A. Proof of Theorem 3.1

Before proving the theorem, we introduce some notations and useful lemmas.

Given a pair of π and f, denote random variable $v_i = K\pi(a_h^i|x_h^i)f(x_{h+1}^i) - f(\tilde{x}_{h+1}^i)$, recall the definition of the utility function $u(\pi, f)$:

$$u(\pi, f) = \frac{1}{N}\sum_{i=1}^{N} K\pi(a_h^i|x_h^i)f(x_{h+1}^i) - \frac{1}{N}\sum_{n=1}^{N} f(\tilde{x}_{h+1}^i) = \frac{1}{N}\sum_{n=1}^{N} v_i.$$

Denote $v = \mathbb{E}_{x\sim\nu_h, a\sim\pi(\cdot|x), x'\sim P_{x,a}}[f(x')] - \mathbb{E}_{x\sim\mu^\star_{h+1}}[f(x)]$. It is easy to verify that $\mathbb{E}_i v_i = v$. We also have $|v_i - v| \leq 4K$. We can further bound the variance of $v_i - v$ as:

$$\begin{aligned}
\mathrm{Var}_i(v_i - v) &= \mathbb{E}_i(v_i - v)^2 = \mathbb{E}_i v_i^2 - v^2 \leq \mathbb{E}_i v_i^2 \\
&= \mathbb{E}_i(K\pi(a_h^i|x_h^i)f(x_{h+1}^i) - f(\tilde{x}_{h+1}^i))^2 \\
&\leq \mathbb{E}_i K^2\pi(a_h^i|x_h^i)f(x_{h+1}^i) - \mathbb{E}_i K\pi(a_h^i|x_h^i)f(x_{h+1}^i)f(\tilde{x}_{h+1}^i) + \mathbb{E}_i f(\tilde{x}_{h+1}^i)^2 \\
&\leq K + 1 + 1 \leq 2K,
\end{aligned}$$

where we used the fact that $|f(x)| \leq 1, \forall x$, $\pi(a|x) \leq 1, \forall x, a$, and the last inequality uses the fact that a_h^i is sampled from a uniform distribution over $\mathcal{A}$. With that, we can apply Bernstein's inequality to $\{v_i\}$ together with a union bound over Π and $\mathcal{F}$, we will have the following lemma:

Lemma A.1. *Given dataset $\mathcal{D} = \{x_h^i, a_h^i, x_{h+1}^i\}$ with $x_h^i \sim \nu_h, a_h^i \sim U(\mathcal{A}), x_{h+1}^i \sim P_{x_h^i, a_h^i}$, and $\mathcal{D}^e = \{\tilde{x}_{h+1}^i\}$ with $\tilde{x}_{h+1}^i \sim \mu^\star_{h+1}$, for any pair $\pi \in \Pi, f \in \mathcal{F}$, with probability at least $1 - \delta$,*

$$\begin{aligned}
&\left|\left(\frac{1}{N}\sum_{i=1}^{N} K\pi(a_h^i|x_h^i)f(x_{h+1}^i) - \frac{1}{N}\sum_{i=1}^{N} f(\tilde{x}_{h+1}^i)\right) - \left(\mathbb{E}_{(x,a,x')\sim\nu_h\pi P^\star}[f(x')] - \mathbb{E}_{x\sim\mu^\star_{h+1}}[f(x)]\right)\right| \\
&\leq 4\sqrt{\frac{2K\log(2|\Pi||\mathcal{F}|/\delta)}{N}} + \frac{8K\log(2|\Pi||\mathcal{F}|/\delta)}{N}.
\end{aligned} \tag{7}$$

Let us define two loss functions for π and f:

$$\begin{aligned}
&\ell_t(\pi) = (1/N)\sum_{i=1}^{N} K\pi(a_h^i|x_h^i)f^t(x_{h+1}^i) \\
&c_t(f) = (1/N)\sum_{i=1}^{N} K\pi^t(a_h^i|x_h^i)f(x_{h+1}^i) - (1/N)\sum_{n=1}^{N} f(\tilde{x}_{h+1}^i).
\end{aligned}$$

For any $f, g : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \to \mathbb{R}$, define $\langle f, g\rangle = \mathbb{E}_{(x,a)\sim\mathcal{D}_{x,a}} f(x,a)g(x,a)$, where we overload the notation and denote $\mathcal{D}$ as the empirical distribution over the dataset $\mathcal{D}$ (i.e., put probability $1/|\mathcal{D}|$ over each data point in $\mathcal{D}$), and $\mathcal{D}_{x,a}$ as the marginal distribution over x, a. With this notation, we can see that $\ell_t(\pi)$ can be written as a linear functional with respect to π:

$$\ell_t(\pi) = \langle \pi, Kf^t\rangle, \tag{8}$$

where Kf_t is defined such that $Kf^t(x,a) = K\sum_{i=1}^{N}\mathbf{1}[x = x_h^i, a = a_h^i]f^t(x_{h+1}^i)$. Under this definition of inner product, we have:

$$\max_{\pi}\|\pi\| \leq 1, \quad \max\|Kf^t\| \leq K.$$

It is easy to verify that Algorithm 1 is running Best Response on loss $\{c_t(f)\}_t$ and running FTRL on loss $\{\ell_t(\pi)\}_t$. Using the no-regret guarantee from FTRL, for $\{\pi^t\}$, we have:

$$\frac{1}{T}\sum_{t=1}^{T}\ell_t(\pi^t) - \min_{\pi\in\Pi}\frac{1}{T}\sum_{t=1}^{T}\ell_t(\pi) \leq \frac{K}{\sqrt{T}}. \tag{9}$$

Denote $\hat{\pi}^\star$ and $\hat{f}^\star$ as the minimizer and maximizer of Eqn 2, i.e.,

$$(\hat{\pi}^\star, \hat{f}^\star) = \arg\min_{\pi\in\Pi}\arg\max_{f\in\mathcal{F}} \left(\frac{1}{N}\sum_{i=1}^{N} K\pi(a_h^i|x_h^i) f(x_{h+1}^i) - \frac{1}{N}\sum_{i=1}^{N} f(\tilde{x}_{h+1}^i)\right). \tag{10}$$

The following lemma quantifies the performance of $\bar{\pi} = \sum_{t=1}^{T}\pi^t/T$ and $\bar{f} = \sum_{t=1}^{T} f^t/T$:

Lemma A.2. *Denote $\bar{\pi} = \sum_{t=1}^{T}\pi^t/T$ and $\bar{f}^\star = \max_{f\in\mathcal{F}}\left(\mathbb{E}_{(x,a,x')\sim\nu_n\bar{\pi}P^\star}[f(x')] - \mathbb{E}_{x\sim\mu_{h+1}^\star}[f(x)]\right)$. We have:*

$$\frac{1}{N}\sum_{i=1}^{N} K\bar{\pi}(a_h^i|x_h^i)\bar{f}^\star(x_{h+1}^i) - \frac{1}{N}\sum_{i=1}^{T}\bar{f}^\star(\tilde{x}_{h+1}^i) \leq \frac{1}{N}\sum_{i=1}^{N} K\hat{\pi}^\star(a_h^i|x_h^i)\hat{f}^\star(x_{h+1}^i) - \frac{1}{N}\sum_{i=1}^{N}\hat{f}^\star(\tilde{x}_{h+1}^i) + \frac{K}{\sqrt{T}},$$

where $\hat{\pi}^\star$, $\hat{f}^\star$ is defined in (10).

Proof. Using the definition of ℓ_t and the no-regret property on $\{\pi_t\}$, we have:

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{N}\sum_{i=1}^{N} K\pi^t(a_h^i|x_h^i) f^t(x_{h+1}^i) - \frac{1}{N}\sum_{i=1}^{N} f^t(\tilde{x}_{h+1}^i)\right) \\
&\leq \min_{\pi\in\Pi}\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{N}\sum_{i=1}^{N} K\pi(a_h^i|x_h^i) f^t(x_{h+1}^i) - \frac{1}{N}\sum_{i=1}^{N} f^t(\tilde{x}_{h+1}^i)\right) + \frac{K}{\sqrt{T}}.
\end{aligned}$$

Since $f^t = \arg\max_{f\in\mathcal{F}} c_t(f)$, we have:

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N} K\bar{\pi}(a_h^i|x_h^i)\bar{f}^\star(x_{h+1}^i) - \frac{1}{N}\sum_{i=1}^{N}\bar{f}^\star(\tilde{x}_{h+1}^i) = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{N}\sum_{i=1}^{N} K\pi^t(a_h^i|x_h^i)\bar{f}^\star(x_{h+1}^i) - \frac{1}{N}\sum_{i=1}^{N}\bar{f}^\star(\tilde{x}_{h+1}^i)\right) \\
&\leq \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{N}\sum_{i=1}^{N} K\pi^t(a_h^i|x_h^i) f^t(x_{h+1}^i) - \frac{1}{N}\sum_{i=1}^{N} f^t(\tilde{x}_{h+1}^i)\right)
\end{aligned}$$

We also have:

$$\begin{aligned}
&\min_{\pi\in\Pi}\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{N}\sum_{i=1}^{N} K\pi(a_h^i|x_h^i) f^t(x_{h+1}^i) - \frac{1}{N}\sum_{i=1}^{N} f^t(\tilde{x}_{h+1}^i)\right) \\
&\leq \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{N}\sum_{i=1}^{N} K\hat{\pi}^\star(a_h^i|x_h^i) f^t(x_{h+1}^i) - \frac{1}{N}\sum_{i=1}^{N} f^t(\tilde{x}_{h+1}^i)\right) \\
&\leq \max_{f\in\{f^1,\ldots,f^T\}}\left(\frac{1}{N}\sum_{i=1}^{N} K\hat{\pi}^\star(a_h^i|x_h^i) f(x_{h+1}^i) - \frac{1}{N}\sum_{i=1}^{N} f(\tilde{x}_{h+1}^i)\right) \\
&\leq \left(\frac{1}{N}\sum_{i=1}^{N} K\hat{\pi}^\star(a_h^i|x_h^i)\hat{f}^\star(x_{h+1}^i) - \frac{1}{N}\sum_{i=1}^{N}\hat{f}^\star(\tilde{x}_{h+1}^i)\right),
\end{aligned}$$

where the first inequality uses the definition of $\min_{\pi\in\Pi}$, the second inequality uses the fact that the maximum is larger than the average, and the last inequality uses the fact that $\hat{f}^\star$ is the maximizer with respect to $\hat{\pi}^\star$.

