Provably Efficient Adversarial Imitation Learning with Unknown Transitions (Supplementary Material)

1 NOTATION

Table 1: Notations

Symbol	Meaning
π^{E}	the expert policy
$V^{\pi,P,r}$	policy value under the transition model P and reward r
ε	the imitation gap
δ	failure probability
$d_h^{\pi}(s)$	state distribution
$d_h^{\pi}(s,a)$	state-action distribution
$\mathbf{tr} = (s_1, a_1, \cdots, s_H, a_H)$	the trajectory
$\mathbf{tr}_h = (s_1, a_1, \cdots, s_h, a_h)$	the truncated trajectory
$\mathbf{tr}_h(\cdot)$	the state at time step h in $\mathbf{tr}$
$\mathbf{tr}_h(\cdot,\cdot)$	the state-action pair at time step h in $\mathbf{tr}$
$\mathbf{tr}(a_h)$	the action at time step h in $\mathbf{tr}$
$\mathcal{D}$	expert dataset
m	number of expert trajectories
$\widehat{d}_h^{\pi^{\mathrm{E}}}(s,a)$	maximum likelihood estimator of $d_h^{\pi^{\mathrm{E}}}$ in Equation (4)
$\widetilde{d}_h^{\pi^{\mathrm{E}}}(s,a)$	transition-aware estimator in Equation (7)
$\mathbb{P}^{\pi^{\mathrm{E}}}(\mathbf{tr})$	probability of the trajectory $\mathbf{tr}$ under the expert policy π^{E}
$\mathbb{P}^{\pi^{\mathrm{E}}}(\mathbf{tr}_h)$	probability of the truncated trajectory $\mathbf{tr}_h$ under the expert policy π^{E}
$\mathcal{S}_h(\mathcal{D})$	the set of states visited in time step h in dataset $\mathcal{D}$
$\mathrm{Tr}_h^{\mathcal{D}}$	the trajectories along which each state has been visited in $\mathcal{D}$ up to time step h
$\pi^{(t)}$	the policy obtained in the iteration t
$w^{(t)}$	the reward function learned in the iteration t
$\eta^{(t)}$	the step size in the iteration t
$f^{(t)}(w)$	the objective function in the iteration t in Equation (9)
$\bar{d}_h(s,a)$	the averaged state-action distribution in Algorithm 2
$\overline{\pi}$	the policy derived by the averaged state-action distribution in Algorithm 2
$\Pi_{\mathrm{BC}}(\mathcal{D}_1)$	the set of policies which take the expert action on states covered in $\mathcal{D}_1$
$\widehat{P}$	the empirical transition function
$d_h^{\pi,\widehat{P}}(s,a)$	the state-action distribution of π under the empirical transition function $\widehat{P}$

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

2 FROM REGRET GUARANTEE TO PAC GUARANTEE

Shani et al. [2022] proved a regret guarantee for their OAL algorithm. In particular, Shani et al. [2022] showed that with probability at least $1-\delta'$, we have

$$\sum_{k=1}^{K} V^{\pi^{\mathrm{E}}} - V^{\pi_k} \leq \widetilde{\mathcal{O}}\left(\sqrt{H^4|\mathcal{S}|^2|\mathcal{A}|K} + \sqrt{H^3|\mathcal{S}||\mathcal{A}|K^2/m}\right), \tag{1}$$

where π^k is the policy obtained at episode k, K is the number of interaction episodes, and m is the number of expert trajectories. We would like to comment that the second term in (1) involves the statistical estimation error about the expert policy. Furthermore, this term reduces to $\widetilde{\mathcal{O}}(\sqrt{H^2|\mathcal{S}|K^2/m})$ under the assumption that the expert policy is deterministic.

To further convert this regret guarantee to the PAC guarantee considered in this paper, we can apply Markov's inequality as suggested by [Jin et al., 2018]. Concretely, let $\overline{\pi}$ be the policy that randomly chosen from $\{\pi^1, \pi^2, \cdots, \pi^K\}$ with equal probability, then we have

$$\mathbb{P}\left(V^{\pi^{\mathrm{E}}} - V^{\overline{\pi}} \geq \varepsilon\right) \leq \frac{1}{\varepsilon}\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K} V^{\pi^{\mathrm{E}}} - V^{\pi_k}\right] \leq \frac{1}{\varepsilon}\left(\widetilde{\mathcal{O}}\left(\sqrt{\frac{H^4|\mathcal{S}|^2|\mathcal{A}|}{K}} + \sqrt{H^2|\mathcal{S}|/m}\right) + \delta' H\right),$$

Therefore, if we set $\delta' = \varepsilon\delta/(3H)$, and

$$K = \widetilde{\mathcal{O}}\left(\frac{H^4|\mathcal{S}|^2|\mathcal{A}|}{\varepsilon^2\delta^2}\right), \quad m = \widetilde{\mathcal{O}}\left(\frac{H^2|\mathcal{S}|}{\varepsilon^2}\right),$$

we obtain that $\mathbb{P}(V^{\pi^{\mathrm{E}}} - V^{\overline{\pi}} \geq \varepsilon) \leq \delta$.

3 PROOF OF RESULTS IN SECTION 4

3.1 PROOF OF LEMMA 1

Proof. The proof starts with the dual representation of policy value (see Equation (1)).

$$\begin{aligned}
V^{\pi^{\mathrm{E}}} - V^{\overline{\pi}} &= \sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left(d_h^{\pi^{\mathrm{E}}}(s,a) - d_h^{\overline{\pi}}(s,a)\right) r_h(s,a) \\
&\overset{(a)}{\leq} \sum_{h=1}^{H}\left\|d_h^{\pi^{\mathrm{E}}} - d_h^{\overline{\pi}}\right\|_1 \\
&\leq \sum_{h=1}^{H}\left\|d_h^{\pi^{\mathrm{E}}} - \widetilde{d}_h^{\pi^{\mathrm{E}}}\right\|_1 + \sum_{h=1}^{H}\left\|d_h^{\overline{\pi}} - \widetilde{d}_h^{\pi^{\mathrm{E}}}\right\|_1,
\end{aligned}$$

where inequality (a) is based on the assumption that $r_h(s,a) \in [0,1]$. For the two terms in RHS, according to Definition 1 and Definition 2, we have

$$\sum_{h=1}^{H}\left\|d_h^{\pi^{\mathrm{E}}} - \widetilde{d}_h^{\pi^{\mathrm{E}}}\right\|_1 \leq \varepsilon_{\mathrm{EST}}, \ \sum_{h=1}^{H}\left\|d_h^{\overline{\pi}} - \widetilde{d}_h^{\pi^{\mathrm{E}}}\right\|_1 \leq \min_{\pi\in\Pi}\sum_{h=1}^{H}\left\|d_h^{\pi} - \widetilde{d}_h^{\pi^{\mathrm{E}}}\right\|_1 + \varepsilon_{\mathrm{OPT}}.$$

With the above two inequalities, we further obtain

$$\begin{aligned}
V^{\pi^{\mathrm{E}}} - V^{\overline{\pi}} &\leq \varepsilon_{\mathrm{EST}} + \min_{\pi\in\Pi}\sum_{h=1}^{H}\left\|d_h^{\pi} - \widetilde{d}_h^{\pi^{\mathrm{E}}}\right\|_1 + \varepsilon_{\mathrm{OPT}} \\
&\overset{(a)}{\leq} \varepsilon_{\mathrm{EST}} + \sum_{h=1}^{H}\left\|d_h^{\pi^{\mathrm{E}}} - \widetilde{d}_h^{\pi^{\mathrm{E}}}\right\|_1 + \varepsilon_{\mathrm{OPT}} \\
&\leq 2\varepsilon_{\mathrm{EST}} + \varepsilon_{\mathrm{OPT}}.
\end{aligned}$$

Inequality (a) holds since $\pi^{\mathrm{E}} \in \Pi$. We complete the proof.

□

4 PROOF OF RESULTS IN SECTION 5

4.1 PROOF OF PROPOSITION 1

Proof. Let $\widetilde{d}_h^{\pi^{\mathrm{E}}}(s,a)$ be an expert state-action distribution estimator and $\widehat{P}$ be a transition model learned by a reward-free method. Notice that reward-free exploration methods also enable uniform policy evaluation with respect to *any* reward function; see Definition 4. That is, with probability at least $1-\delta_{\mathrm{RFE}}$, for any reward function r and policy π, we have $|V^{\pi,P,r} - V^{\pi,\widehat{P},r}| \leq \varepsilon_{\mathrm{RFE}}$. Then we define the following two events.

$$E_{\mathrm{EST}} := \left\{ \sum_{h=1}^{H} \left\| \widetilde{d}_h^{\pi^{\mathrm{E}}} - d_h^{\pi^{\mathrm{E}}} \right\|_1 \leq \varepsilon_{\mathrm{EST}} \right\},$$
$$E_{\mathrm{RFE}} := \left\{ \forall r = (r_1, \ldots, r_H), \forall \pi \in \Pi : \left| V^{\pi,P,r} - V^{\pi,\widehat{P},r} \right| \leq \varepsilon_{\mathrm{RFE}} \right\}.$$

According to assumption (a) and (b), we have that $\mathbb{P}(E_{\mathrm{EST}}) \geq 1-\delta_{\mathrm{EST}}$ and $\mathbb{P}(E_{\mathrm{RFE}}) \geq 1-\delta_{\mathrm{RFE}}$. Applying union bound yields

$$\mathbb{P}(E_{\mathrm{EST}} \cap E_{\mathrm{RFE}}) \geq 1 - \delta_{\mathrm{EST}} - \delta_{\mathrm{RFE}}.$$

The following analysis is established on the event $E_{\mathrm{EST}} \cap E_{\mathrm{RFE}}$. Let $\overline{\pi}$ be the output of Algorithm 1.

$$\left| V^{\pi^{\mathrm{E}},P} - V^{\overline{\pi},P} \right| \leq \left| V^{\pi^{\mathrm{E}},P} - V^{\overline{\pi},\widehat{P}} \right| + \left| V^{\overline{\pi},\widehat{P}} - V^{\overline{\pi},P} \right| \leq \left| V^{\pi^{\mathrm{E}},P} - V^{\overline{\pi},\widehat{P}} \right| + \varepsilon_{\mathrm{RFE}}.$$

The last inequality follows the event E_{RFE}. Then we consider the error $|V^{\pi^{\mathrm{E}},P} - V^{\overline{\pi},\widehat{P}}|$. From the dual form of the policy value in Equation (1), we have that

$$\left| V^{\pi^{\mathrm{E}},P} - V^{\overline{\pi},\widehat{P}} \right| = \left| \sum_{h=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left(d_h^{\pi^{\mathrm{E}},P}(s,a) - d_h^{\overline{\pi},\widehat{P}}(s,a) \right) r_h(s,a) \right| \leq \sum_{h=1}^{H} \left\| d_h^{\pi^{\mathrm{E}},P} - d_h^{\overline{\pi},\widehat{P}} \right\|_1,$$

where $d_h^{\overline{\pi},\widehat{P}}(s,a)$ is the state-action distribution of the policy $\overline{\pi}$ under the transition model $\widehat{P}$. Then we get that

$$\begin{aligned}
\sum_{h=1}^{H} \left\| d_h^{\pi^{\mathrm{E}},P} - d_h^{\overline{\pi},\widehat{P}} \right\|_1 &\leq \sum_{h=1}^{H} \left\| d_h^{\pi^{\mathrm{E}},P} - \widetilde{d}_h^{\pi^{\mathrm{E}}} \right\|_1 + \sum_{h=1}^{H} \left\| \widetilde{d}_h^{\pi^{\mathrm{E}}} - d_h^{\overline{\pi},\widehat{P}} \right\|_1 \\
&\leq \varepsilon_{\mathrm{EST}} + \sum_{h=1}^{H} \left\| \widetilde{d}_h^{\pi^{\mathrm{E}}} - d_h^{\overline{\pi},\widehat{P}} \right\|_1.
\end{aligned}$$

The last inequality follows the event E_{EST}. Combining the above three inequalities yields

$$\left| V^{\pi^{\mathrm{E}},P} - V^{\overline{\pi},P} \right| \leq \sum_{h=1}^{H} \left\| \widetilde{d}_h^{\pi^{\mathrm{E}}} - d_h^{\overline{\pi},\widehat{P}} \right\|_1 + \varepsilon_{\mathrm{EST}} + \varepsilon_{\mathrm{RFE}}.$$

According to assumption (c), with the estimator $\widetilde{d}_h^{\pi^{\mathrm{E}}}(s,a)$ and transition model $\widehat{P}$, algorithm C solves the optimization problem in Equation (3) up to an error $\varepsilon_{\mathrm{OPT}}$ and $\overline{\pi}$ is the output of the algorithm C. Formally,

$$\sum_{h=1}^{H} \left\| \widetilde{d}_h^{\pi^{\mathrm{E}}} - d_h^{\overline{\pi},\widehat{P}} \right\|_1 \leq \min_{\pi\in\Pi} \sum_{h=1}^{H} \left\| \widetilde{d}_h^{\pi^{\mathrm{E}}} - d_h^{\pi,\widehat{P}} \right\|_1 + \varepsilon_{\mathrm{OPT}}.$$

