learned policy, avoiding the concentration on large Q-values, leading to a conservative policy.

C) We have

$$CQL = \min_{Q} \alpha \left(E_{s\sim D,\, a\sim \mu^*} [Q(s,a)] + \tfrac{1}{2} E_{s,a,s'\sim D} [(Q - \hat{\mathcal{B}}^{\pi_k} \hat{Q}^k)^2] \right) + \mathcal{R}(\mu^*)$$

for a fixed s, a we have

$$\alpha \mu^*(a|s) Q(s,a) + \tfrac{1}{2} \hat{\pi}_\beta(a|s) \left(Q(s,a) - \hat{\mathcal{B}}^{\pi_k} \hat{Q}^k(s,a)\right)^2$$

$$\xrightarrow{\nabla_Q} \alpha \mu^*(a|s) + \hat{\pi}_\beta(a|s) \left(Q(s,a) - \hat{\mathcal{B}}^{\pi_k} \hat{Q}^k(s,a)\right) = 0$$

$$\Rightarrow \hat{Q}^{k+1}(s,a) \sim Q(s,a) = \hat{\mathcal{B}}^{\pi_k} \hat{Q}^k(s,a) - \alpha \frac{\mu^*(a|s)}{\hat{\pi}_\beta(a|s)}$$

As we see, in each iteration of the Bellman operator, we have $Q^{k+1}(s,a) \leq \hat{\mathcal{B}}^{\pi_k} \hat{Q}^k(s,a)$, and we really are understimating the next Q-value.

D) $\forall_{a,s,a} \in D$, $|\hat{\mathcal{B}}^{\pi} Q(s,a) - \mathcal{B}^{\pi} Q(s,a)| \leq C_\delta(s,a)$. By the above update rule,

$(\mathcal{B}^{\pi} Q = r + \gamma P^{\pi} Q)$

$$\hat{Q}^{\pi}(s,a) \leq \mathcal{B}^{\pi} \hat{Q}^{\pi}(s,a) - \alpha \frac{\mu^*(a|s)}{\hat{\pi}_\beta(a|s)} + C_\delta(s,a)$$

$$\Rightarrow \hat{Q}^{\pi} \leq (I - \gamma P^{\pi})^{-1} \left[R - \alpha \frac{\mu^*}{\hat{\pi}_\beta} + C_\delta \right]$$

↓ Rewards (Q-value function)

$$\Rightarrow \hat{Q}^{\pi}(s,a) \leq Q^{\pi}(s,a) - \alpha \left[(I - \gamma P^{\pi})^{-1} \left[\frac{\mu^*}{\hat{\pi}_\beta} \right] \right](s,a) + \left[(I - \gamma P^{\pi})^{-1} C_\delta \right](s,a)$$

Q2

A) BC only tries to mimic the experts behavior. But GAIL learns the reward function and a policy at least as good as the expert in a robust manner. GAIL also mitigates the compounding error problem. GALI has a better generalization to unseen states, since it focuses on matching the overall distribution of trajectories. GAIL has a better exploration ability due to the entrop regularization term.

B) I IRI, multiple reward functions can describe the same expert policy. This is called the inherent ambiguity in IRL. The maximum entropy principle, states the among the reward functions describing the same optimal policy, the one with the highest trajectory entropy.

The method:

the trajectory distribution is considered to be of the form:

$$P(\tau|R) \propto \exp\left(\sum_{t=0}^{T} R(s_t, a_t)\right)$$

and the objective is to maximize:

$$\max_R \sum_{\tau \in D} \log P(\tau|R) - \lambda \sum P(s,a) R(s,a)$$

This principle ensures that the policy is as random as possible, while remaining consistent with the observed behaviour, avoiding unnecessary assumptions about the expert.

C) We have: $J(\pi_E, f) \equiv E_{\pi_E}[Log(1 - D(s,a))]$ where the D, is the discriminator function and plays the role of "f" in J.

as for the $J(\pi, f) = \lambda \mathcal{H}(\pi) - E_{\pi}[log(D(s,a))]$.

From another point of view we have:

D: as the discriminator, tries to minimize the discrimination for th policy with respect the expert, while maximizing it for the learned policy. This is adversarial part. On the other hand, the π tries to minimize this discrimination, while maximizing $\mathcal{H}(\pi)$. Here, $\mathcal{H}$ addresses the ambiguity problem, which by MaxEnt, we prefer the policy with max $\mathcal{H}(\pi)$.

D) These methods, suffer from the limit in choice of features, and the complexity of solving a more complicated optimization problem.