Combining the above results, we have:

$$\frac{1}{N}\sum_{i=1}^{N} K\bar{\pi}(a_h^i|x_h^i)\bar{f}^\star(x_{h+1}^i) - \frac{1}{N}\sum_{i=1}^{N}\bar{f}^\star(\tilde{x}_{h+1}^i) \leq \frac{1}{N}\sum_{i=1}^{N} K\hat{\pi}^\star(a_h^i|x_h^i)\hat{f}^\star(x_{h+1}^i) - \frac{1}{N}\sum_{i=1}^{N}\hat{f}^\star(\tilde{x}_{h+1}^i) + \frac{K}{\sqrt{T}}.$$

□

Denote $\pi^\star$ and $f^\star$ as

$$\pi^\star, f^\star = \arg\min_{\pi\in\Pi} \arg\max_{f\in\mathcal{F}} \left(\mathbb{E}_{x\sim v, a\sim\pi, x'\sim P_{x,a}}[f(x')] - \mathbb{E}_{x\sim\mu^{\pi^\star}_{h+1}}[f(x)]\right) \tag{11}$$

Now we are ready to prove Theorem 3.1

Proof of Theorem 3.1. Denote $C_N = 4\sqrt{\frac{2K\log(2|\Pi||\mathcal{F}|/\delta)}{N}} + \frac{8K\log(2|\Pi||\mathcal{F}|/\delta)}{N}$. First, using the concentration result from Lemma A.1, we have:

$$\begin{aligned}
&\left|\left(\frac{1}{N}\sum_{i=1}^{N} K\bar{\pi}(a_h^i|x_h^i)\bar{f}^\star(x_{h+1}^i) - \frac{1}{N}\sum_{i=1}^{T}\bar{f}^\star(\tilde{x}_{h+1}^i)\right) - \left(\mathbb{E}_{(x,a,x')\sim\nu_h\bar{\pi}P^\star}[\bar{f}^\star(x')] - \mathbb{E}_{x\sim\mu^\star_{h+1}}[\bar{f}^\star(x)]\right)\right| \\
&\leq \frac{1}{T}\sum_{t=1}^{T}\left|\left(\frac{1}{N}\sum_{i=1}^{N} K\pi^t(a_h^i|x_h^i)\bar{f}^\star(x_{h+1}^i) - \frac{1}{N}\sum_{i=1}^{T}\bar{f}^\star(\tilde{x}_{h+1}^i)\right) - \left(\mathbb{E}_{(x,a,x')\sim\nu_h\pi^t P^\star}[\bar{f}^\star(x')] - \mathbb{E}_{x\sim\mu^\star_{h+1}}[\bar{f}^\star(x)]\right)\right| \\
&\leq \frac{1}{T}\sum_{t=1}^{T}(C_N) = C_N.
\end{aligned}$$

On the other hand, for $\pi^\star$, $f^\star$, we have:

$$\left|\left(\frac{1}{N}\sum_{i=1}^{N} K\pi^\star(a_h^i|x_h^i)f^\star(x_{h+1}^i) - \frac{1}{N}\sum_{i=1}^{N}f^\star(\tilde{x}_{h+1}^i)\right) - \left(\mathbb{E}_{(x,a,x')\sim\nu_h\pi^\star P^\star}[f^\star(x')] - \mathbb{E}_{x\sim\mu^\star_{h+1}}[f^\star(x)]\right)\right| \tag{12}$$
$$\leq C_N. \tag{13}$$

Define $\hat{f}' = \max_{f\in\mathcal{F}}(\frac{1}{N}\sum_{i=1}^{N}K\pi^\star(a_h^i|x_h^i)f(x_{h+1}^i) - \frac{1}{N}\sum_{i=1}^{N}f(\tilde{x}_{h+1}^i))$. Combine the above inequalities together, we have:

$$\begin{aligned}
&\max_{f\in\mathcal{F}}\mathbb{E}_{(x,a,x')\sim\nu_h\bar{\pi}P^\star}[f(x')] - \mathbb{E}_{x\sim\mu^\star_{h+1}}[f(x)] = \mathbb{E}_{(x,a,x')\sim\nu_h\bar{\pi}P^\star}[\bar{f}^\star(x')] - \mathbb{E}_{x\sim\mu^\star_{h+1}}[\bar{f}^\star(x)] \\
&\leq \frac{1}{N}\sum_{i=1}^{N}K\bar{\pi}(a_h^i|x_h^i)\bar{f}^\star(x_{h+1}^i) - \frac{1}{N}\sum_{i=1}^{T}\bar{f}^\star(\tilde{x}_{h+1}^i) + C_N \\
&\leq \frac{1}{N}\sum_{i=1}^{N}K\hat{\pi}^\star(a_h^i|x_h^i)\hat{f}^\star(x_{h+1}^i) - \frac{1}{N}\sum_{i=1}^{N}\hat{f}^\star(\tilde{x}_{h+1}^i) + \frac{K}{\sqrt{T}} + C_N \\
&\leq \frac{1}{N}\sum_{i=1}^{N}K\pi^\star(a_h^i|x_h^i)\hat{f}'(x_{h+1}^i) - \frac{1}{N}\sum_{n=1}^{N}\hat{f}'(\tilde{x}_{h+1}^i) + \frac{K}{\sqrt{T}} + C_N \\
&\leq \mathbb{E}_{(x,a,x')\sim\nu_h\pi^\star P^\star}[\hat{f}'(x')] - \mathbb{E}_{x\sim\mu^\star_{h+1}}[\hat{f}'(x)] + 2C_N + \frac{K}{\sqrt{T}} \\
&\leq \mathbb{E}_{(x,a,x')\sim\nu_h\pi^\star P^\star}[f^\star(x')] - \mathbb{E}_{x\sim\mu^\star_{h+1}}[f^\star(x)] + 2C_N + \frac{K}{\sqrt{T}},
\end{aligned}$$

where the first equality uses the definition of $\bar{f}^\star$, the second inequality uses Lemma A.2, the third inequality uses the fact that $\hat{\pi}^\star$ and $\hat{f}^\star$ are the min-max solution of (10), and the fifth inequality uses the fact that $f^\star$ is the maximizer of (11) given $\pi^\star$. Hence, we prove the theorem.

□

B. Proof of Theorem 3.2

Lemma B.1. *There exists a distribution $D \in \Delta(\mathcal{X})$, such that for any two datasets $S_1 = \{x_1, \ldots, x_M\}$ and $S_2 = \{x'_1, \ldots, x'_M\}$ where x_i and x'_i are drawn i.i.d from D, as long as $M = O(\log(|\mathcal{X}|))$, then:*

$$\lim_{|\mathcal{X}|\to\infty}\Pr\left([S_1\cap S_2 = \emptyset]\right) = 1.$$

Proof. We simply set D to be a uniform distribution of $\mathcal{X}$. Denote $|\mathcal{X}| = N$, and $M = O(\log(N))$. The probability of S_1 and S_2 does not have any overlap samples can be easily computed as:

$$\mathrm{P}(S_1 \cap S_2 = \emptyset) \geq \mathrm{P}(S_1 \cap S_2 = \emptyset \text{ and } S_1 \text{ does not have repeated samples}).$$

Note that the probability that S_1 does not have repeated samples can be computed as:

$$\mathrm{P}(S_1 \text{ does not have repeated samples}) = (1 - 1/N)(1 - 2/N)(1 - (M-1)/N).$$

When $N \to \infty$ and $M = O(\log N)$, we have:

$$\lim_{N\to\infty} \mathrm{P}(S_1 \text{ does not have repeated samples}) = 1.$$

Now, conditioned on the event that S_1 does not contain repeated samples, we have:

$$\mathrm{P}(S_1 \cap S_2 = \emptyset) = (1 - M/N)^M = (1 - M/N)^{(N/M)*(M^2/N)}$$

Take $N \to \infty$, we know that $\lim_{x\to\infty}(1 - 1/x)^x = 1/e$ and $\lim_{N\to\infty} M^2/N = 0$, hence we have:

$$\lim_{N\to\infty}(1 - M/N)^K = \lim_{N\to\infty}(1 - M/N)^{(N/M)(M^2/N)} = \lim_{N\to\infty}(1/e)^{M^2/N} = 1.$$

Hence we prove the lemma by coming two results above. □

We construct the MDP using the above lemma. The MDP has $H = 2$, two actions $\{a, a^\star\}$, and the initial distribution ρ assigns probability one to a unique state $\hat{x} \in \mathcal{X}$. The expert policy $\pi^\star$ is designed to be $\pi^\star(a^\star|\hat{x}) = 1$, i.e., the expert's action at time step $h = 1$ is $a^\star$. We split the state space $\mathcal{X}$ into half and half, denoted as $\mathcal{X}_1$ and $\mathcal{X}_2$, such that $\mathcal{X}_1 \cap \mathcal{X}_2 = \emptyset$ and $|\mathcal{X}_1| = |\mathcal{X}_2| = N/2$. We design the MDP's dynamics such that $P(\cdot|\hat{x}, a)$ assigns probability $2/N$ to each state in $\mathcal{X}_1$ and assigns probability 0 to any other state in $\mathcal{X}_2$. We design $P(\cdot|\hat{x}, a^\star)$ such that it assigns probability $2/N$ to each state in $\mathcal{X}_2$ and zero to each state in $\mathcal{X}_1$.

Denote $\mathcal{D}^\star = \{\tilde{x}_2^{(i)}\}_{i=1}^M$ as the states generated from the expert by executing $a^\star$ at $\hat{x}$. For any two policies π and π', such that $\pi(a^\star|\hat{x}) = 1$ and $\pi'(a|\hat{x}) = 1$, denote $\mathcal{D} = \{x_2^i\}_{i=1}^M$ as the dataset sampled from executing π at $\hat{x}$ M many times, and $\mathcal{D}' = \{x_2'^{(i)}\}_{i=1}^M$ as the dataset sampled from executing π' at $\hat{x}$ M many times. From Lemma B.1, we know that

$$\lim_{N\to\infty} \mathrm{P}(\mathcal{D} \cap \mathcal{D}^\star = \emptyset) = 1, \quad \lim_{N\to\infty} \mathrm{P}(\mathcal{D}^* \cap \mathcal{D}' = \emptyset) = 1.$$

Hence asymptotically either $\mathcal{D}$ nor $\mathcal{D}'$ will overlap with $\mathcal{D}^\star$, unless $M = \Omega(\mathrm{poly}(N)) = \Omega(\mathrm{poly}(|\mathcal{X}|))$.

C. Proof of Theorem 3.3

We first present some extra notations and useful lemmas below.