Then we get that

$$\begin{aligned}
\left| V^{\pi^{\mathrm{E}},P} - V^{\overline{\pi},P} \right| &\leq \sum_{h=1}^{H} \left\| \widetilde{d}_h^{\pi^{\mathrm{E}}} - d_h^{\overline{\pi},\widehat{P}} \right\|_1 + \varepsilon_{\mathrm{EST}} + \varepsilon_{\mathrm{RFE}} \\
&\leq \min_{\pi\in\Pi} \sum_{h=1}^{H} \left\| \widetilde{d}_h^{\pi^{\mathrm{E}}} - d_h^{\pi,\widehat{P}} \right\|_1 + \varepsilon_{\mathrm{OPT}} + \varepsilon_{\mathrm{EST}} + \varepsilon_{\mathrm{RFE}}
\end{aligned}$$

$$
\begin{aligned}
&\overset{(a)}{\leq} \sum_{h=1}^{H} \left\| \widetilde{d}_h^{\pi^{\mathrm{E}}} - d_h^{\pi^{\mathrm{E}},\widehat{P}} \right\|_1 + \varepsilon_{\mathrm{OPT}} + \varepsilon_{\mathrm{EST}} + \varepsilon_{\mathrm{RFE}} \\
&\leq \sum_{h=1}^{H} \left\| \widetilde{d}_h^{\pi^{\mathrm{E}}} - d_h^{\pi^{\mathrm{E}},P} \right\|_1 + \sum_{h=1}^{H} \left\| d_h^{\pi^{\mathrm{E}},P} - d_h^{\pi^{\mathrm{E}},\widehat{P}} \right\|_1 + \varepsilon_{\mathrm{OPT}} + \varepsilon_{\mathrm{EST}} + \varepsilon_{\mathrm{RFE}} \\
&\overset{(b)}{\leq} \sum_{h=1}^{H} \left\| d_h^{\pi^{\mathrm{E}},P} - d_h^{\pi^{\mathrm{E}},\widehat{P}} \right\|_1 + \varepsilon_{\mathrm{OPT}} + 2\varepsilon_{\mathrm{EST}} + \varepsilon_{\mathrm{RFE}},
\end{aligned}
$$

where inequality (a) holds since $\pi^{\mathrm{E}} \in \Pi$ and inequality (b) follows the event E_{EST}. With the dual representation of ℓ_1-norm, we have that

$$
\begin{aligned}
\sum_{h=1}^{H} \left\| d_h^{\pi^{\mathrm{E}},P} - d_h^{\pi^{\mathrm{E}},\widehat{P}} \right\|_1 &= \max_{w \in \mathcal{W}} \sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} w_h(s,a) \left(d_h^{\pi^{\mathrm{E}},P}(s,a) - d_h^{\pi^{\mathrm{E}},\widehat{P}}(s,a) \right) \\
&= \max_{w \in \mathcal{W}} \sum_{h=1}^{H} V^{\pi^{\mathrm{E}},P,w} - V^{\pi^{\mathrm{E}},\widehat{P},w} \leq \varepsilon_{\mathrm{RFE}},
\end{aligned}
$$

where $\mathcal{W} = \{w = (w_1, \ldots, w_H) : w_h \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}, \|w_h\|_\infty \leq 1\}$, $V^{\pi^{\mathrm{E}},\widehat{P},w}$ is the value of policy π^{E} with the transition model $\widehat{P}$ and reward function w. The last inequality follows the event E_{RFE}. Then we prove that

$$
\left| V^{\pi^{\mathrm{E}},P} - V^{\overline{\pi},P} \right| \leq 2\varepsilon_{\mathrm{EST}} + 2\varepsilon_{\mathrm{RFE}} + \varepsilon_{\mathrm{OPT}}.
$$

□

4.2 REWARD-FREE EXPLORATION METHOD

In this part, we present the RF-Express algorithm in [Ménard et al., 2021] with our notations. Please see Algorithm 1.

Algorithm 1 RF-Express

Input: Failure probability δ, function $\beta(n, \delta) = \log(3|\mathcal{S}||\mathcal{A}|H/\delta) + |\mathcal{S}| \log(8e(n+1))$.

for $t = 0, 1, 2, \cdots$ **do**

Update the counter and the empirical transition model:

$$
\begin{aligned}
&n_h^t(s,a) = \sum_{i=1}^{t} \mathbb{I}\{s_h^i = s, a_h^i = a\},\ n_h^t(s,a,s') = \sum_{i=1}^{t} \mathbb{I}\{s_h^i = s, a_h^i = a, s_{h+1}^i = s'\}, \\
&\widehat{P}_h^t(s'|s,a) = \frac{n_h^t(s,a,s')}{n_h^t(s,a)},\ \text{if } n_h^t(s,a) > 0 \text{ and } \widehat{P}_h^t(s'|s,a) = \frac{1}{|\mathcal{S}|},\ \forall s' \in \mathcal{S} \text{ otherwise.}
\end{aligned}
$$

Define $W_{H+1}^t(s,a) = 0,\ \forall (s,a) \in \mathcal{S} \times \mathcal{A}$.

for $h = H, H-1, \cdots, 1$ **do**

$W_h^t(s,a) = \min\left(H, 15H^2 \frac{\beta\left(n_h^t(s,a),\delta\right)}{n_h^t(s,a)} + \left(1 + \frac{1}{H}\right) \sum_{s' \in \mathcal{S}} \widehat{P}_h^t(s'|s,a) \max_{a'} W_{h+1}^t(s',a') \right)$.

end for

Derive the greedy policy: $\pi_h^{t+1}(s) = \mathrm{argmax}_{a \in \mathcal{A}} W_h^t(s,a), \forall s \in \mathcal{S}, \forall h \in [H]$.

if $3e\sqrt{W_1^t(s_1, \pi_1^{t+1,}(s_1))} + W_1^t(s_1, \pi_1^{t+1}(s_1)) \leq \varepsilon/2$ **then**

break

end if

Rollout π^{t+1} to collect a trajectory $\tau^{t+1} = (s_1^t, a_1^t, s_2^t, a_2^t, \cdots, s_H^t, a_H^t)$.

end for

Output: Transition model $\widehat{P}^t$.

4.3 PROOF OF LEMMA 4

Prior to proving Lemma 4, we first prove that the estimator shown in (7) is an unbiased estimation. We consider the decomposition of $d_h^{\pi^{\mathrm{E}}}(s,a)$.

$$\begin{aligned} d_h^{\pi^{\mathrm{E}}}(s,a) &= \sum_{\mathbf{tr}_h \in \mathrm{Tr}_h^{\mathcal{D}_1}} \mathbb{P}^{\pi^{\mathrm{E}}}(\mathbf{tr}_h)\mathbb{I}\left\{\mathbf{tr}_h(\cdot,\cdot)=(s,a)\right\} + \sum_{\mathbf{tr}_h \notin \mathrm{Tr}_h^{\mathcal{D}_1}} \mathbb{P}^{\pi^{\mathrm{E}}}(\mathbf{tr}_h)\mathbb{I}\left\{\mathbf{tr}_h(\cdot,\cdot)=(s,a)\right\} \\ &= \sum_{\mathbf{tr}_h \in \mathrm{Tr}_h^{\mathcal{D}_1}} \mathbb{P}^{\pi'}(\mathbf{tr}_h)\mathbb{I}\left\{\mathbf{tr}_h(\cdot,\cdot)=(s,a)\right\} + \sum_{\mathbf{tr}_h \notin \mathrm{Tr}_h^{\mathcal{D}_1}} \mathbb{P}^{\pi^{\mathrm{E}}}(\mathbf{tr}_h)\mathbb{I}\left\{\mathbf{tr}_h(\cdot,\cdot)=(s,a)\right\}, \end{aligned} \tag{2}$$

where $\pi' \in \Pi_{\mathrm{BC}}(\mathcal{D}_1)$ and the last equality follows Lemma 1.

Lemma 1. *We define $\Pi_{BC}(\mathcal{D}_1)$ as the set of policies, each of which takes expert action on states contained in $\mathcal{D}_1$. For each $\pi \in \Pi_{BC}(\mathcal{D}_1)$, $\forall h \in [H]$ and $(s,a) \in \mathcal{S}\times\mathcal{A}$, we have*

$$\sum_{\mathbf{tr}_h \in \mathrm{Tr}_h^{\mathcal{D}_1}} \mathbb{P}^{\pi^{\mathrm{E}}}(\mathbf{tr}_h)\mathbb{I}\left\{\mathbf{tr}_h(\cdot,\cdot)=(s,a)\right\} = \sum_{\mathbf{tr}_h \in \mathrm{Tr}_h^{\mathcal{D}_1}} \mathbb{P}^{\pi}(\mathbf{tr}_h)\mathbb{I}\left\{\mathbf{tr}_h(\cdot,\cdot)=(s,a)\right\}.$$

Proof. The proof is based on the fact that any $\pi \in \Pi_{\mathrm{BC}}(\mathcal{D}_1)$ takes the same action with the expert on trajectories in $\mathrm{Tr}_h^{\mathcal{D}_1}$. More concretely, for any $\mathbf{tr}_h \in \mathrm{Tr}_h^{\mathcal{D}_1}$, we have

$$\begin{aligned} &\mathbb{P}^{\pi^{\mathrm{E}}}(\mathbf{tr}_h) \\ &= \rho(\mathbf{tr}_h(s_1))\pi_1^{\mathrm{E}}\left(\mathbf{tr}_h(a_1)|\mathbf{tr}(s_1)\right)\prod_{\ell=1}^{h-1} P_\ell\left(\mathbf{tr}_h(s_{\ell+1})|\mathbf{tr}_h(s_\ell),\mathbf{tr}_h(a_\ell)\right)\pi_{\ell+1}^{\mathrm{E}}\left(\mathbf{tr}_h(a_{\ell+1})|\mathbf{tr}_h(s_{\ell+1})\right) \\ &= \rho(\mathbf{tr}_h(s_1))\pi_1\left(\mathbf{tr}_h(a_1)|\mathbf{tr}(s_1)\right)\prod_{\ell=1}^{h-1} P_\ell\left(\mathbf{tr}_h(s_{\ell+1})|\mathbf{tr}_h(s_\ell),\mathbf{tr}_h(a_\ell)\right)\pi_{\ell+1}\left(\mathbf{tr}_h(a_{\ell+1})|\mathbf{tr}_h(s_{\ell+1})\right) \\ &= \mathbb{P}^{\pi}(\mathbf{tr}_h), \end{aligned}$$

which completes the proof. □

Now we proceed to prove Lemma 4.

Proof of Lemma 4. We aim to upper bound the estimation error $\sum_{h=1}^H \|\widetilde{d}_h^{\pi^{\mathrm{E}}} - d_h^{\pi^{\mathrm{E}}}\|_1$. Recall the definition of the estimator $\widetilde{d}_h^{\pi^{\mathrm{E}}}(s,a)$ in Equation (7):

$$\widetilde{d}_h^{\pi^{\mathrm{E}}}(s,a) := \frac{\sum_{\mathbf{tr}_h \in \mathcal{D}'_{\mathrm{env}}} \mathbb{I}\left\{\mathbf{tr}_h(\cdot,\cdot)=(s,a), \mathbf{tr}_h \in \mathrm{Tr}_h^{\mathcal{D}_1}\right\}}{|\mathcal{D}'_{\mathrm{env}}|} + \frac{\sum_{\mathbf{tr}_h \in \mathcal{D}_1^c} \mathbb{I}\left\{\mathbf{tr}_h(\cdot,\cdot)=(s,a), \mathbf{tr}_h \notin \mathrm{Tr}_h^{\mathcal{D}_1}\right\}}{|\mathcal{D}_1^c|}.$$

Using Equation (2), for any $h \in [H]$ and $(s,a) \in \mathcal{S}\times\mathcal{A}$, we have

$$\begin{aligned} &\left|\widetilde{d}_h^{\pi^{\mathrm{E}}}(s,a) - d_h^{\pi^{\mathrm{E}}}(s,a)\right| \\ &\le \left|\frac{\sum_{\mathbf{tr}_h \in \mathcal{D}'_{\mathrm{env}}} \mathbb{I}\left\{\mathbf{tr}_h(\cdot,\cdot)=(s,a), \mathbf{tr}_h \in \mathrm{Tr}_h^{\mathcal{D}_1}\right\}}{|\mathcal{D}'_{\mathrm{env}}|} - \sum_{\mathbf{tr}_h \in \mathrm{Tr}_h^{\mathcal{D}_1}} \mathbb{P}^{\pi'}(\mathbf{tr}_h)\mathbb{I}\left\{\mathbf{tr}_h(\cdot,\cdot)=(s,a)\right\}\right| \\ &+ \left|\frac{\sum_{\mathbf{tr}_h \in \mathcal{D}_1^c} \mathbb{I}\left\{\mathbf{tr}_h(\cdot,\cdot)=(s,a), \mathbf{tr}_h \notin \mathrm{Tr}_h^{\mathcal{D}_1}\right\}}{|\mathcal{D}_1^c|} - \sum_{\mathbf{tr}_h \notin \mathrm{Tr}_h^{\mathcal{D}_1}} \mathbb{P}^{\pi^{\mathrm{E}}}(\mathbf{tr}_h)\mathbb{I}\left\{\mathbf{tr}_h(\cdot,\cdot)=(s,a)\right\}\right|. \end{aligned}$$

Thus, we can upper bound the estimation error.

$$\sum_{h=1}^H \left\|\widetilde{d}_h^{\pi^{\mathrm{E}}} - d_h^{\pi^{\mathrm{E}}}\right\|_1$$

$$\leq \underbrace{\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left|\frac{\sum_{\mathbf{tr}_h\in\mathcal{D}'_{\text{env}}}\mathbb{I}\left\{\mathbf{tr}_h(\cdot,\cdot)=(s,a),\mathbf{tr}_h\in\text{Tr}_h^{\mathcal{D}_1}\right\}}{|\mathcal{D}'_{\text{env}}|}-\sum_{\mathbf{tr}_h\in\text{Tr}_h^{\mathcal{D}_1}}\mathbb{P}^{\pi'}(\mathbf{tr}_h)\mathbb{I}\left\{\mathbf{tr}_h(\cdot,\cdot)=(s,a)\right\}\right|}_{\text{Error A}}$$

$$+\underbrace{\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left|\frac{\sum_{\mathbf{tr}_h\in\mathcal{D}_1^c}\mathbb{I}\left\{\mathbf{tr}_h(\cdot,\cdot)=(s,a),\mathbf{tr}_h\notin\text{Tr}_h^{\mathcal{D}_1}\right\}}{|\mathcal{D}_1^c|}-\sum_{\mathbf{tr}_h\notin\text{Tr}_h^{\mathcal{D}_1}}\mathbb{P}^{\pi^{\text{E}}}(\mathbf{tr}_h)\mathbb{I}\left\{\mathbf{tr}_h(\cdot,\cdot)=(s,a)\right\}\right|}_{\text{Error B}}.$$

We first analyze the term Error A. Trajectories in $\mathcal{D}'_{\text{env}}$ are collected by π' via interacting with the environment. Thus, we have the estimator in Error A is unbiased, i.e., for any $(s,a)\in\mathcal{S}\times\mathcal{A}$ and $h\in[H]$,

$$\mathbb{E}\left[\frac{\sum_{\mathbf{tr}_h\in\mathcal{D}'_{\text{env}}}\mathbb{I}\left\{\mathbf{tr}_h(\cdot,\cdot)=(s,a),\mathbf{tr}_h\in\text{Tr}_h^{\mathcal{D}_1}\right\}}{|\mathcal{D}'_{\text{env}}|}\right]=\sum_{\mathbf{tr}_h\in\text{Tr}_h^{\mathcal{D}_1}}\mathbb{P}^{\pi'}(\mathbf{tr}_h)\mathbb{I}\left\{\mathbf{tr}_h(\cdot,\cdot)=(s,a)\right\},$$

where the expectation is taken over the randomness of collecting $\mathcal{D}'_{\text{env}}$. The above equality holds because the stochastic processes on the both sides are induced by π'. Then we leverage Chernoff's bound to upper bound Error A.