In contrast, GAIL is more felicible in terms of features. e.g the discriminator D can be a neural network. Also it deals with the discrimination directly.

Adversarial training also improves both policy and discriminator.

Q3

A) Pure model methods suffer from model inaccuracies, with generalizaicon errors for unseen data. This leads to suboptimal policies. Pure environmetal methods require lots of computions whit extensive interaction with the environment. It can be time-costly and sample inefficient.
Hybrid methods, leverage the strength in both. e.g using the model to generate and simulat environment interaction, increasing the sample efficiency and reducing compution costs, while holding to the ground truth environment, mitigating the model inaccuracies over long-term.

B) MBPO integrates model simulaticon rollouces with real-world interactions. It selects random points on the real world trajetories and strats to rollout using the moled, from the selected points, and generate new trajectories. MBPO usually uses rollout with 1-3 steps. longer rollouts lead to accumulated errors by the model, causing unreal hallucination. But shorter rollouts may result in data with less diversity.

C) Since MBPO relies on real-time interactions with the environment and the model simulation, it needs to operate online.

D) MOReL measures the model uncertainty by computing the variance in prediction using an ensemble of learned models. When a model is uncertion about a state-action pair, treats it with caution. As mentioned, a method for measuring the uncertainty, is to use the variane of an ensemble of models on the state-action pairs predictions.

E) COMBO leverages the Model-based Dyna-style algorithm to generate data, and adds it to the total data buffer, which is a mixture of real-world data and the model rollouts. Then CQL is applied to penalize the overstimation of Q-values by enforcing a conservative objective. This is for ensuring the robustness of the model when dealing with the data generated by the Dyna component.

F)

Conservative methods: pros: Robustness to mode inaccuracies. Reduced overstimation. Avoiding risky actions.
cons: computational complexities. Slow convergence Hyperparamer sesitivity.

Uncertainty based methods: pros: Exploration encouragement. Handling uncertainty.
cons: overexploration. complex implementio.

Conservative methods tend to be more exploitive, an robust to inaccuracy, but sensitive to hyper parameter. Uncertainty methotd on the othe hand are explorative, and loss hyperparameter sensitive.

Q4

A) The Hoeffding's inequality states that given a sequence of random variables $X_1, \ldots, X_n$, with $S_n = \sum_{i=1}^{n} X_i$, we have

$$P(|S_n - E[S_n]| \geq t) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

where $\forall i, a_i \leq X_i \leq b_i$.

If X_is are i.i.d, we can rewrite the inequality as

$$P(|1/n \sum_{i=1}^{n} X_i - E[X]| \geq \varepsilon) \leq 2\exp\left(-\frac{2n\varepsilon^2}{(b-a)^2}\right)$$

In our case, since rewards are in $[-1, 1]$, thus $(b-a)^2 = 4$, and we can have

$$P(|1/n \sum_{i=1}^{n} X_i - E[X]| \geq \varepsilon) \leq 2\exp(-n\varepsilon^2/2)$$

Letting $\varepsilon = \sqrt{\frac{2\log(1/\delta)}{n}}$, we obtain:

$$P(|1/n \sum_{i=1}^{n} X_i - E[X]| \geq \sqrt{\frac{2\log(1/\delta)}{n}}) \leq 2\exp(-\log 1/\delta) = 2\delta$$

So

$$P(|1/n \sum_{i=1}^{n} X_i - \mu| \leq \sqrt{\frac{2\log(1/\delta)}{n}}) \geq 1 - 2\delta$$

Here $1/n \sum_{i=1}^{n} X_i \equiv \hat{\mu}_i(t-1)$, $n \equiv T_i(t-1)$ and μ is the true mean reward for action i. Hence:

$$P\left(|\hat{\mu}_i(t-1) - \mu_i| \le \sqrt{\frac{2\log(1/\delta)}{T_i(t-1)}}\right) \ge 1-2\delta$$

So w.h.p $(1-2\delta)$, $\hat{\mu}_i(t-1) + \sqrt{\frac{2\log(1/\delta)}{T_i(t-1)}}$ is an upper bound for μ_i.

B) Let arm 1 have highest reward. Let's define event G_i as

$$G_i = \left\{\mu_1 \le \min_{1\le t\le n} UCB_i(t,\delta)\right\} \cap \left\{\hat{\mu}_{i u_i} + \sqrt{\frac{2}{u_i}\log(1/\delta)} < \mu_1\right\}$$

For some constant u_i.

G_i is the event of not underestimating μ_1, while the UCB for the ith arm is below the mean reward of the optimal arm μ_1.