Lemma C.1 (Performance Difference Lemma (Kakade & Langford, 2002))**.** *Consider a policy* $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_H\}$ *and* $\boldsymbol{\pi}^\star = \{\pi_1^\star, \ldots, \pi_H^\star\}$*. We have:*

$$J(\boldsymbol{\pi}) - J(\boldsymbol{\pi}^\star) = \sum_{h=1}^H \mathbb{E}_{x\sim\mu_h^{\boldsymbol{\pi}}}\left[\mathbb{E}_{a\sim\pi_h(\cdot|x)} Q_h^\star(x,a) - V_h^\star(x)\right].$$

Note that under our setting, i.e., the cost function does not depend on actions, the above equation can be simplified to:

$$J(\boldsymbol{\pi}) - J(\boldsymbol{\pi}^\star) = \sum_{h=1}^H \mathbb{E}_{x\sim\mu_h^{\boldsymbol{\pi}}}\left[\mathbb{E}_{a\sim\pi_h(\cdot|x)} Q_h^\star(x,a) - V_h^\star(x)\right] \tag{14}$$

$$= \sum_{h=1}^H \mathbb{E}_{x\sim\mu_h^{\boldsymbol{\pi}}}\left[\mathbb{E}_{a\sim\pi_h, x'\sim P_{x,a}}\left[V_{h+1}^\star(x')\right] - \mathbb{E}_{a\sim\pi_h^\star, x'\sim P_{x,a}}\left[V_{h+1}^\star(x')\right]\right], \tag{15}$$

where we use Bellman equations, i.e., $Q^\star_h(x,a) = c(x) + \mathbb{E}_{x'\sim P_{x,a}} V^\star_{h+1}(x')$ and $V^\star_h(x) = c(x) + \mathbb{E}_{a\sim \pi^\star_h, x'\sim P_{x,a}} V^\star_{h+1}(x')$.[7]

Note that for any h, we have:

$$
\begin{aligned}
&\left|\mathbb{E}_{x\sim\mu^{\boldsymbol{\pi}}_h}\left[\mathbb{E}_{a\sim\pi_h, x'\sim P_{x,a}}\left[V^\star_{h+1}(x')\right] - \mathbb{E}_{a\sim\pi^\star_h, x'\sim P_{x,a}}\left[V^\star_{h+1}(x')\right]\right]\right| \\
&\leq \left|\mathbb{E}_{x\sim\mu^{\boldsymbol{\pi}}_h}\left[\mathbb{E}_{a\sim\pi_h, x'\sim P_{x,a}}\left[V^\star_{h+1}(x')\right]\right] - \mathbb{E}_{x\sim\mu^\star_h}\mathbb{E}_{a\sim\pi^\star_h, x'\sim P_{x,a}}\left[V^\star_{h+1}(x')\right]\right| \\
&\quad + \left|\mathbb{E}_{x\sim\mu^\star_h}\mathbb{E}_{a\sim\pi^\star_h, x'\sim P_{x,a}}\left[V^\star_{h+1}(x')\right] - \mathbb{E}_{x\sim\mu^{\boldsymbol{\pi}}_h}\mathbb{E}_{a\sim\pi^\star_h, x'\sim P_{x,a}}\left[V^\star_{h+1}(x')\right]\right| \\
&\leq d_{\mathcal{F}_{h+1}}(\pi_h|\mu^{\boldsymbol{\pi}}_h, \mu^\star_{h+1}) + \left|\mathbb{E}_{x\sim\mu^\star_h}\mathbb{E}_{a\sim\pi^\star_h, x'\sim P_{x,a}}\left[V^\star_{h+1}(x')\right] - \mathbb{E}_{x\sim\mu^{\boldsymbol{\pi}}_h}\mathbb{E}_{a\sim\pi^\star_h, x'\sim P_{x,a}}\left[V^\star_{h+1}(x')\right]\right| \\
&\leq d_{\mathcal{F}_{h+1}}(\pi_h|\mu^{\boldsymbol{\pi}}_h, \mu^\star_{h+1}) + \Delta_h + 2\epsilon_{\mathrm{be}}, \qquad (16)
\end{aligned}
$$

where the first inequality comes from the triangle inequality, the second inequality comes from the fact that $V^\star_{h+1} \in \mathcal{F}_{h+1}$, and in the third inequality, we denote $\Delta_h = \max_{f\in\mathcal{F}} \left|\mathbb{E}_{x\sim\mu^\star_h}[f(x)] - \mathbb{E}_{x\sim\mu^{\boldsymbol{\pi}}_h}[f(x)]\right|$, and ϵ_{be} is introduced because $\Gamma_h V^\star_{h+1}$ might not in $\mathcal{F}_h$.

Now we are ready to prove the main theorem.

Proof of Theorem 3.3. We consider the h'th iteration. Let us denote $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_{h-1}\}$ and $\mu^{\boldsymbol{\pi}}_h$ as the observation distribution at time step h of following policies $\boldsymbol{\pi}$ starting from the initial distribution ρ. Denote $\mu^\star_{h+1}$ as the observation distribution of the expert policy $\boldsymbol{\pi}^\star$ at time step $h+1$ starting from the initial distribution ρ. Note that the dataset $\{\tilde{x}^{(i)}_{h+1}\}_{i=1}^n$ is generated from distribution $\mu^\star_{h+1}$. The data at $\mathcal{D}$ is generated i.i.d by first drawing sample $x^{(i)}_h$ from $\mu^{\boldsymbol{\pi}}_h$ (i.e., executing $\pi_1, \ldots \pi_{h-1}$), and then sample action $a^{(i)}_h \sim U(\mathcal{A})$, and then sample $x^{(i)}_{h+1}$ from the real system $P_{x^{(i)}_h, a^{(i)}_h}$.

Mapping the above setup to the setup in Theorem 3.1, i.e., set

$$
\nu_h = \mu^{\boldsymbol{\pi}}_h, \;\; T = \Theta(4K^2/\epsilon^2), \;\; N = \Theta(K\log(|\Pi||\mathcal{F}|/\delta)/\epsilon^2),
$$

Algorithm 1 will output a policy π_h such that with probability at least $1-\delta$:

$$
d_{\mathcal{F}_{h+1}}(\pi_h|\mu^{\boldsymbol{\pi}}_h, \mu^\star_{h+1}) \leq \min_{\pi\in\Pi_h} d_{\mathcal{F}_{h+1}}(\pi|\mu^{\boldsymbol{\pi}}_h, \mu^\star_{h+1}) + \epsilon.
$$

Recall the definition of refined inherent Bellman Error ϵ'_{be} with respect to $\mathcal{F}_h$ and $\boldsymbol{\pi}^\star$:

$$
\epsilon'_{\mathrm{be},h} = \max_{g\in\mathcal{F}_{h+1}} \min_{f\in\mathcal{F}_h} \|f - \Gamma_h g\|_{(\mu^{\boldsymbol{\pi}}_h + \mu^\star_h)/2}.
$$

Denote $\hat{f}$ as:

$$
\hat{f} = \arg\max_{f\in\mathcal{F}_{h+1}} \left(\mathbb{E}_{x\sim\mu^{\boldsymbol{\pi}}_h}\left[\mathbb{E}_{a\sim\pi^\star_h(\cdot|x), x'\sim P_{x,a}}[f(x')]\right] - \mathbb{E}_{x\sim\mu^\star_h}\left[\mathbb{E}_{a\sim\pi^\star_h(\cdot|x), x'\sim P_{x,a}}[f(x')]\right]\right),
$$

and $\hat{g}$ as:

$$
\hat{g} = \arg\min_{g\in\mathcal{F}_h} \|g - \Gamma_h \hat{f}\|_{(\mu^{\boldsymbol{\pi}}_h + \mu^\star_h)/2}
$$

[7] Note that here we actually prove the theorem under a more general setting where we could have cost functions at any time step h.

Now we upper bound $\min_{\pi\in\Pi_h} d_{\mathcal{F}_{h+1}}(\pi|\mu_h^{\boldsymbol{\pi}}, \mu_{h+1}^\star)$ as follows.

$$\begin{aligned}
&\min_{\pi\in\Pi_h} d_{\mathcal{F}_{h+1}}(\pi|\mu_h^{\boldsymbol{\pi}}, \mu_{h+1}^\star) \leq d_{\mathcal{F}_{h+1}}(\pi_h^\star|\mu_h^{\boldsymbol{\pi}}, \mu_{h+1}^\star)\\
&= \max_{f\in\mathcal{F}_{h+1}} \left|\mathbb{E}_{x\sim\mu_h^{\boldsymbol{\pi}}, a\sim\pi_h^\star, x'\sim P_{x,a}} f(x') - \mathbb{E}_{x\sim\mu_h^\star, a\sim\pi_h^\star, x'\sim P_{x,a}} f(x')\right|\\
&= \max_{f\in\mathcal{F}_{h+1}} \left|\mathbb{E}_{x\sim\mu_h^{\boldsymbol{\pi}}}\left[\mathbb{E}_{a\sim\pi_h^*(\cdot|x), x'\sim P_{x,a}}[f(x')]\right] - \mathbb{E}_{x\sim\mu_h^\star}\left[\mathbb{E}_{a\sim\pi_h^*(\cdot|x), x'\sim P_{x,a}}[f(x')]\right]\right|\\
&= \left|\mathbb{E}_{x\sim\mu_h^{\boldsymbol{\pi}}}\left[\mathbb{E}_{a\sim\pi_h^\star(\cdot|x), x'\sim P_{x,a}}[\hat{f}(x')]\right] - \mathbb{E}_{x\sim\mu_h^\star}\left[\mathbb{E}_{a\sim\pi_h^\star(\cdot|x), x'\sim P_{x,a}}[\hat{f}(x')]\right]\right|\\
&\leq \left|\mathbb{E}_{x\sim\mu_h^{\boldsymbol{\pi}}}[\hat{g}(x)] - \mathbb{E}_{x\sim\mu_h^\star}[\hat{g}(x)]\right| + \left|\mathbb{E}_{x\sim\mu_h^{\boldsymbol{\pi}}}[\hat{g}(x) - \mathbb{E}_{a\sim\pi_h^\star, x'\sim P_{x,a}}\hat{f}(x')]\right| + \left|\mathbb{E}_{x\sim\mu_h^\star}[\hat{g}(x) - \mathbb{E}_{a\sim\pi_h^\star, x'\sim P_{x,a}}\hat{f}(x')]\right|\\
&\leq \max_{f\in\mathcal{F}_h}\left|\mathbb{E}_{x\sim\mu_h^{\boldsymbol{\pi}}}[f(x)] - \mathbb{E}_{x\sim\mu_h^\star}[f(x)]\right| + 2\mathbb{E}_{x\sim(\mu_h^{\boldsymbol{\pi}}+\mu_h^\star)/2}\left[|\hat{g}(x) - \mathbb{E}_{a\sim\pi_h^\star, x'\sim P_{x,a}}\hat{f}(x')|\right]\\
&\leq \max_{f\in\mathcal{F}_h}\left|\mathbb{E}_{x\sim\mu_h^{\boldsymbol{\pi}}}[f(x)] - \mathbb{E}_{x\sim\mu_h^\star}[f(x)]\right| + 2\epsilon'_{\text{be}}\\
&= \Delta_h + 2\epsilon'_{\text{be}},
\end{aligned}$$

where the first inequality comes from the realizable assumption that $\pi_h^\star \in \Pi$, the second inequality comes from an application of triangle inequality, and the third inequality comes from the definition of ϵ_{be} and the fact that $\hat{g} \in \mathcal{F}_h$.