Lemma 2 (Chernoff's bound [Vershynin, 2018])**.** *Let $\overline{X}=1/n\cdot\sum_{i=1}^{n}X_i$, where X_i is a Bernoulli random variable with $\mathbb{P}(X_i=1)=p_i$ and $\mathbb{P}(X_i=0)=1-p_i$, for $i\in[n]$. Furthermore, assume these random variables are independent. Let $\mu=\mathbb{E}[\overline{X}]=1/n\cdot\sum_{i=1}^{n}p_i$. Then for $0<t\leq 1$,*

$$\mathbb{P}\left(\left|\overline{X}-\mu\right|\geq t\mu\right)\leq 2\exp\left(-\frac{\mu n t^2}{3}\right).$$

First, for each $s\in\mathcal{S}$ and $h\in[H]$, for any non-expert action $a\neq\pi_h^{\text{E}}(s)$, we have that

$$\sum_{\mathbf{tr}_h\in\text{Tr}_h^{\mathcal{D}_1}}\mathbb{P}^{\pi'}(\mathbf{tr}_h)\mathbb{I}\left\{\mathbf{tr}_h(\cdot,\cdot)=(s,a)\right\}=0.$$

This is because on the trajectory $\mathbf{tr}_h\in\text{Tr}_h^{\mathcal{D}_1}$, the state s in time step h is covered in $\mathcal{D}_1$. As a result, the BC policy π' learned from $\mathcal{D}_1$ must take the expert action $\pi_h^{\text{E}}(s)$ on such a state and thus $\mathbb{P}^{\pi'}(\mathbf{tr}_h)\mathbb{I}\left\{\mathbf{tr}_h(\cdot,\cdot)=(s,a)\right\}=0$. Second, since the estimator of

$$\frac{\sum_{\mathbf{tr}_h\in\mathcal{D}'_{\text{env}}}\mathbb{I}\left\{\mathbf{tr}_h(\cdot,\cdot)=(s,a),\mathbf{tr}_h\in\text{Tr}_h^{\mathcal{D}_1}\right\}}{|\mathcal{D}'_{\text{env}}|}$$

is an unbiased estimator and is non-negative almost surely. Therefore, for each $s\in\mathcal{S}$ and $h\in[H]$, for any non-expert action $a\neq\pi_h^{\text{E}}(s)$, with probability of 1,

$$\frac{\sum_{\mathbf{tr}_h\in\mathcal{D}'_{\text{env}}}\mathbb{I}\left\{\mathbf{tr}_h(\cdot,\cdot)=(s,a),\mathbf{tr}_h\in\text{Tr}_h^{\mathcal{D}_1}\right\}}{|\mathcal{D}'_{\text{env}}|}=0.$$

Based on the above two claims, we have that

$$\text{Error A}=\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\left|\frac{\sum_{\mathbf{tr}_h\in\mathcal{D}'_{\text{env}}}\mathbb{I}\left\{\mathbf{tr}_h(\cdot,\cdot)=(s,\pi_h^{\text{E}}(s)),\mathbf{tr}_h\in\text{Tr}_h^{\mathcal{D}_1}\right\}}{|\mathcal{D}'_{\text{env}}|}-\sum_{\mathbf{tr}_h\in\text{Tr}_h^{\mathcal{D}_1}}\mathbb{P}^{\pi'}(\mathbf{tr}_h)\mathbb{I}\left\{\mathbf{tr}_h(\cdot,\cdot)=(s,\pi_h^{\text{E}}(s))\right\}\right|.$$

Let ${E'}_h^s$ be the event that $\mathbf{tr}_h\in\mathcal{D}'_{\text{env}}$ agrees with expert policy at state s at time step h and also appears in $\text{Tr}_h^{\mathcal{D}_1}$. Formally,

$${E'}_h^s=\mathbb{I}\{\mathbf{tr}_h(\cdot,\cdot)=(s,\pi_h^{\text{E}}(s))\cap\mathbf{tr}_h\in\text{Tr}_h^{\mathcal{D}_1}\}.$$

By Lemma 2, for each $s \in \mathcal{S}$ and $h \in [H]$, with probability at least $1 - \frac{\delta}{2|\mathcal{S}|H}$ over the randomness of $\mathcal{D}'_{\text{env}}$, we have

$$\left| \frac{\sum_{\mathbf{tr}_h \in \mathcal{D}'_{\text{env}}} \mathbb{I}\left\{\mathbf{tr}_h(\cdot,\cdot) = (s, \pi_h^{\text{E}}(s)), \mathbf{tr}_h \in \text{Tr}_h^{\mathcal{D}_1}\right\}}{|\mathcal{D}'_{\text{env}}|} - \sum_{\mathbf{tr}_h \in \text{Tr}_h^{\mathcal{D}_1}} \mathbb{P}^{\pi'}(\mathbf{tr}_h) \mathbb{I}\left\{\mathbf{tr}_h(\cdot,\cdot) = (s, \pi_h^{\text{E}}(s))\right\} \right|$$
$$\leq \sqrt{\mathbb{P}^{\pi'}\left(E'^s_h\right) \frac{3\log\left(4|\mathcal{S}|H/\delta\right)}{n'}}.$$

By union bound, with probability at least $1 - \frac{\delta}{2}$ over the randomness of $\mathcal{D}'_{\text{env}}$, we have

$$\begin{aligned}\text{Error A} &\leq \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} \sqrt{\mathbb{P}^{\pi'}\left(E'^s_h\right) \frac{3\log\left(4|\mathcal{S}|H/\delta\right)}{n'}} \\ &\leq \sum_{h=1}^{H} \sqrt{|\mathcal{S}|} \sqrt{\sum_{s \in \mathcal{S}} \mathbb{P}^{\pi'}\left(E'^s_h\right) \frac{3\log\left(4|\mathcal{S}|H/\delta\right)}{n'}}\end{aligned}$$

The last inequality follows the Cauchy-Schwartz inequality. It remains to upper bound $\sum_{s \in \mathcal{S}} \mathbb{P}^{\pi^{\text{E}}}(E'^s_h)$ for all $h \in [H]$. To this end, we define the event $G'^{\mathcal{D}_1}_h$ that policy π' visits states covered in $\mathcal{D}_1$ up to time step h. Formally, $G'^{\mathcal{D}_1}_h = \mathbb{I}\{\forall h' \leq h, s_{h'} \in \mathcal{S}_{h'}(\mathcal{D}_1)\}$, where $\mathcal{S}_h(\mathcal{D}_1)$ is the set of states in $\mathcal{D}_1$ at time step h, where s'_h comes from $\mathbf{tr}_h \in \mathcal{D}'_{\text{env}}$. Then, for all $h \in [H]$, we have

$$\sum_{s \in \mathcal{S}} \mathbb{P}^{\pi'}\left(E'^s_h\right) = \mathbb{P}^{\pi'}(G'^{\mathcal{D}_1}_h) \leq \mathbb{P}(G'^{\mathcal{D}_1}_1).$$

The last inequality holds since $G'^{\mathcal{D}_1}_h \subseteq G'^{\mathcal{D}_1}_1$ for all $h \in [H]$. Then we have that

$$\text{Error A} \leq H\sqrt{\frac{3|\mathcal{S}|\log\left(4|\mathcal{S}|H/\delta\right)}{n'}}.$$

When the interaction complexity satisfies that $n' \gtrsim \frac{|\mathcal{S}|H^2}{\varepsilon^2} \log\left(\frac{|\mathcal{S}|H}{\delta}\right)$, with probability at least $1 - \frac{\delta}{2}$ over the randomness of $\mathcal{D}'$, we have Error A $\leq \frac{\varepsilon}{2}$.

For the term Error B, we utilize [Rajaraman et al., 2020, Lemma A.11]. When the expert sample complexity satisfies that $m \gtrsim \frac{|\mathcal{S}|H^{3/2}}{\varepsilon} \log\left(\frac{|\mathcal{S}|H}{\delta}\right)$, with probability at least $1 - \frac{\delta}{2}$ over the randomness of $\mathcal{D}$, we have Error B $\leq \frac{\varepsilon}{2}$. Applying union bound finishes the proof.

□

4.4 PROOF OF LEMMA 5

Before we prove Lemma 5, we first state the following key lemma.

Lemma 3. *Consider Algorithm 2, we have*

$$\sum_{t=1}^{T} f^{(t)}\left(w^{(t)}\right) - \min_{w \in \mathcal{W}} \sum_{t=1}^{T} f^{(t)}(w) \leq 2H\sqrt{2|\mathcal{S}||\mathcal{A}|T},$$

where $f^{(t)}(w) = \sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} w_h(s,a)(d_h^{\pi^{(t)}, \widehat{P}}(s,a) - \widetilde{d}_h^{\pi^{\text{E}}}(s,a))$.

Proof. Lemma 3 is a direct consequence of the regret bound of online gradient descent [Shalev-Shwartz, 2012]. To apply such a regret bound, we need to verify that 1) the iterate norm $\|w\|_2$ has an upper bound; 2) the gradient norm $\|\nabla_w f^{(t)}(w)\|_2$ also has an upper bound. The first point is easy to show, i.e., $\|w\|_2 \leq \sqrt{H|\mathcal{S}||\mathcal{A}|}$ by the condition that

$w \in \mathcal{W} = \{w = (w_1, \ldots, w_H) : \|w_h\|_\infty \leq 1,\ \forall h \in [H]\}$. For the second point, let $\widetilde{d}_h^1$ and $\widetilde{d}_h^2$ be the first and the second part in $\widetilde{d}_h^{\pi^{\mathrm{E}}}$ defined in (7). Then,

$$\begin{aligned}
\left\|\nabla_w f^{(t)}(w)\right\|_2 &= \sqrt{\sum_{h=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left(d_h^{\pi^{(t)},\widehat{P}}(s,a) - \widetilde{d}_h^{\pi^{\mathrm{E}}}(s,a)\right)^2} \\
&= \sqrt{\sum_{h=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left(d_h^{\pi^{(t)},\widehat{P}}(s,a) - \widetilde{d}_h^1(s,a) - \widetilde{d}_h^2(s,a)\right)^2} \\
&\leq \sqrt{\sum_{h=1}^{H} 3 \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left(d_h^{\pi^{(t)},\widehat{P}}(s,a)\right)^2 + \left(\widetilde{d}_h^1(s,a)\right)^2 + \left(\widetilde{d}_h^2(s,a)\right)^2} \\
&\leq \sqrt{\sum_{h=1}^{H} 3 \left(\left\|d_h^{\pi^{(t)},\widehat{P}}\right\|_1 + \left\|\widetilde{d}_h^1\right\|_1 + \left\|\widetilde{d}_h^2\right\|_1\right)} \\
&\leq 2\sqrt{H},
\end{aligned}$$

where the first inequality follows $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$ and the second inequality is based on that $x^2 \leq |x|$ if $0 \leq x \leq 1$.