We can write $E[T_i(n)]$ as:

$$E[T_i(n)] = E[\mathbb{1}(G_i)T_i(n)] + E[\mathbb{1}(G_i^c)T_i(n)]$$

Let's first compute an upper bound for $E[\mathbb{1}(G_i)T_i(n)]$. We first show that when G_i holds, then $T_i(n) \le u_i$. Let G_i hold. Suppose $T_i(n) > u_i$. This means arm i has been played more than u_i times. So there is some round $t \in \{1,\dots,n\}$, which $T_i(t-1) = u_i$ and $A_t = i$. Therefor

$$UCB_i(t-1,\delta) = \hat{\mu}_i(t-1) + \sqrt{\frac{2\log(1/\delta)}{T_i(t-1)}} =$$

$$= \hat{\mu}_{i u_i} + \sqrt{\frac{2\log(1/\delta)}{u_i}}$$

$$< u_1$$

$$< UCB_1(t-1,\delta)$$

So $A_t = \arg\max_j UCB_j(t-1, \delta) \neq i$, which is a contradiction. On the other hand, we have:

$$G_i^c = \underbrace{\{\mu_1 \geq \min_{1 \leq t \leq n} UCB_1(t, \delta)\}}_{A} \cup \underbrace{\{\hat{\mu}_{i u_i} + \sqrt{\frac{2\log(1/\delta)}{u_i}} \geq \mu_1\}}_{B}$$

For A we have: $(T_i(t) < t)$

$$A = \{\mu_1 \geq \min_{1 \leq t \leq n} UCB_1(t, \delta)\} \subset \{\mu_1 \geq \min_{1 \leq t \leq n} \hat{\mu}_{1t} + \sqrt{\frac{2\log(1/\delta)}{t}}\}$$
$$= \bigcup_{1 \leq t \leq n} \{\mu_1 \geq \hat{\mu}_{1t} + \sqrt{\frac{2\log(1/\delta)}{t}}\}$$

By the union bound we have

$$P(\mu_1 \geq \min_t UCB_1(t, t)) \leq P\left(\bigcup_{1 \leq t \leq n} \{\mu_1 \geq \hat{\mu}_{1t} + \sqrt{\frac{2\log(1/\delta)}{t}}\}\right)$$
$$\leq \sum_t P\left(\{\mu_1 \geq \hat{\mu}_{1t} + \sqrt{\frac{2\log(1/\delta)}{t}}\}\right)$$

Lemma 1: Let $X \in [-2, 2]$, then we have:

$$P(X \geq \varepsilon) = P(\exp(tX) \geq \exp(t\varepsilon))$$
$$\leq E[\exp(tX)]\exp(-t\varepsilon) \quad \text{(Markov's inequality)}$$
$$\leq E[\exp(2t - t\varepsilon)] \quad \text{letting } t = \frac{\varepsilon^2}{2(\varepsilon - 2)}$$
$$= \exp(-\varepsilon^2/2)$$

Thus

$$P\left(\{\mu_1 \geq \hat{\mu}_{1t} + \sqrt{\frac{2\log(1/\delta)}{t}}\}\right) \leq \exp\left(-\frac{\sqrt{2\log(1/\delta)}^2}{2}\right) = \delta$$

And

$$P(\mu_1 \geq \min_t UCB_1(t,t)) \leq n\delta$$

As for B, let u_i be chosen large enough so that:

$$\Delta_i - \sqrt{\frac{2\log(1/\delta)}{u_i}} \geq c\Delta_i \qquad c \in (0,1) \qquad (1)$$

then since $\mu_1 = \mu_i + \Delta_i$, then again using lemma 1,

$$P\left(\hat{\mu}_{i u_i} + \sqrt{\frac{2\log(1/\delta)}{u_i}} \geq \mu_1\right) = P\left(\hat{\mu}_{i u_i} - \mu_i + \sqrt{\frac{2\log(1/\delta)}{u_i}} \geq u_1 - \mu_i\right)$$

$$= P\left(\hat{\mu}_{i u_i} - \mu_i \geq \Delta_i - \sqrt{\frac{2\log(1/\delta)}{u_i}}\right)$$

$$\leq P(\hat{\mu}_{i u_i} - \mu_i \geq c\Delta_i)$$

$$\leq \exp\left(\frac{-u_i c^2 \Delta_i^2}{2}\right) \qquad \text{(by just setting } t = \frac{u_i \varepsilon^2}{2(\varepsilon - 2)} \text{ (in lemma 1)}$$