After learn π_h, $\boldsymbol{\pi}$ is updated to $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_h\}$. For Δ_{h+1}, we have:

$$\begin{aligned}
\Delta_{h+1} &= \max_{f\in\mathcal{F}_{h+1}}\left|\mathbb{E}_{x\sim\mu_{h+1}^{\boldsymbol{\pi}}}[f(x)] - \mathbb{E}_{x\sim\mu_{h+1}^\star}[f(x)]\right|\\
&= \max_{f\in\mathcal{F}_{h+1}}\left|\mathbb{E}_{x\sim\mu_h^{\boldsymbol{\pi}}, a\sim\pi_h, x'\sim P_{x,a}}[f(x')] - \mathbb{E}_{x\sim\mu_h^\star, a\sim\pi_h^\star, x'\sim P_{x,a}}[f(x')]\right|\\
&= d_{\mathcal{F}_{h+1}}(\pi_h|\mu_h^{\boldsymbol{\pi}}, \mu_{h+1}^\star) \leq \min_{\pi\in\Pi_h} d_{\mathcal{F}_{h+1}}(\pi|\mu_h^{\boldsymbol{\pi}}, \mu_{h+1}^\star) + O(\epsilon) \leq \Delta_h + 2\epsilon'_{\text{be}} + O(\epsilon).
\end{aligned}$$

Define $\Delta_0 = \max_f |\mathbb{E}_{x\sim\rho}[f(x)] - \mathbb{E}_{x\sim\rho}[f(x)]| = 0$, we have for any h,

$$\Delta_h \leq 2h\epsilon'_{\text{be}} + O(h\epsilon).$$

Now we link Δ_h to the performance of the policy $J(\boldsymbol{\pi})$. From (16), we know that:

$$\begin{aligned}
&\left|\mathbb{E}_{x\sim\mu_h^{\boldsymbol{\pi}}}\left[\mathbb{E}_{a\sim\pi_h, x'\sim P_{x,a}}\left[V_{h+1}^\star(x')\right] - \mathbb{E}_{a\sim\pi_h^\star, x'\sim P_{x,a}}\left[V_{h+1}^\star(x')\right]\right]\right|\\
&\leq d_{\mathcal{F}_{h+1}}(\pi_h|\mu_h^{\boldsymbol{\pi}}, \mu_{h+1}^\star) + \Delta_h + 2\epsilon_{\text{be}}\\
&\leq \Delta_h + O(\epsilon) + 2\epsilon'_{\text{be}} + \Delta_h + 2\epsilon'_{\text{be}} = 2\Delta_h + 4\epsilon'_{\text{be}} + O(\epsilon)\\
&\leq 4h\epsilon'_{\text{be}} + O(2h\epsilon) + 4\epsilon'_{\text{be}} + O(\epsilon) = O(h\epsilon'_{\text{be}}) + O(h\epsilon).
\end{aligned}$$

Using Performance Difference Lemma (Lemma C.1), we know that:

$$\begin{aligned}
J(\boldsymbol{\pi}) - J(\boldsymbol{\pi}^\star) &\leq \sum_{h=1}^{H}\left|\mathbb{E}_{x\sim\mu_h^{\boldsymbol{\pi}}}\left[\mathbb{E}_{a\sim\pi_h, x'\sim P_{x,a}}\left[V_{h+1}^\star(x')\right] - \mathbb{E}_{a\sim\pi_h^\star, x'\sim P_{x,a}}\left[V_{h+1}^\star(x')\right]\right]\right|\\
&\leq \sum_{h=1}^{H} 4h\epsilon'_{\text{be}} + 4\epsilon'_{\text{be}} + O(2h\epsilon) + O(\epsilon)\\
&\leq 4H^2\epsilon'_{\text{be}} + 2H\epsilon'_{\text{be}} + O(2H^2\epsilon) + O(H\epsilon) = O(H^2\epsilon'_{\text{be}}) + O(H^2\epsilon)
\end{aligned}$$

□

D. Proof of Proposition 4.1

We first show the construction of the MDP below. The MDP has horizon H, $2^H - 1$ many states, and two actions $\{l, r\}$ standing for *go left* and *go right* respectively. All states are organized in a perfect balanced binary tree, with 2^{H-1} many leafs at level $h = H$, and the first level $h = 1$ contains only a root. The transition is deterministic such that at any internal state, taking action l leads to the state's left child, and taking action r leads to the state's right child. Each internal node has cost zero, and all leafs will have nonzero cost which we will specify later. Note that in such MDP, any sequence of actions $\{a_1, \dots, a_{H-1}\}$ with $a_i \in \{l, r\}$ deterministically leads to one and only one leaf, and the total cost of the sequence of actions is only revealed once the leaf is reached.

The first part of the proposition is proved by reducing the problem to Best-arm identification in multi-armed bandit (MAB) setting. We use the following lower bound of best-arm identification in MAB from (Krishnamurthy et al., 2016):

Lemma D.1 (Lower bound for best arm identification in stochastic bandits from (Krishnamurthy et al., 2016))**.** *For any $K \geq 2$ and $\epsilon \in (0, \sqrt{1/8}]$, and any best-arm identification algorithm, there exists a multi-armed bandit problem for which the best arm $i^\star$ is ϵ better than all others, but for which the estimate $\hat{i}$ of the best arm must have $\mathrm{P}(\hat{i} \neq i^\star) \geq 1/3$ unless the number of samples collected T is at least $K/(72\epsilon^2)$.*

Given any MAB problem with K arms, without loss of generality, let us assume $K = 2^H - 1$ for some $H \in \mathbb{N}^+$. Any such MAB problem can be reduced to the above constructed binary tree MDP with horizon H, and $2^H - 1$ leafs. Each arm in the original MAB will be encoded by a unique sequence of actions $\{a_h\}_{h=1}^{H-1}$ with $a_h \in \{l, r\}$, and its corresponding leaf. We assign each leaf the cost distribution of the corresponding arm. The optimal policy in the MDP, i.e., the sequence of actions leading to the leaf that has the smallest expected cost, is one-to-one corresponding to the best arm, i.e., the arm that has the smallest expected cost in the MAB. Hence, without any further information about the MDP, any RL algorithm that aims to find the near-optimal policy must suffer the lower bound presented in Lemma D.1, as otherwise one can solve the original MAB by first converting the MAB to the MDP and then running an RL algorithm. Hence, we prove the first part of the proposition.

For the second part, let us denote the sequence of the observations from the expert policy as $\{\tilde{x}_h\}_{h=1}^{H}$, i.e., the sequence of states corresponding to the optimal sequence of actions where the last state $\tilde{x}_H$ has the smallest expected cost. We design an IL algorithm as follows.

We initialize a sequence of actions $\boldsymbol{a} = \emptyset$. At every level h, staring at $h = 1$, we try any sequence of actions with prefix $\boldsymbol{a} \circ l$ ($\boldsymbol{a} \circ a$ means we append action a to end of the sequence $\boldsymbol{a}$), record the observed observation x_{h+1}^l; we then reset and try any sequence of actions with prefix $\boldsymbol{a} \circ r$, and record the observed observation x_{h+1}^l. If $x_{h+1}^l = \tilde{x}_{h+1}$, then we append l to $\boldsymbol{a}$, i.e., $\boldsymbol{a} = \boldsymbol{a} \circ l$, otherwise, we append r, i.e., $\boldsymbol{a} = \boldsymbol{a} \circ r$. We continue the above procedure until $h = H - 1$, and we output the final action sequence $\boldsymbol{a}$.

Due to the deterministic transition, by induction, it is easy to verify that the outputted sequence of actions $\boldsymbol{a}$ is exactly equal to the optimal sequence of actions executed by the expert policy. Note that in each level h, we only generate two trajectories from the MDP. Hence the total number trajectories before finding the optimal sequence of actions is at most $2(H-1)$. Hence we prove the proposition.

E. Reduction to LP

Let us denote a set $\{y_i\}_{i=1}^{2N}$ such that $\{y_1, \dots, y_N\} = \{x_1, \dots, x_N\}$, and $\{y_{N+1} \dots, y_{2N}\} = \{x'_1, \dots, x'_N\}$. Denote $d_{i,j} = \mathcal{D}(y_i, y_j)$ for $i \neq j$, and $c_i = 1/N$ for $i \in [N]$ and $c_i = -1/N$ for $i \in [N+1, 2N]$. We formulate the following LP with $2N$ variables and $O(N^2)$ many constraints:

$$\max_{\alpha_1, \dots, \alpha_{2N}} \sum_{i=1}^{2N} c_i \alpha_i, \quad s.t., \forall i \neq j, -L d_{i,j} \leq \alpha_i - \alpha_j \leq L d_{i,j}, \quad \forall i, -1 \leq \alpha_i \leq 1. \tag{17}$$

Denote the solution of the above LP as $\alpha_i^\star$. We will have the following claim:

Claim E.1 (LP Oracle)**.** *Given $\mathcal{F}$ in (5), $\{x_i\}_{i=1}^N$, and $\{x'_i\}_{i=1}^N$, denote $\{\alpha_i^\star\}_{i=1}^{2N}$ as the solution of the LP from (17), we have:* $\sup_{f \in \mathcal{F}} \left(\sum_{i=1}^N f(x_i)/N - \sum_{i=1}^N f(x'_i)/N \right) = \sum_{i=1}^{2N} c_i \alpha_i^\star$.

Proof of Claim E.1. Given the solutions $\{\alpha_i^\star\}_{i=1}^{2N}$, we first are going to construct a function $\hat{f} : \mathcal{X} \to \mathbb{R}$, such that for any y_i, we have $\hat{f}(y_i) = \alpha_i^\star$, and $\hat{f} \in \mathcal{F}$. Denote $L^\star = \max_{i \neq j} |\alpha_i^\star - \alpha_j^\star| / d_{i,j}$. Note that $L^\star \leq L$. The function $\hat{f}$ is constructed as:

$$\hat{f}(x) = \max\left(-1, \min\left(1, \min_{i\in[2N]} L^{\star}\mathcal{D}(y_i, x) + \alpha_i^{\star}\right)\right)$$

First of all, we show that for any y_i, we have $\hat{f}(y_i) = \alpha_i^{\star}$. For any $j \neq i$, we have:

$$L^{\star}\mathcal{D}(y_j, y_i) + \alpha_j^{\star} \geq |\alpha_j^{\star} - \alpha_i^{\star}| + \alpha_j^{\star} \geq \alpha_i^{\star},$$

where the first inequality uses the definition of $L^{\star}$. Also we know that $-1 \leq \alpha_i^{\star} \leq 1$. Hence we have that for y_i, $\max(-1, \min(1, \min_{j\in[2N]} L^{\star}\mathcal{D}(y_j, y_i) + \alpha_j^{\star})) = \alpha_i^{\star}$.