Invoking Corollary 2.7 in [Shalev-Shwartz, 2012] with $B = \sqrt{H|\mathcal{S}||\mathcal{A}|}$ and $L = 2\sqrt{H}$ finishes the proof. □

Proof of Lemma 5. With the dual representation of ℓ_1-norm, we have

$$\min_{\pi\in\Pi} \sum_{h=1}^{H} \left\|d_h^{\pi,\widehat{P}} - \widetilde{d}_h^{\pi^{\mathrm{E}}}\right\|_1 = \min_{\pi\in\Pi} \max_{w\in\mathcal{W}} \sum_{h=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} w_h(s,a) \left(\widetilde{d}_h^{\pi^{\mathrm{E}}}(s,a) - d_h^{\pi,\widehat{P}}(s,a)\right).$$

Since the above objective is linear w.r.t both w and d_h^π, invoking the minimax theorem [Bertsekas, 2016] yields

$$\begin{aligned}
&\min_{\pi\in\Pi} \max_{w\in\mathcal{W}} \sum_{h=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} w_h(s,a) \left(\widetilde{d}_h^{\pi^{\mathrm{E}}}(s,a) - d_h^{\pi,\widehat{P}}(s,a)\right) \\
&= \max_{w\in\mathcal{W}} \min_{\pi\in\Pi} \sum_{h=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} w_h(s,a) \left(\widetilde{d}_h^{\pi^{\mathrm{E}}}(s,a) - d_h^{\pi,\widehat{P}}(s,a)\right) \\
&= -\min_{w\in\mathcal{W}} \max_{\pi\in\Pi} \sum_{h=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} w_h(s,a) \left(d_h^{\pi,\widehat{P}}(s,a) - \widetilde{d}_h^{\pi^{\mathrm{E}}}(s,a)\right),
\end{aligned}$$

where the last step follows the property that for a function f, $-\max_x f(x) = \min_x -f(x)$. Therefore, we have

$$\min_{\pi\in\Pi} \sum_{h=1}^{H} \left\|d_h^{\pi,\widehat{P}} - \widetilde{d}_h^{\pi^{\mathrm{E}}}\right\|_1 = -\min_{w\in\mathcal{W}} \max_{\pi\in\Pi} \sum_{h=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} w_h(s,a) \left(d_h^{\pi,\widehat{P}}(s,a) - \widetilde{d}_h^{\pi^{\mathrm{E}}}(s,a)\right). \tag{3}$$

Then we consider the term $\min_{w\in\mathcal{W}} \max_{\pi\in\Pi} \sum_{h=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} w_h(s,a) \left(d_h^{\pi,\widehat{P}}(s,a) - \widetilde{d}_h^{\pi^{\mathrm{E}}}(s,a)\right)$.

$$\begin{aligned}
&\min_{w\in\mathcal{W}} \max_{\pi\in\Pi} \sum_{h=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} w_h(s,a) \left(d_h^{\pi,\widehat{P}}(s,a) - \widetilde{d}_h^{\pi^{\mathrm{E}}}(s,a)\right) \\
&\leq \max_{\pi\in\Pi} \sum_{h=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left(\frac{1}{T}\sum_{t=1}^{T} w_h^{(t)}(s,a)\right) \left(d_h^{\pi,\widehat{P}}(s,a) - \widetilde{d}_h^{\pi^{\mathrm{E}}}(s,a)\right) \\
&\leq \frac{1}{T}\sum_{t=1}^{T} \max_{\pi\in\Pi} \sum_{h=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} w_h^{(t)}(s,a) \left(d_h^{\pi,\widehat{P}}(s,a) - \widetilde{d}_h^{\pi^{\mathrm{E}}}(s,a)\right).
\end{aligned}$$

At iteration t, $\pi^{(t)}$ is the approximately optimal policy regarding reward function $w^{(t)}$ with an optimization error of $\varepsilon_{\mathrm{RL}}$. Then we obtain that

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\max_{\pi\in\Pi}\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} w_h^{(t)}(s,a)\left(d_h^{\pi,\widehat{P}}(s,a)-\widetilde{d}_h^{\pi^{\mathrm{E}}}(s,a)\right)\\
\leq&\frac{1}{T}\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} w_h^{(t)}(s,a)\left(d_h^{\pi^{(t)},\widehat{P}}(s,a)-\widetilde{d}_h^{\pi^{\mathrm{E}}}(s,a)\right)+\varepsilon_{\mathrm{RL}}.
\end{aligned}$$

Applying Lemma 3 yields that

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} w_h^{(t)}(s,a)\left(d_h^{\pi^{(t)},\widehat{P}}(s,a)-\widetilde{d}_h^{\pi^{\mathrm{E}}}(s,a)\right)\\
\leq&\min_{w\in\mathcal{W}}\frac{1}{T}\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} w_h(s,a)\left(d_h^{\pi^{(t)},\widehat{P}}(s,a)-\widetilde{d}_h^{\pi^{\mathrm{E}}}(s,a)\right)+2H\sqrt{\frac{2|\mathcal{S}||\mathcal{A}|}{T}}\\
=&\min_{w\in\mathcal{W}}\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} w_h(s,a)\left(\frac{1}{T}\sum_{t=1}^{T}d_h^{\pi^{(t)},\widehat{P}}(s,a)-\widetilde{d}_h^{\pi^{\mathrm{E}}}(s,a)\right)+2H\sqrt{\frac{2|\mathcal{S}||\mathcal{A}|}{T}}\\
=&\min_{w\in\mathcal{W}}\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} w_h(s,a)\left(d_h^{\overline{\pi},\widehat{P}}(s,a)-\widetilde{d}_h^{\pi^{\mathrm{E}}}(s,a)\right)+2H\sqrt{\frac{2|\mathcal{S}||\mathcal{A}|}{T}}.
\end{aligned}$$

Note that $\overline{\pi}$ is induced by the mean state-action distribution, i.e., $\overline{\pi}_h(a|s)=\overline{P}_h(s,a)/\sum_a\overline{P}_h(s,a)$, where $\overline{P}_h(s,a)=\frac{1}{T}\sum_{t=1}^{T}d_h^{\pi^{(t)},\widehat{P}}(s,a)$. Based on Proposition 3.1 in [Ho and Ermon, 2016], we have that $d_h^{\overline{\pi},\widehat{P}}(s,a)=\overline{P}_h(s,a)$, and hence the last equation holds. Combined with Equation (3), we have that

$$\begin{aligned}
&\min_{\pi\in\Pi}\sum_{h=1}^{H}\left\|d_h^{\pi,\widehat{P}}-\widetilde{d}_h^{\pi^{\mathrm{E}}}\right\|_1\\
\geq&-\min_{w\in\mathcal{W}}\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} w_h(s,a)\left(d_h^{\overline{\pi},\widehat{P}}(s,a)-\widetilde{d}_h^{\pi^{\mathrm{E}}}(s,a)\right)-2H\sqrt{\frac{2|\mathcal{S}||\mathcal{A}|}{T}}-\varepsilon_{\mathrm{RL}}\\
=&\max_{w\in\mathcal{W}}\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} w_h(s,a)\left(\widetilde{d}_h^{\pi^{\mathrm{E}}}(s,a)-d_h^{\overline{\pi},\widehat{P}}(s,a)\right)-2H\sqrt{\frac{2|\mathcal{S}||\mathcal{A}|}{T}}-\varepsilon_{\mathrm{RL}}\\
=&\left\|\widetilde{d}_h^{\pi^{\mathrm{E}}}-d_h^{\overline{\pi},\widehat{P}}\right\|_1-2H\sqrt{\frac{2|\mathcal{S}||\mathcal{A}|}{T}}-\varepsilon_{\mathrm{RL}},
\end{aligned}$$

where the last step again utilizes the dual representation of ℓ_1-norm. If we take $\varepsilon_{\mathrm{RL}}\leq\varepsilon/2$, $T\gtrsim H^2|\mathcal{S}||\mathcal{A}|/\varepsilon^2$ and $\eta^{(t)}:=\sqrt{|\mathcal{S}||\mathcal{A}|/(8T)}$, then we have

$$\sum_{h=1}^{H}\left\|d_h^{\overline{\pi},\widehat{P}}-\widetilde{d}_h^{\pi^{\mathrm{E}}}\right\|_1\leq\min_{\pi\in\Pi}\sum_{h=1}^{H}\left\|d_h^{\pi,\widehat{P}}-\widetilde{d}_h^{\pi^{\mathrm{E}}}\right\|_1+\varepsilon.$$

We complete the proof. □

4.5 PROOF OF THEOREM 1

Proof of Theorem 1. Firstly, we verify assumption (a) in Proposition 1. With Lemma 2, when the number of trajectories collected by RF-Express satisfies

$$n\gtrsim\frac{H^3|\mathcal{S}||\mathcal{A}|}{\varepsilon^2}\left(|\mathcal{S}|+\log\left(\frac{|\mathcal{S}|H}{\delta}\right)\right),$$

for any policy $\pi \in \Pi$ and reward function $w : \mathcal{S} \times \mathcal{A} \to [0, 1]$, with probability at least $1 - \delta/2$, $|V^{\pi,P,w} - V^{\pi,\widehat{P},w}| \leq \varepsilon/16 = \varepsilon_{\mathrm{RFE}}$. In a word, the assumption (a) in Proposition 1 holds with $\delta_{\mathrm{RFE}} = \delta/2$ and $\varepsilon_{\mathrm{RFE}} = \varepsilon/16$.

Secondly, we note that the assumption (b) in Proposition 1 holds by Lemma 4. More concretely, if the expert sample complexity and interaction complexity satisfies

$$m \gtrsim \frac{H^{3/2}|\mathcal{S}|}{\varepsilon} \log\left(\frac{|\mathcal{S}|H}{\delta}\right),\ n' \gtrsim \frac{H^2|\mathcal{S}|}{\varepsilon^2} \log\left(\frac{|\mathcal{S}|H}{\delta}\right),$$

with probability at least $1 - \delta/2$, $\sum_{h=1}^{H} \|\widetilde{d}_h^{\pi^{\mathrm{E}}} - d_h^{\pi^{\mathrm{E}}}\|_1 \leq \varepsilon/16 = \varepsilon_{\mathrm{EST}}$. Hence, the assumption (b) in Proposition 1 holds with $\delta_{\mathrm{EST}} = \delta/2$ and $\varepsilon_{\mathrm{EST}} = \varepsilon/16$.

Thirdly, we aim to verify that the assumption (c) in Proposition 1 holds with $\widetilde{d}_h^{\pi^{\mathrm{E}}}(s, a)$ and $\widehat{P}$. When $\varepsilon_{\mathrm{RL}} \leq \varepsilon/2$ and $T \gtrsim |\mathcal{S}||\mathcal{A}|H^2/\varepsilon^2$ such that $2H\sqrt{2|\mathcal{S}||\mathcal{A}|/T} \leq \varepsilon/4$, we have that

$$\sum_{h=1}^{H} \left\| d_h^{\overline{\pi},\widehat{P}} - \widetilde{d}_h^{\pi^{\mathrm{E}}} \right\|_1 - \min_{\pi \in \Pi} \sum_{h=1}^{H} \left\| d_h^{\pi,\widehat{P}} - \widetilde{d}_h^{\pi^{\mathrm{E}}} \right\|_1 \leq \frac{3\varepsilon}{4} = \varepsilon_{\mathrm{OPT}}.$$

Therefore, the assumption (c) in Proposition 1 holds with $\varepsilon_{\mathrm{OPT}} = 3\varepsilon/4$. Now, we summarize the conditions what we have obtained.

- The assumption (a) in Proposition 1 holds with $\delta_{\mathrm{RFE}} = \delta/2$ and $\varepsilon_{\mathrm{RFE}} = \varepsilon/16$.
- The assumption (b) in Proposition 1 holds with $\delta_{\mathrm{EST}} = \delta/2$ and $\varepsilon_{\mathrm{EST}} = \varepsilon/16$.
- The assumption (c) in Proposition 1 holds with $\varepsilon_{\mathrm{OPT}} = 3\varepsilon/4$.

Applying Proposition 1 finishes the proof. With probability at least $1 - \delta$,

$$V^{\pi^{\mathrm{E}}} - V^{\overline{\pi}} \leq 2\varepsilon_{\mathrm{RFE}} + 2\varepsilon_{\mathrm{EST}} + \varepsilon_{\mathrm{OPT}} = \varepsilon.$$

□

5 PROOF OF RESULTS IN SECTION 6

5.1 PROBLEM SETUP

To facilitate later analysis, we introduce some useful notations widely used in the literature [Li et al., 2006, Jiang et al., 2015]. In this part, for a function f that operates on the original state space $\mathcal{S}$, we add a superscript ϕ (i.e., f^ϕ) to denote the counterpart that operates on the abstract state space Φ. Inversely, for a function f^ϕ that operates on the abstract state space, we use $[f^\phi]^M$ to denote its lifted version, which is defined as $[f^\phi]^M(s) = f^\phi(\phi(s))$. Notice that $[f^\phi]^M$ is a function over $\mathcal{S}$.

Definition 1 (Abstract MDP)**.** *Under Assumption 1, for the original MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, H, \rho)$, we define the abstract MDP $\mathcal{M}^\phi = (\Phi, \mathcal{A}, P^\phi, r^\phi, H, \rho^\phi)$. In particular,*

- $P_h^\phi(x'|x, a) = \sum_{s' \in \phi_h^{-1}(x')} P_h(s'|s, a)$*, for an arbitrary* $s \in \phi_h^{-1}(x)$.
- $r_h^\phi(x, a) = r_h(s, a)$*, for an arbitrary* $s \in \phi_h^{-1}(x)$.
- $\rho^\phi(x) = \sum_{s \in \phi_1^{-1}(x)} \rho(s, a)$.

Here $\phi_h^{-1}(x) = \{s \in \mathcal{S} : \phi_h(s) = x\}$.

We clarify that there is no ambiguity in Definition 1 because of Assumption 1. The bisimulation condition enables that $s \in \phi_h^{-1}(x)$ are equivalent under the reward-consistent and transition-consistent conditions. With the abstract MDP $\mathcal{M}^\phi$, for any abstract policy π^ϕ, we utilize $V_h^{\pi^\phi,\mathcal{M}^\phi}(x)$ to denote the corresponding value function. Similarly, with the original MDP $\mathcal{M}$, for any policy π, we use $V_h^{\pi,\mathcal{M}}(s)$ to denote the corresponding value function.