So overall

$$P(G_i^c) \leq n\delta + \exp\left(\frac{-u_i c^2 \Delta_i^2}{2}\right)$$

And

$$E[T_i(n)] \leq u_i + n\left(n\delta + \exp\left(\frac{-u_i c^2 \Delta_i^2}{2}\right)\right)$$

choosing the smallest u_i s.t. (1) holds, we get $u_i = \left\lceil \frac{2\log(1/\delta)}{(1-c^2)\Delta_i^2} \right\rceil$

and $\delta = 1/n^2$, we obtain

$$E[T_i(n)] \leq u_i + 1 + n e^{\frac{-u_i c^2 \Delta_i^2}{2}} = \left\lceil \frac{2\log(1/\delta)}{(1-c^2)\Delta_i^2} \right\rceil + 1 + n e^{\frac{-u_i c^2 \Delta_i^2}{2}}$$

and setting $c = 1/2$, leads to

$$\left\lceil \frac{4\log n}{1/4\,\Delta_i^2} \right\rceil \leq \frac{16\log n}{\Delta_i^2}$$

and

$$n e^{-\frac{u_i c^2 \Delta_i^2}{2}} \le e^{\log n - \frac{u_i c^2 \Delta_i^2}{2}}$$

$$\le n^{1 - \frac{c^2 \Delta_i^2 \, 4 \log n}{2 \log n \, (1-c^2) \Delta_i^2}}$$

$$= n^{1 - \frac{2c^2}{(1-c^2)}}$$

by just setting $c \ge 2$: ≤ 1

So

$$E[T_i(n)] \le 3 + \frac{16 \log n}{\Delta_i^2}$$

c)

Lemma 2: for any policy π, countable action set $\mathcal{A}$ and horizon n, the regret R_n for this policy can be decomposed as

$$R_n = \sum_{a \in \mathcal{A}} \Delta_a E[T_a(n)]$$

Proof: For any fixed t we have $\sum_{a \in \mathcal{A}} \mathbb{1}\{A_t = a\} = 1$. Therefore

$S_n = \sum_t X_t = \sum_t \sum_a X_t \mathbb{1}\{A_t = a\}$. Thus

$$R_n = n\mu^* - E[S_n] = \sum_{t=1}^{n} \sum_{a \in \mathcal{A}} E[(\mu^* - X_t)\mathbb{1}\{A_t = a\}]$$

$$E[(\mu^* - X_t)\mathbb{1}\{A_t = a\}] = E[E[(\mu^* - X_t)\mathbb{1}\{A_t = a\} \mid A_t]]$$

$$= E[\mathbb{1}\{A_t = a\}]E[(\mu^* - X_t) \mid A_t]$$

$$= E[\mathbb{1}\{A_t = a\}](\mu^* - \mu_{A_t=a})$$

$$= E[\mathbb{1}\{A_t = a\}]\Delta_a$$

Therefor

$$R_n = \sum_{a \in A} \sum_{t=1}^{n} E[\mathbb{1}\{A_t = a\}]\Delta_a = \sum_{a \in A} E[T_a(n)]\Delta_a$$

So by lemma 2 we have $R_n = \sum_{i=1}^{K} E[T_i(n)]\Delta_i$. In part B we had $\delta = 1/n^2$. So:

$$R_n = \sum_{i=1}^{K} E[T_i(n)]\Delta_i \leq \sum_{i:1}^{K} \left(3 + \frac{16 \log n}{\Delta_i^2}\right)\Delta_i$$

$$= \sum_{i=1}^{K} 3\Delta_i + \sum_{i:\Delta_i > 0} \frac{16 \log n}{\Delta_i}$$

D) By part B, C and lemma 2 we have:

$$R_n = \sum_{i:1}^{K} \Delta_i E[T_i(n)] = \sum_{i, \Delta_i < \Delta} \Delta_i \cdot E[T_i(n)] + \sum_{i, \Delta_i \geq \Delta} \Delta_i E[T_i(n)]$$

(For some $\Delta > 0$ to be determined later)

$$\leq n\Delta + \sum_{i, \Delta_i \geq \Delta} \left(3\Delta_i + \frac{16 \log n}{\Delta_i}\right)$$

$$\leq n\Delta + 3\sum_{i} \Delta_i + \frac{16K \log n}{\Delta}$$

$$\leq 8\sqrt{nK \log(n)} + 3\sum_{i} \Delta_i$$

By choosing

$$\Delta = \sqrt{\frac{16K \log n}{n}}$$