Now we need to prove that $\hat{f}$ is L-Lipschitz continuous. Note that we just need to prove that $\min_i L^{\star}\mathcal{D}(y_i, x) + \alpha_i^{\star}$ is L-Lipschitz continuous, since for any L-Lipschitz continuous function $f(x)$, we have $\max(1, f(x))$ and $\min(-1, f(x))$ to be L-Lipschitz continuous as well.

Consider any two points x and x' such that $x \neq x'$. Denote $\hat{i}$ as $\arg\min_i L^{\star}\mathcal{D}(y_i, x) + \alpha_i^{\star}$ and $\hat{i}' = \arg\min_i L^{\star}\mathcal{D}(y_i, x') + \alpha_i^{\star}$. We have:

$$\begin{aligned}
&\hat{f}(x) - \hat{f}(x') = L^{\star}\mathcal{D}(y_{\hat{i}}, x) + \alpha_{\hat{i}}^{\star} - (L^{\star}\mathcal{D}(y_{\hat{i}'}, x') + \alpha_{\hat{i}'}^{\star})\\
&\leq L^{\star}\mathcal{D}(y_{\hat{i}'}, x) + \alpha_{\hat{i}'}^{\star} - (L^{\star}\mathcal{D}(y_{\hat{i}'}, x') + \alpha_{\hat{i}'}^{\star})\\
&\leq L^{\star}\mathcal{D}(x, x'),
\end{aligned}$$

where for the first inequality we used the definition of $\hat{i}$, and the second inequality uses the triangle inequality. Similarly, one can show that

$$\hat{f}(x) - \hat{f}(x') \geq -L^{\star}\mathcal{D}(x, x').$$

Combine the above two inequalities and the fact that $L^{\star} \leq L$, we conclude that $\hat{f}$ is L-Lipschitiz continuous.

Now we have constructed $\hat{f}$ such that $\hat{f}(y_i) = \alpha_i^{\star}$ for all $i \in [2N]$ and $\hat{f} \in \mathcal{F}$. Now suppose that there exists a function $f' \in \mathcal{F}$, such that $|\sum_{i=1}^{N} f'(x_i)/N - \sum_{i=1}^{N} f'(x_i')/N| > |\sum_{i=1}^{N} \hat{f}(x_i)/N - \sum_{i=1}^{N} \hat{f}'(x_i')/N|$, then we must have for some $i \in [2N]$, $f'(y_i) \neq \alpha_i^{\star}$. However, since $f' \in \mathcal{F}$, we must have that $\{f'(y_i)\}_{i=1}^{2N}$ satisfies all constrains in the LP in (17). Hence the assumption that $\sum_{i=1}^{2N} c_i f'(y_i) > \sum_{i=1}^{2N} c_i \alpha_i^{\star}$ contradicts to the fact that $\{\alpha_i\}_{i=1}^{2N}$ is the maximum solution of the LP formulation in (17). Hence we prove the claim.

□

F. Proof of Corollary 5.1

Since in this setting, $\mathcal{F}_h$ for all $h \in [H]$ contains infinitely many functions, we need to discretize $\mathcal{F}_h$ before we can apply the proof techniques from the proof of Theorem 3.3. We use covering number.

Denote $\mathcal{N}(\mathcal{X}, \epsilon, \mathcal{D})$ as the ϵ-cover of the metric space $(\mathcal{X}, \mathcal{D})$. Namely, for any $x \in \mathcal{X}$, there exists a $x' \in \mathcal{N}(\mathcal{X}, \epsilon, \mathcal{D})$ such that $D(x', x) \leq \epsilon$. Consider any function class $\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R}, \|f\|_L \leq L, \|f\|_\infty \leq 1\}$ with $L \in \mathbb{R}^+$. Below we construct the ϵ-cover over $\mathcal{F}$.

For any $f \in \mathcal{F}$, denote $\bar{f} \in \mathbb{R}^{|\mathcal{N}(\mathcal{X},\epsilon,\mathcal{D})|}$ with the i-th element $\bar{f}_i$ being the function value $f(x_i)$ measured at the i-th element x_i from $\mathcal{N}(\mathcal{X}, \epsilon, \mathcal{D})$. Hence $\bar{\mathcal{F}} \triangleq \{\bar{f} : f \in \mathcal{F}\} \in \mathbb{R}^{|\mathcal{N}(\mathcal{X},\epsilon,\mathcal{D})|}$, and $\|\bar{f}\|_\infty \leq C$ for any $\bar{f} \in \bar{\mathcal{F}}$. Denote $\bar{\mathcal{N}}(\bar{\mathcal{F}}, \alpha, \|\cdot\|_\infty)$ as the α-cover of $\bar{\mathcal{F}}$. Let us denote the set $\mathcal{N} \triangleq \{f \in \mathcal{F} : \bar{f} \in \bar{\mathcal{N}}(\tilde{\mathcal{F}}, \alpha, \|\cdot\|_\infty)\}$.

Claim F.1. *With the above set up, for $\mathcal{F}$'s $(\alpha + 2L\epsilon)$-cover, i.e., $\mathcal{N}(\mathcal{F}, \alpha + 2L\epsilon, \|f\|_\infty)$, we have*

$$|\mathcal{N}(\mathcal{F}, \alpha + 2L\epsilon, \|\cdot\|_\infty)| \leq \left|\bar{\mathcal{N}}(\bar{\mathcal{F}}, \alpha, \|\cdot\|_\infty)\right| \leq \left(\frac{1}{\alpha}\right)^{|\mathcal{N}(\mathcal{X},\epsilon,\mathcal{D})|}.$$

Proof. By definition, we know that for any $\bar{f} \in \bar{\mathcal{F}}$, we have that there exists a $\bar{f}^{\star} \in \bar{\mathcal{F}}$ such that $\|\bar{f} - \bar{f}^{\star}\|_\infty \leq \alpha$. Now consider $\|f - f^\star\|_\infty$. Denote $x^\star = \arg\max_x |f(x) - f^\star(x)|$ and $x^{\star\prime}$ is its closest point in $\mathcal{N}(\mathcal{X}, \epsilon, \mathcal{D})$. We have:

$$\begin{aligned}
&\sup_x |f(x) - f^\star(x)| = |f(x^\star) - f^\star(x^\star)| \le \left|f(x^\star) - f(x^{\star\prime})\right| + \left|f(x^{\star\prime}) - f^\star(x^\star)\right| \\
&\le \left|f(x^\star) - f(x^{\star\prime})\right| + \left|f(x^{\star\prime}) - f^\star(x^{\star\prime})\right| + \left|f^\star(x^{\star\prime}) - f^\star(x^\star)\right| \\
&\le L\epsilon + \alpha + L\epsilon = 2L\epsilon + \alpha,
\end{aligned}$$

where the last inequality comes from the fact that $f, f^\star$ are L-Lipschitz continuous, $\mathcal{D}(x^\star, x^{\star\prime}) \le \epsilon$, $\|\bar{f} - \bar{f}^\star\|_\infty \le \alpha$ and $x^{\star\prime} \in \mathcal{N}(\mathcal{X}, \epsilon, \mathcal{D})$. Hence, we just identify a subset of $\mathcal{F}$ such that it forms a $\alpha + 2L\epsilon$ cover for $\mathcal{F}$ under norm $\|\cdot\|_\infty$.

Note that $\bar{\mathcal{N}}$ is a α-cover with $\|\cdot\|_\infty$ for $\bar{\mathcal{F}}$ which is a subset of $|\mathcal{N}(\mathcal{X}, \epsilon, \mathcal{D})|$-dimension space. By standard discretization along each dimension, we prove the claim. □

From the above claim, setting up α and ϵ properly, we have:

$$|\mathcal{N}(\mathcal{F}, \epsilon/K, \|\cdot\|_\infty)| \le \left(\frac{K}{\epsilon}\right)^{|\mathcal{N}(\mathcal{X}, \epsilon/(3KL), \mathcal{D})|}.$$

Extending the analysis of Theorem 3.3 simply results extending the concentration result in Lemma A.1. Specifically via Bernstein's inequality and a union bound over $\Pi \times \mathcal{N}(\mathcal{F}, \epsilon/K, \|\cdot\|_\infty)$, we have that for any $\pi \in \Pi$, $\tilde{f} \in \mathcal{N}(\mathcal{F}, \epsilon/K, \|\cdot\|_\infty)$, with probability at least $1 - \delta$,

$$\begin{aligned}
&\left|\left(\frac{1}{N}\sum_{i=1}^N K\pi(a_h^i|x_h^i)\tilde{f}(x_{h+1}^i) - \frac{1}{N}\sum_{i=1}^N \tilde{f}(\tilde{x}_{h+1}^i)\right) - \left(\mathbb{E}_{(x,a,x')\sim\nu_h\pi P^\star}[\tilde{f}(x')] - \mathbb{E}_{x\sim\mu_{h+1}^\star}[\tilde{f}(x)]\right)\right| \\
&\le 4\sqrt{\frac{2K\,|\mathcal{N}(\mathcal{X}, \epsilon/(3KL), \mathcal{D})|\log(2|\Pi|K/\epsilon(\delta))}{N}} + \frac{8K\,|\mathcal{N}(\mathcal{X}, \epsilon/(3KL), \mathcal{D})|\log(2|\Pi|K/(\epsilon\delta))}{N}.
\end{aligned}$$

Now using the fact that $\mathcal{N}(\mathcal{F}, \epsilon/K, \|\cdot\|_\infty)$ is an ϵ-cover under norm $\|\cdot\|_\infty$, we have that for any $\pi \in \Pi$, $f \in \mathcal{F}$, with probability at least $1 - \delta$,

$$\begin{aligned}
&\left|\left(\frac{1}{N}\sum_{i=1}^N K\pi(a_h^i|x_h^i)f(x_{h+1}^i) - \frac{1}{N}\sum_{i=1}^N f(\tilde{x}_{h+1}^i)\right) - \left(\mathbb{E}_{(x,a,x')\sim\nu_h\pi P^\star}[f(x')] - \mathbb{E}_{x\sim\mu_{h+1}^\star}[f(x)]\right)\right| \\
&\le 4\sqrt{\frac{2K\,|\mathcal{N}(\mathcal{X}, \epsilon/(3KL), \mathcal{D})|\log(2|\Pi|3C/\epsilon(\delta))}{N}} + \frac{8K\,|\mathcal{N}(\mathcal{X}, \epsilon/(3KL), \mathcal{D})|\log(2|\Pi|3C/(\epsilon\delta))}{N} + 2\epsilon.
\end{aligned}$$

The rest of the proof is the same as the proof of Theorem 3.3.