Definition 2 (Abstract Expert Policy)**.** *Under Assumption 1, for the original expert policy π^{E}, we define the abstract expert policy $\pi^{\mathrm{E},\phi}$. In particular, for any $(x,h) \in \Phi \times [H]$, it holds that*

$$\pi_h^{\mathrm{E},\phi}(x) = \pi_h^{\mathrm{E}}(s), \text{ for an arbitrary } s \in \phi_h^{-1}(x).$$

Besides, for any policy $\pi \in \Pi$, we utilize $d_h^{\pi,\phi} \in \Delta(\Phi \times \mathcal{A})$ to denote the abstract state-action distribution.

$$d_h^{\pi,\phi}(x,a) = \mathbb{P}^{\pi}(\phi_h(s_h) = x, a_h = a|P) = \sum_{s \in \phi_h^{-1}(x)} d_h^{\pi}(s,a).$$

For any abstract policy $\pi^{\phi} \in \Pi^{\phi}$ and abstract transition function P^{ϕ}, we utilize $d_h^{\pi^{\phi},P^{\phi}} \in \Delta(\Phi \times \mathcal{A})$ to denote the abstract state-action distribution induced by π^{ϕ} in P^{ϕ}. In particular,

$$d_h^{\pi^{\phi},P^{\phi}}(x,a) = \mathbb{P}^{\pi^{\phi}}(x_h = x, a_h = a|P^{\phi}).$$

5.2 MB-TAIL WITH STATE ABSTRACTION

Before presenting MB-TAIL with state abstraction, we first develop a meta-algorithm for AIL with state abstractions when the transition function is unknown.

Algorithm 2 Meta-algorithm for AIL with State Abstractions and Unknown Transitions

Input: Expert demonstrations $\mathcal{D}$, a set of state abstractions $\{\phi_h\}_{h=1}^H$.

1: $\widehat{P}^{\phi} \leftarrow$ Invoke a reward-free exploration method to collect n trajectories and learn an *abstract* transition model.

2: $\widetilde{d}_h^{\pi^{\mathrm{E}},\phi} \leftarrow$ Estimate the *abstract* expert state-action distribution.

3: $\overline{\pi}^{\phi} \leftarrow$ Apply an AIL approach to perform imitation with the expert estimation $\widetilde{d}_h^{\pi^{\mathrm{E}},\phi}$ under transition model $\widehat{P}^{\phi}$.

Output: Policy $[\overline{\pi}^{\phi}]^M$.

In the sequel, we present three main algorithmic designs that appeared in Line 1, Line 2 and Line 3 in Algorithm 2 in the setting with state abstraction.

The Transition-aware Estimator with State Abstraction. Here we present the transition-aware estimator with state abstraction. The key idea of the construction of the estimator is similar to that discussed in Section 5.2. However, unlike the original estimator in (7), the transition-aware estimator with state abstraction is a distribution over the abstract space $\Phi \times \mathcal{A}$. We present our adaptions to the setting with state abstraction in the following part.

Similar to the procedure presented in Section 5.2, we randomly divide the expert dataset into two equal parts, i.e., $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_1^c$ and $\mathcal{D}_1 \cap \mathcal{D}_1^c = \emptyset$ with $|\mathcal{D}_1| = |\mathcal{D}_1^c| = m/2$. First, with state abstractions $\{\phi_h\}_{h=1}^H$, we first apply BC on $\mathcal{D}_1$ to learn the abstract policy $\pi'^{,\phi}$.

$$\pi_h'^{,\phi}(a|x) = \begin{cases} \frac{n_h^1(x,a)}{n_h^1(x)} & \text{if } n_h^1(x) > 0 \\ \frac{1}{|\mathcal{A}|} & \text{otherwise} \end{cases} \tag{4}$$

Here $n_h^1(x,a) = \sum_{\mathbf{tr} \in \mathcal{D}_1} \mathbb{I}\{\phi_h(\mathbf{tr}_h(\cdot)) = x, \mathbf{tr}_h(a_h) = a\}$ and $n_h^1(x) = \sum_{\mathbf{tr} \in \mathcal{D}_1} \mathbb{I}\{\phi_h(\mathbf{tr}_h(\cdot)) = x\}$. Intuitively, $n_h^1(x,a)$ ($n_h^1(x)$) is the number of abstract-state-action (abstract state) pairs that appeared in $\mathcal{D}_1$ in step h.

Second, we utilize the lifted policy $[\pi'^{,\phi}]^M$ to interact with the environment to collect a new dataset $\mathcal{D}'_{\mathrm{env}}$. Notice that $[\pi'^{,\phi}]^M$ is a policy defined in the original state space $\mathcal{S}$. Finally, we can establish the following estimator with state abstractions $\{\phi_h\}_{h=1}^H$.

$$\begin{aligned} \widetilde{d}_h^{\pi^{\mathrm{E}},\phi}(x,a) = & \frac{\sum_{\mathbf{tr}_h \in \mathcal{D}'_{\mathrm{env}}} \mathbb{I}\{\phi_h(\mathbf{tr}_h(\cdot)) = x, \mathbf{tr}_h(a_h) = a, \mathbf{tr}_h \in \mathrm{Tr}_h^{\mathcal{D}_1,\phi}\}}{|\mathcal{D}'_{\mathrm{env}}|} \\ & + \frac{\sum_{\mathbf{tr}_h \in \mathcal{D}_1^c} \mathbb{I}\{\phi_h(\mathbf{tr}_h(\cdot)) = x, \mathbf{tr}_h(a_h) = a, \mathbf{tr}_h \notin \mathrm{Tr}_h^{\mathcal{D}_1,\phi}\}}{|\mathcal{D}_1^c|}. \end{aligned} \tag{5}$$

Here

$$\mathrm{Tr}_h^{\mathcal{D},\phi} = \{\mathbf{tr}_h = (s_1, a_1, \ldots, s_h, a_h) : \phi_\ell(s_\ell) \in \Phi_\ell(\mathcal{D}), \forall \ell \in [h]\},\ \Phi_h(\mathcal{D}) = \{x \in \Phi : \exists \mathbf{tr} \in \mathcal{D}, \phi_h(\mathbf{tr}_h(\cdot)) = x\}.$$

Intuitively, $\Phi_h(\mathcal{D})$ is the set of abstract states visited in $\mathcal{D}$ in time step h. $\mathrm{Tr}_h^{\mathcal{D},\phi}$ is the set of truncated trajectories of length h, along which each abstract state is visited in $\mathcal{D}$.

Reward-free Exploration with State Abstraction. In this part, we adapt the reward-free exploration method RF-Express to the setting with state abstraction; see Algorithm 3. The main difference is that we learn the abstract transition model and abstract exploration policy. Nevertheless, when interacting with the original environment, we need to transfer the abstract policy $\pi^{\phi,t+1}$ to the lifted version $[\pi^{\phi,t+1}]^M$.

Algorithm 3 RF-Express with State Abstraction

Input: A set of state abstractions $\{\phi_h\}_{h=1}^H$, failure probability δ, and function $\beta(n,\delta) = \log(3|\Phi||\mathcal{A}|H/\delta) + |\Phi|\log(8e(n+1))$.

1: **for** $t = 0, 1, 2, \cdots$ **do**

2: Update the abstract counter and abstract empirical transition model:

$$n_h^t(x,a) = \sum_{i=1}^t \mathbb{I}\{\phi_h(s_h^i) = x, a_h^i = a\},\ n_h^t(x,a,x') = \sum_{i=1}^t \mathbb{I}\{\phi_h(s_h^i) = x, a_h^i = a, \phi_{h+1}(s_{h+1}^i) = x'\},$$

$$\widehat{P}_h^{\phi,t}(x'|x,a) = \frac{n_h^t(x,a,x')}{n_h^t(x,a)},\ \text{if } n_h^t(x,a) > 0 \text{ and } \widehat{P}_h^{\phi,t}(x'|x,a) = \frac{1}{|\mathcal{S}|},\ \forall x' \in \Phi \text{ otherwise.}$$

3: Define $W_{H+1}^t(x,a) = 0,\ \forall (x,a) \in \Phi \times \mathcal{A}$.

4: **for** $h = H, H-1, \cdots, 1$ **do**

5: $W_h^t(x,a) = \min\left(H, 15H^2 \frac{\beta(n_h^t(x,a),\delta)}{n_h^t(x,a)} + \left(1 + \frac{1}{H}\right)\sum_{x' \in \Phi} \widehat{P}_h^{\phi,t}(x'|x,a) \max_{a'} W_{h+1}^t(x',a')\right)$.

6: **end for**

7: Derive the greedy policy: $\pi_h^{\phi,t+1}(x) = \mathrm{argmax}_{a \in \mathcal{A}} W_h^t(x,a), \forall x \in \Phi, \forall h \in [H]$.

8: **if** $3e\sqrt{W_1^t(\phi_1(s_1), \pi_1^{\phi,t+1}(\phi_1(s_1)))} + W_1^t(\phi_1(s_1), \pi_1^{\phi,t+1}(\phi_1(s_1))) \le \varepsilon/2$ **then**

9: **break**

10: **end if**

11: Rollout $[\pi^{\phi,t+1}]^M$ to collect a trajectory $\tau^{t+1} = (s_1^{t+1}, a_1^{t+1}, s_2^{t+1}, a_2^{t+1}, \cdots, s_H^{t+1}, a_H^{t+1})$.

12: **end for**

Output: Transition model $\widehat{P}^{\phi,t}$.

Gradient-based Optimization. For Line 3 in Algorithm 2, we aim to solve the following state-action distribution matching problem.

$$\min_{\pi^\phi \in \Pi^\phi} \sum_{h=1}^H \left\| \widetilde{d}_h^{\pi^{\mathrm{E}},\phi} - d_h^{\pi^\phi, \widehat{P}^\phi} \right\|_1,$$

Notice that this is precisely the optimization problem of projecting $\widetilde{d}_h^{\pi^{\mathrm{E}},\phi}$ on the set of all feasible *abstract* state-action distributions. We can still apply Algorithm 2 with inputs of $\widehat{P}^\phi$ and $\widetilde{d}_h^{\pi^{\mathrm{E}},\phi}$ to solve this optimization problem.

Finally, we combine the above three algorithmic designs under the developed framework (Algorithm 2), which yields the final algorithm.

5.3 PROOF OF THEOREM 2

Prior to proving Theorem 2, we provide a theoretical guarantee for the meta-algorithm presented in Algorithm 2. The algorithm constructs an abstract transition model, an abstract state-action distribution and an abstract policy. Finally, the algorithm outputs a policy that can operate in the original state space. To accomplish this, we introduce specialized analysis tools to connect these concepts in both the original and abstract spaces.

Algorithm 4 Model-based Transition-aware AIL with State Abstractions

Input: Expert demonstrations $\mathcal{D}$, and a set of state abstractions $\{\phi_h\}_{h=1}^H$.

1: Randomly split $\mathcal{D}$ into two equal parts: $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_1^c$.
2: Learn an abstract policy $\pi'^{,\phi} \in \Pi_{\mathrm{BC}}(\mathcal{D}_1)$ by BC with $\{\phi_h\}_{h=1}^H$ and roll out $[\pi'^{,\phi}]^M$ to obtain dataset $\mathcal{D}'_{\mathrm{env}}$ with $|\mathcal{D}'_{\mathrm{env}}| = n'$.
3: Obtain the abstract estimator $\widetilde{d}_h^{\pi^{\mathrm{E}},\phi}$ in (5) with $\mathcal{D}$, $\mathcal{D}'_{\mathrm{env}}$ and $\{\phi_h\}_{h=1}^H$.
4: Invoke Algorithm 3 to collect n trajectories and learn an abstract empirical transition function $\widehat{P}^\phi$.
5: $\overline{\pi}^\phi \leftarrow$ Apply Algorithm 2 with the estimation $\widetilde{d}_h^{\pi^{\mathrm{E}},\phi}$ under transition model $\widehat{P}^\phi$.

Output: Policy $[\overline{\pi}^\phi]^M$.

Proposition 1. *Suppose that*

(a) an algorithm A solves the reward-free exploration problem on the abstract *MDP $\mathcal{M}^\phi$ (see Definition 3) up to an error $\varepsilon_{\mathrm{RFE}}$ with probability at least $1-\delta_{\mathrm{RFE}}$.*

(b) an algorithm B has an abstract *state-action distribution estimator for $d_h^{\pi^{\mathrm{E}},\phi}$, which satisfies $\sum_{h=1}^H \|\widetilde{d}_h^{\pi^{\mathrm{E}},\phi} - d_h^{\pi^{\mathrm{E}},\phi}\|_1 \le \varepsilon_{\mathrm{EST}}$, with probability at least $1-\delta_{\mathrm{EST}}$;*

(c) with the abstract *transition model in (a) and the* abstract *estimator in (b), an algorithm C solves the following optimization problem up to an error $\varepsilon_{\mathrm{OPT}}$.*

$$\min_{\pi^\phi \in \Pi^\phi} \sum_{h=1}^H \left\| \widetilde{d}_h^{\pi^{\mathrm{E}},\phi} - d_h^{\pi^\phi,\widehat{P}^\phi} \right\|_1, \tag{6}$$

where Π^ϕ is the set of all abstract policies and $d_h^{\pi^\phi,\widehat{P}^\phi}$ is the abstract state-action distribution induced by the abstract policy π^ϕ and abstract transition function $\widehat{P}^\phi$.

Then applying algorithms A, B and C under the framework in Algorithm 2 could return a policy $[\overline{\pi}^\phi]^M$, which has a policy value gap (i.e., $V^{\pi^{\mathrm{E}}} - V^{[\overline{\pi}^\phi]^M}$) at most $2\varepsilon_{\mathrm{EST}} + 2\varepsilon_{\mathrm{RFE}} + \varepsilon_{\mathrm{OPT}}$, with probability at least $1-\delta_{\mathrm{EST}}-\delta_{\mathrm{RFE}}$.

Proof. The proof idea is similar to that in Section 4.1. Additionally, we leverage the analysis techniques in state abstraction. We want to upper bound the imitation gap $V^{\pi^{\mathrm{E}},\mathcal{M}} - V^{[\overline{\pi}^\phi]^M,\mathcal{M}}$, where $V^{\pi,\mathcal{M}}$ represents the policy value of π on the original MDP $\mathcal{M}$. We consider the following two events.

$$E_{\mathrm{EST}} = \left\{ \sum_{h=1}^H \left\| \widetilde{d}_h^{\pi^{\mathrm{E}},\phi} - d_h^{\pi^{\mathrm{E}},\phi} \right\|_1 \le \varepsilon_{\mathrm{EST}} \right\}$$
$$E_{\mathrm{RFE}} = \left\{ \forall \pi^\phi \in \Pi^\phi,\ r^\phi = (r_1^\phi, \ldots, r_H^\phi),\ r_h^\phi : \Phi \times \mathcal{A} \to [0,1],\ |V^{\pi^\phi,P^\phi,r^\phi} - V^{\pi^\phi,\widehat{P}^\phi,r^\phi}| \le \varepsilon_{\mathrm{RFE}} \right\}.$$

With condition (a) and condition (b), we obtain $\mathbb{P}(E_{\mathrm{EST}} \cap E_{\mathrm{RFE}}) \ge 1 - \delta_{\mathrm{RFE}} - \delta_{\mathrm{EST}}$. The following analysis is established on the event $E_{\mathrm{EST}} \cap E_{\mathrm{RFE}}$.

By Lemma 6, we have $V^{[\overline{\pi}^\phi]^M,\mathcal{M}} = V^{\overline{\pi}^\phi,\mathcal{M}^\phi}$, where $\mathcal{M}^\phi$ is the abstract MDP in Definition 1. Then we can upper bound the term $V^{\pi^{\mathrm{E}},\mathcal{M}} - V^{\overline{\pi}^\phi,\mathcal{M}^\phi}$. On the event E_{RFE}, we further have

$$\begin{aligned}
V^{\pi^{\mathrm{E}},\mathcal{M}} - V^{\overline{\pi}^\phi,\mathcal{M}^\phi} &\le V^{\pi^{\mathrm{E}},\mathcal{M}} - V^{\overline{\pi}^\phi,\widehat{P}^\phi,r^\phi} + \varepsilon_{\mathrm{RFE}} \\
&= \sum_{h=1}^H \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^{\mathrm{E}}}(s,a) r_h(s,a) - \sum_{h=1}^H \sum_{(x,a)\in\Phi\times\mathcal{A}} d_h^{\overline{\pi}^\phi,\widehat{P}^\phi}(x,a) r_h^\phi(x,a) + \varepsilon_{\mathrm{RFE}} \\
&= \sum_{h=1}^H \sum_{(x,a)\in\Phi\times\mathcal{A}} d_h^{\pi^{\mathrm{E}},\phi}(x,a) r_h^\phi(x,a) - \sum_{h=1}^H \sum_{(x,a)\in\Phi\times\mathcal{A}} d_h^{\overline{\pi}^\phi,\widehat{P}^\phi}(x,a) r_h^\phi(x,a) + \varepsilon_{\mathrm{RFE}}.
\end{aligned}$$

Here $d_h^{\overline{\pi}^\phi, \widehat{P}^\phi}$ is the abstract state-action distribution of $\overline{\pi}^\phi$ in $\widehat{P}^\phi$ and $d_h^{\pi^{\mathrm{E}},\phi}(x,a) = \sum_{s\in\phi_h^{-1}(x)} d_h^{\pi^{\mathrm{E}}}(s,a)$. The last equation holds due to the reward-consistent condition in (10). Then we can obtain

$$\begin{aligned}
V^{\pi^{\mathrm{E}},\mathcal{M}} - V^{\overline{\pi}^\phi,\mathcal{M}^\phi} &\le \sum_{h=1}^{H} \sum_{(x,a)\in\Phi\times\mathcal{A}} d_h^{\pi^{\mathrm{E}},\phi}(x,a) r_h^\phi(x,a) - \sum_{h=1}^{H} \sum_{(x,a)\in\Phi\times\mathcal{A}} d_h^{\overline{\pi}^\phi,\widehat{P}^\phi}(x,a) r_h^\phi(x,a) + \varepsilon_{\mathrm{RFE}} \\
&\overset{(a)}{\le} \sum_{h=1}^{H} \left\| d_h^{\pi^{\mathrm{E}},\phi} - d_h^{\overline{\pi}^\phi,\widehat{P}^\phi} \right\|_1 + \varepsilon_{\mathrm{RFE}} \\
&\le \sum_{h=1}^{H} \left\| d_h^{\pi^{\mathrm{E}},\phi} - \widetilde{d}_h^{\pi^{\mathrm{E}},\phi} \right\|_1 + \sum_{h=1}^{H} \left\| \widetilde{d}_h^{\pi^{\mathrm{E}},\phi} - d_h^{\overline{\pi}^\phi,\widehat{P}^\phi} \right\|_1 + \varepsilon_{\mathrm{RFE}} \\
&\overset{(b)}{\le} \sum_{h=1}^{H} \left\| \widetilde{d}_h^{\pi^{\mathrm{E}},\phi} - d_h^{\overline{\pi}^\phi,\widehat{P}^\phi} \right\|_1 + \varepsilon_{\mathrm{EST}} + \varepsilon_{\mathrm{RFE}}.
\end{aligned}$$

Inequality (a) holds due to the dual representation of ℓ_1-norm and inequality (b) holds due to the event E_{EST}. Because $\overline{\pi}^\phi$ is an $\varepsilon_{\mathrm{OPT}}$-optimal solution of the optimization problem in (6), we get that

$$V^{\pi^{\mathrm{E}},\mathcal{M}} - V^{\overline{\pi}^\phi,\mathcal{M}^\phi} \le \min_{\pi^\phi\in\Pi^\phi} \sum_{h=1}^{H} \left\| \widetilde{d}_h^{\pi^{\mathrm{E}},\phi} - d_h^{\pi^\phi,\widehat{P}^\phi} \right\|_1 + \varepsilon_{\mathrm{EST}} + \varepsilon_{\mathrm{RFE}} + \varepsilon_{\mathrm{OPT}}.$$

We consider the abstract expert policy $\pi^{\mathrm{E},\phi}$ in Definition 2. Since $\pi^{\mathrm{E},\phi} \in \Pi^\phi$, it holds that

$$\begin{aligned}
V^{\pi^{\mathrm{E}},\mathcal{M}} - V^{\overline{\pi}^\phi,\mathcal{M}^\phi} &\le \sum_{h=1}^{H} \left\| \widetilde{d}_h^{\pi^{\mathrm{E}},\phi} - d_h^{\pi^{\mathrm{E}},\phi,\widehat{P}^\phi} \right\|_1 + \varepsilon_{\mathrm{EST}} + \varepsilon_{\mathrm{RFE}} + \varepsilon_{\mathrm{OPT}} \\
&\le \sum_{h=1}^{H} \left\| \widetilde{d}_h^{\pi^{\mathrm{E}},\phi} - d_h^{\pi^{\mathrm{E}},\phi,P^\phi} \right\|_1 + \sum_{h=1}^{H} \left\| d_h^{\pi^{\mathrm{E}},\phi,P^\phi} - d_h^{\pi^{\mathrm{E}},\phi,\widehat{P}^\phi} \right\|_1 + \varepsilon_{\mathrm{EST}} + \varepsilon_{\mathrm{RFE}} + \varepsilon_{\mathrm{OPT}}
\end{aligned}$$

Then we upper bound the term $\sum_{h=1}^{H} \| d_h^{\pi^{\mathrm{E}},\phi,P^\phi} - d_h^{\pi^{\mathrm{E}},\phi,\widehat{P}^\phi} \|_1$

$$\begin{aligned}
\sum_{h=1}^{H} \left\| d_h^{\pi^{\mathrm{E}},\phi,P^\phi} - d_h^{\pi^{\mathrm{E}},\phi,\widehat{P}^\phi} \right\|_1 &= \max_{r^\phi\in\mathcal{W}^\phi} \sum_{h=1}^{H} \sum_{(x,a)\in\Phi\times\mathcal{A}} \left(d_h^{\pi^{\mathrm{E}},\phi,P^\phi}(x,a) - d_h^{\pi^{\mathrm{E}},\phi,\widehat{P}^\phi}(x,a) \right) r_h^\phi(x,a) \\
&= \max_{r^\phi\in\mathcal{W}^\phi} V^{\pi^{\mathrm{E}},\phi,P^\phi,r^\phi} - V^{\pi^{\mathrm{E}},\phi,\widehat{P}^\phi,r^\phi} \\
&\le \varepsilon_{\mathrm{RFE}}.
\end{aligned}$$

Here $\mathcal{W}^\phi = \{w^\phi = (w_1^\phi, \ldots, w_H^\phi),\ w_h^\phi : \Phi\times\mathcal{A} \to [0,1], \forall h \in [H]\}$. The last inequality holds due to the event E_{RFE}. Then we obtain

$$V^{\pi^{\mathrm{E}},\mathcal{M}} - V^{\overline{\pi}^\phi,\mathcal{M}^\phi} \le \sum_{h=1}^{H} \left\| \widetilde{d}_h^{\pi^{\mathrm{E}},\phi} - d_h^{\pi^{\mathrm{E}},\phi,P^\phi} \right\|_1 + \varepsilon_{\mathrm{EST}} + 2\varepsilon_{\mathrm{RFE}} + \varepsilon_{\mathrm{OPT}}.$$

Applying Lemma 7 on $\pi^{\mathrm{E},\phi}$ and P^ϕ yields $d_h^{\pi^{\mathrm{E}},\phi,P^\phi} = d_h^{[\pi^{\mathrm{E}},\phi]^M,P,\phi}$. Combined with $[\pi^{\mathrm{E},\phi}]^M = \pi^{\mathrm{E}}$ in Lemma 4, we obtain

$$\begin{aligned}
V^{\pi^{\mathrm{E}},\mathcal{M}} - V^{\overline{\pi}^\phi,\mathcal{M}^\phi} &\le \sum_{h=1}^{H} \left\| \widetilde{d}_h^{\pi^{\mathrm{E}},\phi} - d_h^{\pi^{\mathrm{E}},\phi,P^\phi} \right\|_1 + \varepsilon_{\mathrm{EST}} + 2\varepsilon_{\mathrm{RFE}} + \varepsilon_{\mathrm{OPT}} \\
&= \sum_{h=1}^{H} \left\| \widetilde{d}_h^{\pi^{\mathrm{E}},\phi} - d_h^{\pi^{\mathrm{E}},\phi} \right\|_1 + \varepsilon_{\mathrm{EST}} + 2\varepsilon_{\mathrm{RFE}} + \varepsilon_{\mathrm{OPT}} \\
&\le 2\varepsilon_{\mathrm{EST}} + 2\varepsilon_{\mathrm{RFE}} + \varepsilon_{\mathrm{OPT}},
\end{aligned}$$

where the last inequality holds due to the event E_{EST}. We finish the proof. □

Now, we proceed to prove Theorem 2.

Proof of Theorem 2. First, we verify condition (a) in Proposition 1. We want to demonstrate that Algorithm 3 is equivalent to applying RF-Express (Algorithm 1) on the abstract MDP $\mathcal{M}^\phi$. The only difference lies in the data-collection process. On one hand, in line 11 in Algorithm 3, we roll out the lifted policy $[\pi^{\phi,t+1}]^M$ on the original MDP $\mathcal{M}$. On the other hand, when applying RF-Express (Algorithm 1) on the abstract MDP $\mathcal{M}^\phi$, we rollout the abstract policy $\pi^{\phi,t+1}$ on the abstract MDP $\mathcal{M}^\phi$. We will prove that in the above two data-collection processes, the corresponding abstract-state-action distributions are actually the same. Consequently, Algorithm 3 can be regarded as applying RF-Express (Algorithm 1) on the abstract MDP $\mathcal{M}^\phi$.

In the first process, conditioned on $\pi^{\phi,t+1}$, we consider the probability distribution of $(\phi_h(s_h^{t+1}), a_h^{t+1})$. Recall the definition:

$$d_h^{[\pi^{\phi,t+1}]^M,P,\phi}(x,a) := \mathbb{P}\left(\phi_h(s_h^{t+1}) = x, a_h^{t+1} = a | [\pi^{\phi,t+1}]^M, P\right) = \sum_{s\in\phi_h^{-1}(x)} \mathbb{P}\left(s_h^{t+1} = s, a_h^{t+1} = a | [\pi^{\phi,t+1}]^M, P\right).$$

By Lemma 7, we have that

$$d_h^{[\pi^{\phi,t+1}]^M,P,\phi}(x,a) = d_h^{\pi^{\phi,t+1},P^\phi}(x,a).$$

Notice that the distribution $d_h^{\pi^{\phi,t+1},P^\phi}(x,a)$ is exactly the abstract state-action distribution of $\pi^{\phi,t+1}$ in the abstract MDP $\mathcal{M}^\phi$. Therefore, in the mentioned two data-collection processes, the corresponding abstract-state-action distributions are actually the same. Then we can apply Lemma 2 on the abstract MDP. When the number of trajectories collected by Algorithm 3 satisfies

$$n \gtrsim \frac{H^3|\Phi||\mathcal{A}|}{\varepsilon^2}\left(|\Phi| + \log\left(\frac{|\Phi|H}{\delta}\right)\right),$$

for any policy $\pi^\phi \in \Pi^\phi$ and reward function $r^\phi = (r_1^\phi, \ldots, r_H^\phi),\ r_h^\phi : \Phi \times \mathcal{A} \to [0,1]$, with probability at least $1-\delta/2$, $|V^{\pi^\phi,P^\phi,r^\phi} - V^{\pi^\phi,\widehat{P}^\phi,r^\phi}| \le \varepsilon/16 = \varepsilon_{\mathrm{RFE}}$. In summary, the assumption (a) in Proposition 1 holds with $\delta_{\mathrm{RFE}} = \delta/2$ and $\varepsilon_{\mathrm{RFE}} = \varepsilon/16$.