G. FAIL in Interactive Setting

Recall that with $\{\pi_1, \dots, \pi_{h-1}\}$ being fixed, we denote ν_h as resulting observation distribution resulting at time step h. The interactiveness comes from the ability we can query expert to generate next observation conditioned on states sampled from ν_h—the states that would be visited by learner at time step h. Let us define $d(\pi|\nu_h, \pi_h^\star)$ as:

$$d_{\mathcal{F}_{h+1}}(\pi|\nu_h, \pi_h^\star) \triangleq \max_{f\in\mathcal{F}_{h+1}} \left(\mathbb{E}_{x\sim\nu_h}\mathbb{E}_{a\sim\pi, x'\sim P_{x,a}}[f(x')] - \mathbb{E}_{x\sim\nu_h}\mathbb{E}_{a\sim\pi^\star, x'\sim P_{x,a}}[f(x')]\right).$$

Note that different from $d_{\mathcal{F}_{h+1}}(\pi|\nu_h, \mu_{h+1}^\star)$, in $d_{\mathcal{F}_{h+1}}(\pi|\nu_h, \pi_h^\star)$, the marginal distributions on x are the same for both π and $\pi_h^\star$ and we directly access $\pi_h^\star$ to generate epxert observations at $h+1$ rather than thorough the expert observation distribution $\mu_{h+1}^\star$. In other words, we use IPM to compare the observation distribution at time step $h+1$ after applying π and the observation distribution at time step $h+1$ after applying $\pi^\star$, *conditioned on the distribution* ν_h *generated by the previously learned policies* $\{\pi_1, \dots, \pi_{h-1}\}$. In Algorithm 5, at every time step h, to find a policy π_h that approximately minimizes $d_{\mathcal{F}_{h+1}}(\pi|\nu_h, \pi_h^\star)$, we replace expectations in $d_{\mathcal{F}_{h+1}}(\pi|\nu_h, \pi_h^\star)$ by proper samples (line **??** and line **??**), and then call Algorithm 1 (line **??**).

Algorithm 3 IFAIL(Π, $\mathcal{F}$, ϵ, n, T)

```
Set $\boldsymbol{\pi} = \emptyset$
for $h = 1$ to $H - 1$ do
    $\mathcal{D} = \emptyset, \tilde{\mathcal{D}} = \emptyset$
    for $i = 1$ to $n$ do
        Reset $x_1^{(i)} \sim \rho$ and from $x_i^{(1)}$ execute $\boldsymbol{\pi} = \{\pi_1, \dots, \pi_{h-1}\}$ to generate $x_h^{(i)}$
        Execute $a \sim U(\mathcal{A})$ to generate $x_{h+1}^{(i)}$ and add it to $\mathcal{D}$
        Reset again $\tilde{x}_1^{(i)} \sim \rho$ and from $\tilde{x}_1^{(i)}$ execute $\boldsymbol{\pi} = \{\pi_1, \dots, \pi_{h-1}\}$ to generate $\tilde{x}_h^{(i)}$
        Ask expert to execute its policy at $\tilde{x}_h^{(i)}$ for one step, observe $\tilde{x}_{h+1}^{(i)}$, and add it to $\tilde{\mathcal{D}}$
    end for
    Set $\pi_h$ to be the return of Algorithm 1 with inputs $\left(\tilde{\mathcal{D}}, \mathcal{D}, \Pi, \mathcal{F}, T\right)$
    Append $\pi_h$ to $\boldsymbol{\pi}$
end for
```

Proof. Recall the definition of $d(\pi|\nu_h, \pi_h^\star)$,

$$d_{\mathcal{F}_{h+1}}(\pi|\nu_h, \pi_h^\star) = \max_{f \in \mathcal{F}_{h+1}} \left(\mathbb{E}_{x \sim \nu_h} \mathbb{E}_{a \sim \pi, x' \sim P_{x,a}}[f(x')] - \mathbb{E}_{x \sim \nu_h} \mathbb{E}_{a \sim \pi^\star, x' \sim P_{x,a}}[f(x')]\right),$$

with ν_h being the distribution over $\mathcal{X}_h$ resulting from executing policies $\{\pi_1, \dots, \pi_{h-1}\}$.

We will use Lemma A.1, Lemma A.2, and Lemma C.1 below.

The Performance Difference Lemma (Lemma C.1) tells us that:

$$\begin{aligned} J(\boldsymbol{\pi}) - J(\boldsymbol{\pi}^\star) &= \sum_{h=1}^{H} \mathbb{E}_{x \sim \mu_h^{\boldsymbol{\pi}}} \left[\mathbb{E}_{a \sim \pi_h, x' \sim P_{x,a}} \left[V_{h+1}^\star(x')\right] - \mathbb{E}_{a \sim \pi_h^\star, x' \sim P_{x,a}} \left[V_{h+1}^\star(x')\right]\right] \\ &\leq \sum_{h=1}^{H} \left|\mathbb{E}_{x \sim \mu_h^{\boldsymbol{\pi}}} \left[\mathbb{E}_{a \sim \pi_h, x' \sim P_{x,a}} \left[V_{h+1}^\star(x')\right] - \mathbb{E}_{a \sim \pi_h^\star, x' \sim P_{x,a}} \left[V_{h+1}^\star(x')\right]\right]\right| \\ &\leq \sum_{h=1}^{H} d_{\mathcal{F}_{h+1}}(\pi_h|\mu_h^{\boldsymbol{\pi}}, \pi_h^\star), \end{aligned} \tag{18}$$

where the last inequality comes from the realizable assumption that $V_h^\star \in \mathcal{F}_h$.

At every time step h, mapping to Theorem 3.1 with $\nu_h = \mu_h^{\boldsymbol{\pi}}, T = 4K^2/\epsilon^2$, $n = K\log(|\Pi||\mathcal{F}|/\delta)/\epsilon^2$, we have that with probability at least $1 - \delta$:

$$d_{\mathcal{F}_{h+1}}(\pi_h|\mu_h^{\boldsymbol{\pi}}, \pi_h^\star) \leq \min_{\pi \in \Pi_h} d_{\mathcal{F}_{h+1}}(\pi|\mu_h^{\boldsymbol{\pi}}\pi_h^\star), +\epsilon.$$

Note that $\min_{\pi \in \Pi_h} d_{\mathcal{F}_{h+1}}(\pi|\mu_h^{\boldsymbol{\pi}}, \pi_h^\star) \leq d_{\mathcal{F}_{h+1}}(\pi_h^\star|\mu_h^{\boldsymbol{\pi}}, \pi_h^\star) = 0$, since $\pi_h^\star \in \Pi_h$ by the realizable assumption. Hence, we have that:

$$d_{\mathcal{F}_{h+1}}(\pi_h|\mu_h^{\boldsymbol{\pi}}, \pi_h^\star) \leq \epsilon.$$

Hence, using (18), and a union bound over all time steps $h \in [H]$, we have that with probability at least $1 - \delta$,

$$J(\boldsymbol{\pi}) - J(\boldsymbol{\pi}^\star) \leq H\epsilon,$$

with $T = 4K^2/\epsilon^2$, and $N = K\log(H|\Pi||\mathcal{F}|/\delta)/\epsilon^2$. Since in every round h, we need to draw N many trajectories, hence, the total number of trajectories we need is at most $HK\log(H|\Pi||\mathcal{F}|/\delta)/\epsilon^2$.

□

H. Relaxation of Assumption 2.1

Our theoretical results presented so far rely on the realizable assumption (Assumption 2.1). While equipped with recent advances in powerful non-linear function approximators (e.g., deep neural networks), readability can be ensured, in this section, we relax the realizable assumption and show that our algorithms' performance only degenerates mildly. We relax Assumption 2.1 as follows:

Assumption H.1 (Approximate Realizability)**.** *We assume Π and $\mathcal{F}$ is approximate realizable in a sense that for any $h \in [H]$, we have $\min_{\pi\in\Pi_h} \max_{x,a} \|\pi(a|x) - \pi^\star_h(a|x)\| \leq \epsilon_\Pi$ and $\min_{f\in\mathcal{F}_h} \|f - V_h^\star\|_\infty \leq \epsilon_{\mathcal{F}}$.*

The above assumption does not require $\mathcal{F}$ and Π to contain the exact $V_h^\star$ and $\pi_h^\star$, but assumes $\mathcal{F}$ and Π are rich enough to contain functions that can approximate $V_h^\star$ and $\pi_h^\star$ uniformly well (i.e., $\epsilon_{\mathcal{F}}$ and ϵ_Π are small). Without any further modification of Algorithm 2 and Algorithm 5 for non-interactive and interactive setting, we have the following corollary.

Corollary H.2. *Under Assumption H.1, for $\epsilon \in (0,1)$ and $\delta \in (0,1)$, with $T = \Theta(K/\epsilon^2)$, $n = \Theta(K\log(|\Pi||\mathcal{F}|H/\delta)/\epsilon^2)$, with probability at least $1-\delta$, (1) for non-interactive setting,* FAIL *(Algorithm 2) outputs a policy $\boldsymbol{\pi}$ with $J(\boldsymbol{\pi}) - J(\boldsymbol{\pi}^\star) \leq O\left(H^2(\epsilon_{\text{be}} + \epsilon) + H(\epsilon_{\mathcal{F}} + \epsilon_\Pi)\right)$, and (2) for interactive setting,* IFAIL *(Algorithm 5) outputs a policy $\boldsymbol{\pi}$ with $J(\boldsymbol{\pi}) - J(\boldsymbol{\pi}^\star) \leq O\left(H\epsilon + H\epsilon_{\mathcal{F}} + H\epsilon_\Pi\right)$, by using at most $\tilde{O}((HK/\epsilon^2)\log(|\Pi||\mathcal{F}|/\delta))$ many trajectories under both settings.*

The proof is deferred to Appendix H.1.

H.1. Proof of Corollary H.2

Proof of Corollary H.2. For any h, denote g_h as

$$g_h = \arg\min_{g\in\mathcal{F}} \|g - V_h^\star\|_\infty$$

Below we prove the first bullet in Corollary H.2, i.e., the results for non-interactive setting.