Second, we verify the condition (b) in Proposition 1. Note that the assumption (b) in Proposition 1 holds by Lemma 8. More concretely, if the expert sample complexity and interaction complexity satisfies

$$m \gtrsim \frac{H^{3/2}|\Phi|}{\varepsilon}\log\left(\frac{|\Phi|H}{\delta}\right),\ n' \gtrsim \frac{H^2|\Phi|}{\varepsilon^2}\log\left(\frac{|\Phi|H}{\delta}\right),$$

with probability at least $1-\delta/2$, $\sum_{h=1}^H \|\widetilde{d}_h^{\pi^{\mathrm{E}},\phi} - d_h^{\pi^{\mathrm{E}},\phi}\|_1 \le \varepsilon/16 = \varepsilon_{\mathrm{EST}}$. Hence, the assumption (b) in Proposition 1 holds with $\delta_{\mathrm{EST}} = \delta/2$ and $\varepsilon_{\mathrm{EST}} = \varepsilon/16$.

Third, we validate the condition (c) in Proposition 1. In particular, we apply Algorithm 2 to solve the following abstract state-action distribution matching problem.

$$\min_{\pi^\phi\in\Pi^\phi} \sum_{h=1}^H \left\|\widetilde{d}_h^{\pi^{\mathrm{E}},\phi} - d_h^{\pi^\phi,\widehat{P}^\phi}\right\|_1.$$

Therefore, we can apply Lemma 5. In particular, when $\varepsilon_{\mathrm{RL}} \le \varepsilon/2$ and $T \gtrsim |\Phi||\mathcal{A}|H^2/\varepsilon^2$ such that $2H\sqrt{2|\Phi||\mathcal{A}|/T} \le \varepsilon/4$, we have that

$$\sum_{h=1}^H \left\|\widetilde{d}_h^{\pi^{\mathrm{E}},\phi} - d_h^{\overline{\pi}^\phi,\widehat{P}^\phi}\right\|_1 - \min_{\pi^\phi\in\Pi^\phi}\sum_{h=1}^H \left\|\widetilde{d}_h^{\pi^{\mathrm{E}},\phi} - d_h^{\pi^\phi,\widehat{P}^\phi}\right\|_1 \le \frac{3\varepsilon}{4} = \varepsilon_{\mathrm{OPT}}.$$

In summary, we have established the following conditions:

- Assumption (a) in Proposition 1 holds with $\delta_{\mathrm{RFE}} = \delta/2$ and $\varepsilon_{\mathrm{RFE}} = \varepsilon/16$.
- Assumption (b) in Proposition 1 holds with $\delta_{\mathrm{EST}} = \delta/2$ and $\varepsilon_{\mathrm{EST}} = \varepsilon/16$.
- Assumption (c) in Proposition 1 holds with $\varepsilon_{\mathrm{OPT}} = 3\varepsilon/4$.

By applying Proposition 1, we complete the proof. With probability at least $1-\delta$, we have

$$V^{\pi^{\mathrm{E}}} - V^{[\overline{\pi}^{\phi}]^M} \leq 2\varepsilon_{\mathrm{RFE}} + 2\varepsilon_{\mathrm{EST}} + \varepsilon_{\mathrm{OPT}} = \varepsilon.$$

□

5.4 USEFUL LEMMAS

In this part, we develop specialized analysis tools for AIL with state abstraction. The below lemma indicates that under Assumption 1, the lifted versions of the abstract reward function and abstract transition function are identical to the original reward function and transition function, respectively.

Lemma 4. *For the original MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, H, \rho)$ and expert policy π^{E} that satisfy Assumption 1, we consider the abstract MDP $\mathcal{M}^{\phi} = (\Phi, \mathcal{A}, P^{\phi}, r^{\phi}, H, \rho^{\phi})$ in Definition 1. Then we have that*

$$\forall h \in [H],\ (s,a) \in \mathcal{S} \times \mathcal{A},\ x' \in \Phi,\ r_h(s,a) = [r^{\phi}]_h^M(s,a),\ \sum_{s' \in \phi_{h+1}^{-1}(x')} P_h(s'|s,a) = [P^{\phi}]_h^M(x'|s,a).$$

Here $[r^{\phi}]_h^M(s,a) = r_h^{\phi}(\phi_h(s),a)$ and $[P^{\phi}]_h^M(x'|s,a) = P_h^{\phi}(x'|\phi_h(s),a)$. Furthermore, we consider the abstract expert policy $\pi^{\mathrm{E},\phi}$ in Definition 2. Then we have that

$$\forall h \in [H],\ s \in \mathcal{S},\ \pi_h^{\mathrm{E}}(s) = [\pi^{\mathrm{E},\phi}]_h^M(s),$$

where $[\pi^{\mathrm{E},\phi}]_h^M(s) = \pi_h^{\mathrm{E},\phi}(\phi_h(s))$.

Proof. For the reward function, we have

$$[r_h^{\phi}]_{\mathcal{M}}(s,a) = r_h^{\phi}(\phi_h(s),a) \overset{x:=\phi_h(s)}{=} r_h^{\phi}(x,a).$$

Notice that $r_h^{\phi}(x,a) = r_h(\widehat{s},a)$ for an arbitrary $\widehat{s} \in \phi_h^{-1}(x)$. Moreover, since $s,\ \widehat{s} \in \phi_h^{-1}(x)$ and r satisfies (10), we have $r_h(\widehat{s},a) = r_h(s,a)$.

For the transition function, we have

$$[P^{\phi}]_h^M(x'|s,a) = P_h^{\phi}(x'|\phi_h(s),a) \overset{x:=\phi_h(s)}{=} P_h^{\phi}(x'|x,a).$$

According to Definition 1, we have $P_h^{\phi}(x'|x,a) = \sum_{s' \in \phi_{h+1}^{-1}(x')} P_h(s'|\widetilde{s},a)$ for an arbitrary $\widetilde{s} \in \phi_h^{-1}(x)$. Furthermore, because $s,\ \widetilde{s} \in \phi_h^{-1}(x)$ and P satisfies (11), we have

$$\sum_{s' \in \phi_{h+1}^{-1}(x')} P_h(s'|\widetilde{s},a) = \sum_{s' \in \phi_{h+1}^{-1}(x')} P_h(s'|s,a).$$

Finally, for the expert policy, it holds that

$$[\pi^{\mathrm{E},\phi}]_h^M(s) = \pi_h^{\mathrm{E},\phi}(\phi_h(s)) \overset{x:=\phi_h(s)}{=} \pi_h^{\mathrm{E},\phi}(x).$$

According to Definition 2, we have $\pi_h^{\mathrm{E},\phi}(x) = \pi_h^{\mathrm{E}}(\widetilde{s})$ for an arbitrary $\widetilde{s} \in \phi_h^{-1}(x)$. Notice that $s,\ \widetilde{s} \in \phi_h^{-1}(x)$ and π^{E} satisfies (12). Therefore, we have $\pi_h^{\mathrm{E},\phi}(x) = \pi_h^{\mathrm{E}}(s)$. We finish the proof. □

Lemma 5. *For any function $f : \Phi \to \mathbb{R}$, $g : \mathcal{S} \to \mathbb{R}$ and an state abstraction $\phi : \mathcal{S} \to \Phi$, we define $g^{\phi}(x) := \sum_{s \in \phi^{-1}(x)} g(s)$, then we have*

$$\sum_{x \in \Phi} g^{\phi}(x) f(x) = \sum_{s \in \mathcal{S}} g(s) [f]^M(s),$$

where $[f]^M(s) = f(\phi(s))$.

Proof.

$$\begin{aligned}\sum_{x\in\Phi} g^{\phi}(x)f(x) &= \sum_{x\in\Phi}\sum_{s\in\phi^{-1}(x)} g(s)f(x)\\ &= \sum_{x\in\Phi}\sum_{s\in\mathcal{S}} \mathbb{I}\left\{s\in\phi^{-1}(x)\right\} g(s)f(x)\\ &= \sum_{s\in\mathcal{S}}\sum_{x\in\Phi} \mathbb{I}\left\{x=\phi(s)\right\} g(s)f(x)\\ &= \sum_{s\in\mathcal{S}} g(s)f(\phi(s))\\ &= \sum_{s\in\mathcal{S}} g(s)[f]^M(s).\end{aligned}$$

We complete the proof. □

Lemma 6 indicates that for any abstract policy $\pi^{\phi}\in\Pi^{\phi}$, the value function of $[\pi^{\phi}]^M$ on P equals the lifted version of the value function of π^{ϕ} on P^{ϕ}.

Lemma 6. *For the original MDP $\mathcal{M}=(\mathcal{S},\mathcal{A},P,r,H,\rho)$ and expert policy π^{E} that satisfy Assumption 1, we consider the abstract MDP $\mathcal{M}^{\phi}=(\Phi,\mathcal{A},P^{\phi},r^{\phi},H,\rho^{\phi})$ in Definition 1. Then, for any abstract policy $\pi^{\phi}\in\Pi^{\phi}$, we have*

$$V_h^{[\pi^{\phi}]^M,\mathcal{M}}(s) = [V^{\pi^{\phi},\mathcal{M}^{\phi}}]_h^M(s), \forall s\in\mathcal{S}, h\in[H],$$

where $[V^{\pi^{\phi},\mathcal{M}^{\phi}}]_h^M(s) := V_h^{\pi^{\phi},\mathcal{M}^{\phi}}(\phi_h(s))$, $[\pi^{\phi}]_h^M(a|s)=\pi_h^{\phi}(a|\phi_h(s))$. $V_h^{\pi^{\phi},\mathcal{M}^{\phi}}(s)$ is the value function of π^{ϕ} on $\mathcal{M}^{\phi}$ and $V_h^{[\pi^{\phi}]^M,\mathcal{M}}(s)$ is the value function of $[\pi^{\phi}]^M$ on $\mathcal{M}$. Furthermore, it holds that $V^{[\pi^{\phi}]^M,\mathcal{M}}=V^{\pi^{\phi},\mathcal{M}^{\phi}}$.

Proof. The proof is based on backward induction. For the base case, we prove that

$$V_H^{[\pi^{\phi}]^M,\mathcal{M}}(s) = [V^{\pi^{\phi},\mathcal{M}^{\phi}}]_H^M(s),\ \forall s\in\mathcal{S}.$$

In particular,

$$\begin{aligned}[V^{\pi^{\phi},\mathcal{M}^{\phi}}]_H^M(s) &= V_H^{\pi^{\phi},\mathcal{M}^{\phi}}(\phi_H(s))\\ &= \sum_{a\in\mathcal{A}}\pi_H^{\phi}(a|\phi_H(s))r_H^{\phi}(\phi_H(s),a)\\ &= \sum_{a\in\mathcal{A}}[\pi^{\phi}]_H^M(a|s)[r^{\phi}]_H^M(s,a)\\ &\overset{(a)}{=} \sum_{a\in\mathcal{A}}[\pi^{\phi}]_H^M(a|s)r_H(s,a)\\ &= V_H^{[\pi^{\phi}]^M,\mathcal{M}}(s).\end{aligned}$$

Equation (a) follows Lemma 4. We finish the proof of the base case and continue to prove the induction stage. Assume that $V_{h+1}^{[\pi^{\phi}]^M,\mathcal{M}}(s)=[V^{\pi^{\phi},\mathcal{M}^{\phi}}]_{h+1}^M(s),\forall s\in\mathcal{S}$, we consider the time step h.

$$\begin{aligned}[V^{\pi^{\phi},\mathcal{M}^{\phi}}]_h^M(s) &= V_h^{\pi^{\phi},\mathcal{M}^{\phi}}(\phi_h(s))\\ &= \mathbb{E}_{a\sim\pi_h^{\phi}(\cdot|\phi_h(s))}\left[r_h^{\phi}(\phi_h(s),a)+P_{h+1}^{\phi}V_{h+1}^{\pi^{\phi},\mathcal{M}^{\phi}}(\phi_h(s),a)\right].\end{aligned}$$

Here $P_{h+1}^{\phi}V_{h+1}^{\pi^{\phi},\mathcal{M}^{\phi}}(\phi_h(s),a)=\mathbb{E}_{x'\sim P_{h+1}^{\phi}(\cdot|\phi_h(s),a)}\left[V_{h+1}^{\pi^{\phi},\mathcal{M}^{\phi}}(x')\right]$. For the first term in RHS, we have

$$\mathbb{E}_{a\sim\pi_h^{\phi}(\cdot|\phi_h(s))}\left[r_h^{\phi}(\phi_h(s),a)\right] = \mathbb{E}_{a\sim[\pi^{\phi}]_h^M(\cdot|s)}\left[[r^{\phi}]_h^M(s,a)\right] = \mathbb{E}_{a\sim[\pi^{\phi}]_h^M(\cdot|s)}\left[r_h(s,a)\right].$$

The last equation utilizes Lemma 4. For the term $P_{h+1}^{\phi} V_{h+1}^{\pi^{\phi},\mathcal{M}^{\phi}}(\phi_h(s),a)$, we obtain

$$\begin{aligned}
P_{h+1}^{\phi} V_{h+1}^{\pi^{\phi},\mathcal{M}^{\phi}}(\phi_h(s),a) &= \sum_{x'\in\Phi} P_{h+1}^{\phi}(x'|\phi_h(s),a) V_{h+1}^{\pi^{\phi},\mathcal{M}^{\phi}}(x') \\
&= \sum_{x'\in\Phi} \left(\sum_{s'\in\phi_h^{-1}(x')} P_{h+1}(s'|\phi_h(s),a) \right) V_{h+1}^{\pi^{\phi},\mathcal{M}^{\phi}}(x'), \\
&= \sum_{x'\in\Phi} \left(\sum_{s'\in\phi_h^{-1}(x')} P_{h+1}(s'|\tilde{s},a) \right) V_{h+1}^{\pi^{\phi},\mathcal{M}^{\phi}}(x'), \text{ for an arbitrary } \widetilde{s}\in\phi_h^{-1}(x).
\end{aligned}$$