Non-Interactive Setting Using PDL (Lemma C.1), we have

$$\begin{aligned}
& J(\boldsymbol{\pi}) - J(\boldsymbol{\pi}^\star) \\
& \leq \sum_{h=1}^{H} \left| \mathbb{E}_{x\sim\mu_h^{\boldsymbol{\pi}}}\left[\mathbb{E}_{a\sim\pi_h, x'\sim P_{x,a}}\left[V_{h+1}^\star(x')\right] - \mathbb{E}_{a\sim\pi_h^\star, x'\sim P_{x,a}}\left[V_{h+1}^\star(x')\right]\right]\right| \\
& \leq \sum_{h=1}^{H} \left| \mathbb{E}_{x\sim\mu_h^{\boldsymbol{\pi}}}\left[\mathbb{E}_{a\sim\pi_h, x'\sim P_{x,a}}\left[V_{h+1}^\star(x')\right]\right] - \mathbb{E}_{x\sim\mu_h^\star}\mathbb{E}_{a\sim\pi_h^\star, x'\sim P_{x,a}}\left[V_{h+1}^\star(x')\right]\right| \\
& \quad + \left| \mathbb{E}_{x\sim\mu_h^\star}\mathbb{E}_{a\sim\pi_h^\star, x'\sim P_{x,a}}\left[V_{h+1}^\star(x')\right] - \mathbb{E}_{x\sim\mu_h^{\boldsymbol{\pi}}}\mathbb{E}_{a\sim\pi_h^\star, x'\sim P_{x,a}}\left[V_{h+1}^\star(x')\right]\right| \\
& \leq \sum_{h=1}^{H} \left| \mathbb{E}_{x\sim\mu_h^{\boldsymbol{\pi}}}\left[\mathbb{E}_{a\sim\pi_h, x'\sim P_{x,a}}\left[g_{h+1}(x')\right]\right] - \mathbb{E}_{x\sim\mu_h^\star}\mathbb{E}_{a\sim\pi_h^\star, x'\sim P_{x,a}}\left[g_{h+1}(x')\right]\right| \\
& \quad + \left| \mathbb{E}_{x\sim\mu_h^\star}\mathbb{E}_{a\sim\pi_h^\star, x'\sim P_{x,a}}\left[g_{h+1}(x')\right] - \mathbb{E}_{x\sim\mu_h^{\boldsymbol{\pi}}}\mathbb{E}_{a\sim\pi_h^\star, x'\sim P_{x,a}}\left[g_{h+1}(x')\right]\right| + 4\epsilon_{\mathcal{F}} \\
& \leq \sum_{h=1}^{H} \left(d_{\mathcal{F}_{h+1}}(\pi_h | \mu_h^{\boldsymbol{\pi}}, \mu_h^\star) + \Delta_h + 4\epsilon_{\mathcal{F}} \right).
\end{aligned}$$

Now repeat the same recursive analysis for $d_{\mathcal{F}_{h+1}}(\pi_h|\mu_h^{\boldsymbol{\pi}}, \mu_h^\star)$ as we did in proof of Theorem 3.3 in Appendix C, we can prove the first bullet in the corollary.

Now we prove the second bullet in Corollary H.2, i.e., the results for interactive setting.

Interactive Setting Again, using Performance Difference Lemma (Lemma C.1), we have

$$\begin{aligned}
J(\boldsymbol{\pi}) - J(\boldsymbol{\pi}^\star) &= \sum_{h=1}^{H} \mathbb{E}_{x\sim\mu_h^{\boldsymbol{\pi}}}\left[\mathbb{E}_{a\sim\pi_h, x'\sim P_{x,a}}\left[V_{h+1}^\star(x')\right] - \mathbb{E}_{a\sim\pi_h^\star, x'\sim P_{x,a}}\left[V_{h+1}^\star(x')\right]\right] \\
&\leq \sum_{h=1}^{H} \left|\mathbb{E}_{x\sim\mu_h^{\boldsymbol{\pi}}}\left[\mathbb{E}_{a\sim\pi_h, x'\sim P_{x,a}}\left[V_{h+1}^\star(x')\right] - \mathbb{E}_{a\sim\pi_h^\star, x'\sim P_{x,a}}\left[V_{h+1}^\star(x')\right]\right]\right| \\
&\leq \sum_{h=1}^{H} \left|\mathbb{E}_{x\sim\mu_h^{\boldsymbol{\pi}}}\left[\mathbb{E}_{a\sim\pi_h, x'\sim P_{x,a}}\left[g_{h+1}(x')\right] - \mathbb{E}_{a\sim\pi_h^\star, x'\sim P_{x,a}}\left[g_{h+1}(x')\right]\right]\right| \\
&+ \left|\mathbb{E}_{x\sim\mu_h^{\boldsymbol{\pi}}}\mathbb{E}_{a\sim\pi_h, x'\sim P_{x,a}}[V_{h+1}^\star(x') - g_{h+1}(x')]\right| + \left|\mathbb{E}_{x\sim\mu_h^{\boldsymbol{\pi}}}\mathbb{E}_{a\sim\pi_h^\star, x'\sim P_{x,a}}[V_{h+1}^\star(x') - g_{h+1}(x')]\right| \\
&\leq \sum_{h=1}^{H} \max_{f\in\mathcal{F}_{h+1}} \left|\mathbb{E}_{x\sim\mu_h^{\boldsymbol{\pi}}}\left[\mathbb{E}_{a\sim\pi_h, x'\sim P_{x,a}}\left[f(x')\right] - \mathbb{E}_{a\sim\pi_h^\star, x'\sim P_{x,a}}\left[f(x')\right]\right]\right| + 2\epsilon_{\mathcal{F}} \\
&= \sum_{h=1}^{H} \left(d_{\mathcal{F}_{h+1}}(\pi_h|\mu_h^{\boldsymbol{\pi}}, \pi_h^\star) + 2\epsilon_{\mathcal{F}}\right)
\end{aligned}$$

Now repeat the same steps from the proof of Theorem 5.2 after (18) in proof of Theorem 5.2, we can prove the second bullet in the corollary.

□

I. Missing Details on ILFO with State Abstraction

We consider the bisimulation model from (6). The following proposition summarizes the conclusion in this section.

Proposition I.1. *Assume Bisimulation holds (Eq. 6) and set $\mathcal{F}_h = \{f : \|f\|_\infty \leq 1, f(x) = f(x'), \forall x, x'$ s.t. $\phi(x) = \phi(x')\}, \forall h \in [H]$ to be piece-wise constant functions over the partitions induced from ϕ. We have:*

1. *$V_h^\star$ is a piece-wise constant function for all $h \in [H]$,*

2. *$\epsilon_{\mathrm{be}} = 0$,*

3. *$\sup_{f\in\mathcal{F}_h}(\sum_{i=1}^N f(x_i)/N - \sum_{i=1}^N f(x_i')/N)$ can be solved by LP, for all $h \in [H]$,*

4. *given any $\{x_i\}_{i=1}^N$, the Rademacher complexity of $\mathcal{F}_h$ is in the order of $O(\sqrt{|\mathcal{S}|/N})$, i.e., $(1/N)\mathbb{E}_\sigma[\sup_{f\in\mathcal{F}_h}\sum_{i=1}^N \sigma_i f(x_i)] = O(\sqrt{|\mathcal{S}|}/N)$, with σ_i being a Rademacher number.*

The above proposition states that by leveraging the abstraction, we can design discriminators to be piece-wise constant functions over the partitions induced by ϕ, such that inherent Bellman error is zero, and the discriminator class has bounded statistical complexity. Below we prove the above proposition. The first two points in the above proposition were studied in (Chen & Jiang, 2019). For completeness, we simply prove all four points below.

Piece-wise constant $V^\star$ First, we show that $V_h^\star(x)$ is piece-wise constant over the partitions induced from ϕ. Starting from H, via (6), we know that $V_H^\star(x) = c(x)$, which is piece-wise constant over the partitions induced from ϕ. Then let us assume that for $h+1$, we have $V_{h+1}^\star(x) = V_{h+1}^\star(x')$ for any x, x' s.t. $\phi(x) = \phi(x')$. At time step h, via Bellman equation, we know:

$$V_h^\star(x) = \mathbb{E}_{a\sim\pi^\star(\cdot|x)}\mathbb{E}_{x'\sim P_{x,a}} V_{h+1}^\star(x').$$

Hence for any two x, x' with $\phi(x) = \phi(x')$, we have:

$$\begin{aligned} V_h^\star(x) - V_h^\star(x') &= \mathbb{E}_{a\sim\pi^\star(\cdot|x), x''\sim P_{x,a}} V_{h+1}^\star(x'') - \mathbb{E}_{a\sim\pi^\star(\cdot|x'), x''\sim P_{x',a}} V_{h+1}^\star(x'') \\ &= \sum_a \pi^\star(a|x)\left(\sum_{s\in\mathcal{S}}\sum_{x''\in\phi^{-1}(s)}\left(P(x''|x,a) - P(x''|x',a)\right)V_{h+1}^\star(x'')\right) \\ &= \sum_a \pi^\star(a|x)\left(\sum_{s\in\mathcal{S}} V_{h+1}^\star(s)\sum_{x''\in\phi^{-1}(s)}\left(P(x''|x,a) - P(x''|x',a)\right)\right) = 0, \end{aligned}$$

where the second and the last equality use (6). In the third equality above, we abuse the notation $V_{h+1}^\star(s)$ for $s\in\mathcal{S}$ to denote the value of $V_{h+1}^\star(x)$ for any x such that $\phi(x) = s$.

Inherent Bellman Error With $\mathcal{F}_h = \{f : \|f\|_\infty \leq 1, f(x) = f(x'), \forall x, x' \text{ s.t. } \phi(x) = \phi(x')\}, \forall h\in[H]$, we can show $\epsilon_{\text{be}} = 0$ as follows. For any $x, x' \in \mathcal{X}$ with $\phi(x) = \phi(x'), f\in\mathcal{F}_{h+1}$, we have:

$$\begin{aligned} &\mathbb{E}_{a\sim\pi^\star(\cdot|x)}\mathbb{E}_{x''\sim P_{x,a}} f(x'') - \mathbb{E}_{a\sim\pi^\star(\cdot|x')}\mathbb{E}_{x''\sim P_{x',a}} f(x'') \\ &= \sum_a \pi^\star(a|x)\left(\sum_{s\in\mathcal{S}} f(s)\sum_{x''\in\phi^{-1}(s)}\left(P(x''|x,a) - P(x''|x',a)\right)\right) = 0, \end{aligned}$$

where again we abuse the notation $f(s)$ to denote that value $f(x)$ for any x such that $\phi(x) = s$. Namely, $\Gamma_h f$ is also a piece-wise constant over the partitions induced from ϕ. Since $\|f\|_\infty \leq 1$, it is also easy to see that $\|\Gamma_h f\|_\infty \leq 1$. Hence we have $\Gamma_h f \in \mathcal{F}_h$.

Reduction to LP Regarding evaluating $\sup_{f\in\mathcal{F}_h}\left(\sum_{i=1}^N f(x_i)/N - \sum_{i=1}^N f(x_i')/N\right)$, we can again reduce it an LP. Denote $\alpha\in[-1,1]^{|\mathcal{S}|}$, where the i-th entry in α corresponds to the i-th element in $\mathcal{S}$. We denote α_s as the entry in α that corresponds to the state s in $\mathcal{S}$. Take $\{x_i\}_{i=1}^N$, and compute $c_s = \sum_{i=1}^N \mathbf{1}(\phi(x_i) = s)$ for every $s\in\mathcal{S}$ (i.e., c_s is the number of points mapped to s). Take $\{x_i'\}_{i=1}^N$ and compute $c_s' = \sum_{i=1}^N \mathbf{1}(\phi(x_i') = s)$. We solve the following LP:

$$\begin{aligned} &\max_{\alpha\in\mathbb{R}^{|\mathcal{S}|}} \sum_{s\in\mathcal{S}}\left(c_s\alpha_s/N - c_s'\alpha_s/N\right), \\ &s.t., \alpha_s\in[-1,1], \forall s\in\mathcal{S}. \end{aligned}$$

Denote the solution of the above LP as $\alpha^\star$. Then $f^\star(x) = \alpha^\star_{\phi(x)}$.

Complexity of Discriminators $\mathcal{F}_h$ Regarding the complexity of $\mathcal{F}_h$, note that $\mathcal{F}_h$ essentially corresponds to a $|\mathcal{S}|$-dim box: $[-1,1]^{|\mathcal{S}|}$. Again, consider a dataset $\{x_i\}_{i=1}^N$ and the counts $\{c_s\}_{s\in\mathcal{S}}$. For any f, and Rademacher numbers $\sigma\in\{-1,1\}^N$, we have

$$\sum_{i=1}^N \sigma_i f(x_i) = \sum_{s\in\mathcal{S}} f_s \sum_{i\in\phi^{-1}(s)}\sigma_i \leq \sum_{s\in\mathcal{S}}\left|\sum_{i\in\phi^{-1}(s)}\sigma_i\right|.$$

Note that $(\mathbb{E}_\sigma\left|\sum_{i=1}^N\sigma_i\right|)^2 \leq \mathbb{E}_\sigma(\sum_{i=1}^N\sigma_i)^2 = N$, which implies that $\mathbb{E}_\sigma|\sum_{i=1}^N\sigma_i| \leq \sqrt{N}$. Hence,

$$\mathbb{E}_\sigma\sum_{i=1}^N\sigma_i f(x_i) \leq \sum_{s\in\mathcal{S}}\mathbb{E}_\sigma|\sum_{i\in\phi^{-1}(s)}\sigma_i| \leq \sum_{s\in\mathcal{S}}\sqrt{c_s} \leq \sum_{s\in\mathcal{S}}\sqrt{N/|\mathcal{S}|} = \sqrt{N|\mathcal{S}|}.$$

Now, we can show that the Rademacher complexity of $\mathcal{F}_h$ is bounded as follows:

$$\frac{1}{N}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}_h}\sum_{i=1}^N\sigma_i f(x_i)\right] \leq \sqrt{N|\mathcal{S}|}/N = \sqrt{\frac{|\mathcal{S}|}{N}}.$$

Algorithm 4 Min-Max Game ($\mathcal{D}^\star, \mathcal{D}, \Pi, \mathcal{F}, T, \theta_0$)

```
for $n = 0$ to $T$ do
    $f^n = \arg\max_{f\in\mathcal{F}} u(\pi_{\theta^n}, f)$ (LP Oracle)
    $u^n = u(\pi_{\theta^n}, f^n)$
    $\theta^{n+1} = \theta^n - \nabla_\theta u(\pi_{\theta^n}, f^n)$ (Policy Gradient)
end for
Output: $\pi^{n^\star}$ with $n^\star = \arg\min_{n\in[T]} u^n$
```

J. Additional Experiments

When we design the utility in (3), we sample actions from $U(\mathcal{A})$ and then perform importance weighting. This ensures that in analysis the variance will be bounded by K. In practice, we can use any reference policy to generate actions, and then perform importance weighting accordingly. Assume that we have a dataset $\mathcal{D} = \{x_h^i, a_h^i, p_h^i, x_{h+1}^i\}_{i=1}^N$ and the expert's dataset $\mathcal{D}^\star = \{\tilde{x}_{h+1}^i\}_{i=1}^{N'}$, where p_h^i is the probability of action a_h^i being chosen at x_h^i. We can form the utility as follows:

$$u(\pi, f) \triangleq \sum_{i=1}^{N} (\pi(a_h^i|x_h^i)/p_h^i) f(x_{h+1}^i)/N - \sum_{i=1}^{N'} f(\tilde{x}_{h+1}^i)/N'. \tag{19}$$

As long as the probability of choosing any action at any state is lower bounded, then the variance of the above estimator is upper bounded. This formulation also immediately extends FAIL to continuous action space setting. For a parameterized policy π_θ, given any f, we can compute $\nabla_\theta u(\pi_\theta, f)$ easily. If $a_h \sim \pi_\theta(\cdot|x)$ (i.e., on-policy samples), then for any fixed f, the policy gradient $\nabla_\theta u(\pi_\theta, f)$ can be estimated using the REINFORCE trick:

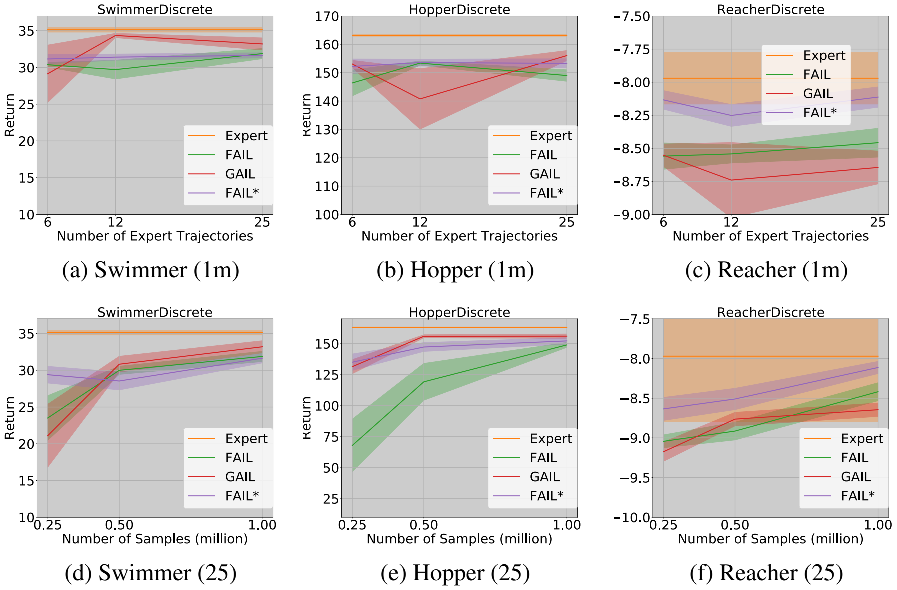
$$\nabla_\theta u(\pi_\theta, f)|_{\theta=\theta_0} = (1/N) \sum_{i=1}^{N} \nabla_\theta (\ln \pi_\theta(a_h^i|x_h^i)|_{\theta=\theta_0}) f(x_{h+1}^i). \tag{20}$$

With the form of $\nabla_\theta u(\pi_\theta, f)$, we can perform the min-max optimization in Alg. 1 by iteratively finding the maximizer $f^n = \arg\max_f u(\pi_{\theta_n}, f)$ using LP oracle, and then perform gradient descent update $\theta^{n+1} = \theta^n - \eta^n \nabla_\theta u(\pi_{\theta^n}, f^n)$. See Algorithm 4 below. Note that in Algorithm 4 the dataset $\mathcal{D} = \{x_h^i, a_h^i, p_h^i, x_{h+1}^i\}$ contains p_h^i which is the probability of a_h^i being chosen at x_h^i. We can integrate Algorithm 4 into the forward training framework.

Algorithm 5 FAIL* ($\{\Pi_h\}_h, \{\mathcal{F}_h\}_h, \epsilon, n, n', T$)

```
1:  Set $\boldsymbol{\pi} = \emptyset$
2:  for $h = 1$ to $H - 1$ do
3:      Initialize $\pi_h$
4:      Extract expert's data at $h + 1$: $\tilde{\mathcal{D}}_{h+1} = \{\tilde{x}_{h+1}^i\}_{i=1}^{n'}$
5:      $\mathcal{D}_1 = \emptyset, \dots \mathcal{D}_h = \emptyset$
6:      for $i = 1$ to $n$ do
7:          Execute $\{\pi_1, \dots, \pi_{h-1}\}$ to generate $\tau^i = \{x_1^i, a_1^i, p_1^i, x_2^i, \dots, x_{h-1}^i, a_{h-1}^i, p_{h-1}^i, x_h^i\}$ with $p_t^i = \pi_t(a_t^i|x_t^i)$
8:          For any $t \in [h-1]$, add $(x_t^i, a_t^i, p_t^i, x_{t+1}^i)$ to $\mathcal{D}_t$
9:          Execute $a_h^i \sim U(\mathcal{A})$ to generate $x_{h+1}^i$ and add $(x_h^i, a_h^i, p_h^i, x_{h+1}^i)$ to $\mathcal{D}_h$ with $p_h^i$ being the probability corresponding to the uniform distribution over $\mathcal{A}$
10:     end for
11:     For all $t \in [h]$, update $\pi_t$ to be the return of Algorithm 4 with inputs $\left(\tilde{D}_{t+1}, \mathcal{D}_t, \Pi_h, \mathcal{F}_{h+1}, T, \pi_t\right)$
12: end for
```

In Algorithm 2, at every time step h, we execute the current sequence of policies $\boldsymbol{\pi} = \{\pi_1, \dots, \pi_{h-1}\}$ to collect samples at time step h, i.e., x_h. We then throw away all generated samples $\{x_1, \dots, x_{h-1}\}$ except x_h. While this simplifies the analysis, in practice, we could leverage these samples $\{x_1, \dots, x_{h-1}\}$ as well, especially now we can form the utility with on-policy samples and compute the corresponding policy gradient ((20)). This leads us to Alg. 5. Namely, in Algorithm 5, when training π_h, we also incrementally update $\pi_1, \dots, \pi_{h-1}$ using their on-policy samples (Line 11 Algorithm 5).



(a) Swimmer (1m) (b) Hopper (1m) (c) Reacher (1m)

(d) Swimmer (25) (e) Hopper (25) (f) Reacher (25)

Figure 3. Performance of expert, FAIL* (Algorithm 5), FAIL(Algorithm 2), and GAIL (without actions) on three control tasks. For the top line, we fix the number of training samples while varying the number of expert demonstrations (6, 12, 25). For the bottom line, we fix the number of expert demonstrations, while varying the number of training samples. All results are averaged over 10 random seeds.

We test Algorithm 5 on the same set of environments (Figure 3) under 10 random rand seeds, with all default parameters. We observe that FAIL* can be more sample efficient especially in small data setting (e.g., 0.25 million training samples). Implementation of Algorithm 5 and scripts for reproducing results can be found in `https://github.com/wensun/Imitation-Learning-from-Observation/tree/fail_star`.