In the last equation, we define $x=\phi_h(s)$. According to $s,\ \widetilde{s}\in\phi_h^{-1}(x)$ and (11) in Assumption 1, we have

$$\begin{aligned}
P_{h+1}^{\phi} V_{h+1}^{\pi^{\phi},\mathcal{M}^{\phi}}(\phi_h(s),a) &= \sum_{x'\in\Phi} \left(\sum_{s'\in\phi_h^{-1}(x')} P_{h+1}(s'|\tilde{s},a) \right) V_{h+1}^{\pi^{\phi},\mathcal{M}^{\phi}}(x') \\
&= \sum_{x'\in\Phi} \left(\sum_{s'\in\phi_h^{-1}(x')} P_{h+1}(s'|s,a) \right) V_{h+1}^{\pi^{\phi},\mathcal{M}^{\phi}}(x').
\end{aligned}$$

Applying Lemma 5 with $f(x)=V_{h+1}^{\pi^{\phi},\mathcal{M}^{\phi}}(x), g(s')=P_{h+1}(s'|s,a), \phi=\phi_{h+1}$ yields that

$$\begin{aligned}
P_{h+1}^{\phi} V_{h+1}^{\pi^{\phi},\mathcal{M}^{\phi}}(\phi_h(s),a) &= \sum_{s'\in\mathcal{S}} P_{h+1}(s'|s,a) \left[V^{\pi^{\phi},\mathcal{M}^{\phi}} \right]_{h+1}^{M}(s') \\
&\overset{(a)}{=} \sum_{s'\in\mathcal{S}} P_{h+1}(s'|s,a) V_{h+1}^{[\pi^{\phi}]^M,\mathcal{M}}(s') \\
&= P_{h+1} V_{h+1}^{[\pi^{\phi}]^M,\mathcal{M}}(s,a).
\end{aligned}$$

In equation (a), we leverage the assumption in time step $h+1$. Then we obtain

$$\begin{aligned}
[V^{\pi^{\phi},\mathcal{M}^{\phi}}]_h^M(s) &= \mathbb{E}_{a\sim\pi_h^{\phi}(\cdot|\phi_h(s))} \left[r_h^{\phi}(\phi_h(s),a) + P_{h+1}^{\phi} V_{h+1}^{\pi^{\phi},\mathcal{M}^{\phi}}(\phi_h(s),a) \right] \\
&= \mathbb{E}_{a\sim[\pi^{\phi}]_h^M(\cdot|s)} \left[r_h(s,a) + P_{h+1} V_{h+1}^{[\pi^{\phi}]^M,\mathcal{M}}(s,a) \right] \\
&= V_h^{[\pi^{\phi}]^M,\mathcal{M}}(s).
\end{aligned}$$

We prove the induction stage and thus finish the proof of the first claim. Furthermore, according to the definition of ρ^{ϕ}, we have

$$V^{\pi^{\phi},\mathcal{M}^{\phi}} = \mathbb{E}_{x\sim\rho^{\phi}} \left[V_1^{\pi^{\phi},\mathcal{M}^{\phi}}(x) \right] = \sum_{x\in\Phi} \rho^{\phi}(x) V_1^{\pi^{\phi},\mathcal{M}^{\phi}}(x) = \sum_{s\in\mathcal{S}} \rho(s) \left[V^{\pi^{\phi},\mathcal{M}^{\phi}} \right]_1^M(s).$$

In the last equation, we apply Lemma 5 with $f(x)=V_1^{\pi^{\phi},\mathcal{M}^{\phi}}(x)$, $g(s)=\rho(s)$ and $\phi=\phi_1$. We have proved that $[V^{\pi^{\phi},\mathcal{M}^{\phi}}]_1^M(s) = V_1^{[\pi^{\phi}]^M,\mathcal{M}}(s)$. Then it holds that

$$V^{\pi^{\phi},\mathcal{M}^{\phi}} = \sum_{s\in\mathcal{S}} \rho(s) V_1^{[\pi^{\phi}]^M,\mathcal{M}}(s) = V_1^{[\pi^{\phi}]^M,\mathcal{M}},$$

which completes the proof. □

Lemma 7. *For the original MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, H, \rho)$ and expert policy π^{E} that satisfy Assumption 1, we consider the abstract MDP $\mathcal{M}^{\phi} = (\Phi, \mathcal{A}, P^{\phi}, r^{\phi}, H, \rho^{\phi})$ in Definition 1. Then, for any abstract policy $\pi^{\phi} \in \Pi^{\phi}$,*

$$\forall h \in [H],\ (x, a) \in \Phi \times \mathcal{A},\ d_h^{\pi^{\phi}, P^{\phi}}(x, a) = d_h^{[\pi^{\phi}]^M, P, \phi}(x, a).$$

Here $d_h^{\pi^{\phi}, P^{\phi}}(x, a) = \mathbb{P}(x_h = x, a_h = a | \pi^{\phi}, P^{\phi})$ and $d_h^{[\pi^{\phi}]^M, P, \phi}(x, a) = \mathbb{P}(\phi_h(s_h) = x, a_h = a | [\pi^{\phi}]^M, P) = \sum_{s \in \phi_h^{-1}(x)} d_h^{[\pi^{\phi}]^M, P}(s, a)$.

Proof. We first prove that for any fixed $x \in \Phi, h \in [H]$,

$$d_h^{\pi^{\phi}, P^{\phi}}(x) = d_h^{[\pi^{\phi}]^M, P, \phi}(x),$$

where $d_h^{\pi^{\phi}, P^{\phi}}(x) = \mathbb{P}\left(x_h = x | \pi^{\phi}, P^{\phi}\right)$ and $d_h^{[\pi^{\phi}]^M, P, \phi}(x) = \mathbb{P}\left(\phi_h(s_h) = x | [\pi^{\phi}]^M, P\right)$. Consider any fixed $x \in \Phi, h \in [H]$, we construct an abstract reward function $\widetilde{r}^{\phi}$.

$$\begin{aligned}
&\widetilde{r}_h^{\phi}(x, a) = 1, \forall a \in \mathcal{A}, \\
&\widetilde{r}_{\ell}^{\phi}(\tilde{x}, a) = 0, \forall \tilde{x} \in \Phi \setminus \{x\}, a \in \mathcal{A}, \ell \in [H] \setminus \{h\}.
\end{aligned}$$

Furthermore, we consider $[\widetilde{r}^{\phi}]^M$, which is the lifted version of $\widetilde{r}^{\phi}$.

$$\begin{aligned}
&\left[\widetilde{r}^{\phi}\right]_h^M (s, a) = 1, \forall s \in \phi_h^{-1}(x), a \in \mathcal{A}, \\
&\left[\widetilde{r}^{\phi}\right]_{\ell}^M (s, a) = 0, \forall s \in \mathcal{S} \setminus \phi_h^{-1}(x), a \in \mathcal{A}, \ell \in [H] \setminus \{h\}.
\end{aligned}$$

On the one hand, according to the dual formulation of policy value in (1), we can get that $d_h^{\pi^{\phi}, P^{\phi}}(x) = V^{\pi^{\phi}, P^{\phi}, \widetilde{r}^{\phi}}$. On the other hand, it holds that

$$d_h^{[\pi^{\phi}]^M, P, \phi}(x) = \sum_{s \in \phi_h^{-1}(x)} d_h^{[\pi^{\phi}]^M, P}(s) = V^{[\pi^{\phi}]^M, P, [\widetilde{r}^{\phi}]^M}.$$

The last equation still follows the dual representation of policy value. Notice that $[\widetilde{r}^{\phi}]^M$ satisfies the reward-consistent condition (i.e., (10) in Assumption 1). With Lemma 6, we get that $V^{[\pi^{\phi}]^M, P, [\widetilde{r}^{\phi}]^M} = V^{\pi^{\phi}, P^{\phi}, \widetilde{r}^{\phi}}$, which implies that $d_h^{\pi^{\phi}, P^{\phi}}(x) = d_h^{[\pi^{\phi}]^M, P, \phi}(x)$. Then we have that

$$d_h^{\pi^{\phi}, P^{\phi}}(x, a) = d_h^{\pi^{\phi}, P^{\phi}}(x) \pi_h^{\phi}(a|x) = d_h^{[\pi^{\phi}]^M, P, \phi}(x) \pi_h^{\phi}(a|x) = d_h^{[\pi^{\phi}]^M, P, \phi}(x) \left[\pi^{\phi}\right]_h^M (a|s) = d_h^{[\pi^{\phi}]^M, P, \phi}(x, a),$$

where $s \in \phi_h^{-1}(x)$. We finish the proof. □

Lemma 8. *Given the expert dataset $\mathcal{D}$, let $\mathcal{D}$ be divided into two equal subsets, i.e., $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_1^c$ and $\mathcal{D}_1 \cap \mathcal{D}_1^c = \emptyset$ with $|\mathcal{D}_1| = |\mathcal{D}_1^c| = m/2$. Let $\pi'^{,\phi}$ be the abstract BC's policy on $\mathcal{D}_1$. Fix $\pi'^{,\phi}$, let $\mathcal{D}'_{\mathrm{env}}$ be the dataset collected by $[\pi'^{,\phi}]^M$ and $|\mathcal{D}'_{\mathrm{env}}| = n'$. Fix $\varepsilon \in (0, 1)$ and $\delta \in (0, 1)$; suppose $H \geq 5$. Consider the abstract state-action distribution estimator $\widetilde{d}_h^{\pi^{\mathrm{E}}, \phi}$ shown in (5), if the expert sample complexity (m) and the interaction complexity (n') satisfy*

$$m \gtrsim \frac{H^{3/2}|\Phi|}{\varepsilon} \log\left(\frac{|\Phi|H}{\delta}\right),\ n' \gtrsim \frac{H^2|\Phi|}{\varepsilon^2} \log\left(\frac{|\Phi|H}{\delta}\right),$$

then with probability at least $1 - \delta$, we have

$$\sum_{h=1}^{H} \left\| \widetilde{d}_h^{\pi^{\mathrm{E}}, \phi} - d_h^{\pi^{\mathrm{E}}, \phi} \right\|_1 \leq \varepsilon.$$

Proof. First, we can obtain that

$$\sum_{h=1}^{H}\left\|\widetilde{d}_h^{\pi^{\mathrm{E}},\phi}-d_h^{\pi^{\mathrm{E}},\phi}\right\|_1=\sum_{h=1}^{H}\sum_{(x,a)\in\Phi\times\mathcal{A}}\left|\widetilde{d}_h^{\pi^{\mathrm{E}},\phi}(x,a)-d_h^{\pi^{\mathrm{E}},\phi}(x,a)\right|$$
$$=\sum_{h=1}^{H}\sum_{x\in\Phi}\left|\widetilde{d}_h^{\pi^{\mathrm{E}},\phi}(x,\pi_h^{\mathrm{E},\phi}(x))-d_h^{\pi^{\mathrm{E}},\phi}(x,\pi_h^{\mathrm{E},\phi}(x))\right|.$$

Here $\pi^{\mathrm{E},\phi}$ is the abstract expert policy in Definition 2. The last equation holds since π^{E} is a deterministic policy and satisfies (12) in Assumption 1. Recall the abstract state-action distribution estimator $\widetilde{d}_h^{\pi^{\mathrm{E}},\phi}$ shown in (5).

$$\widetilde{d}_h^{\pi^{\mathrm{E}},\phi}(x,\pi_h^{\mathrm{E},\phi}(x))=\frac{\sum_{\mathbf{tr}_h\in\mathcal{D}'_{\mathrm{env}}}\mathbb{I}\{\phi_h(\mathbf{tr}_h(\cdot))=x,\mathbf{tr}_h(a_h)=\pi_h^{\mathrm{E},\phi}(x),\mathbf{tr}_h\in\mathrm{Tr}_h^{\mathcal{D}_1,\phi}\}}{|\mathcal{D}'_{\mathrm{env}}|}$$
$$+\frac{\sum_{\mathbf{tr}_h\in\mathcal{D}_1^c}\mathbb{I}\{\phi_h(\mathbf{tr}_h(\cdot))=x,\mathbf{tr}_h(a_h)=\pi_h^{\mathrm{E},\phi}(x),\mathbf{tr}_h\notin\mathrm{Tr}_h^{\mathcal{D}_1,\phi}\}}{|\mathcal{D}_1^c|}.$$

Given $\mathcal{D}_1$, for $d_h^{\pi^{\mathrm{E}},\phi}$, we have the following decomposition.

$$\begin{aligned}
&d_h^{\pi^{\mathrm{E}},\phi}(x,\pi_h^{\mathrm{E},\phi}(x))\\
&=\sum_{\mathbf{tr}_h}\mathbb{P}^{\pi^{\mathrm{E}}}(\mathbf{tr}_h)\mathbb{I}\{\phi_h(\mathbf{tr}_h(\cdot))=x,\mathbf{tr}_h(a_h)=\pi_h^{\mathrm{E},\phi}(x)\}\\
&=\sum_{\mathbf{tr}_h\in\mathrm{Tr}_h^{\mathcal{D}_1,\phi}}\mathbb{P}^{\pi^{\mathrm{E}}}(\mathbf{tr}_h)\mathbb{I}\{\phi_h(\mathbf{tr}_h(\cdot))=x,\mathbf{tr}_h(a_h)=\pi_h^{\mathrm{E},\phi}(x)\}\\
&+\sum_{\mathbf{tr}_h\notin\mathrm{Tr}_h^{\mathcal{D}_1,\phi}}\mathbb{P}^{\pi^{\mathrm{E}}}(\mathbf{tr}_h)\mathbb{I}\{\phi_h(\mathbf{tr}_h(\cdot))=x,\mathbf{tr}_h(a_h)=\pi_h^{\mathrm{E},\phi}(x)\}.
\end{aligned}$$

Then we have that

$$\begin{aligned}
&|\widetilde{d}_h^{\pi^{\mathrm{E}},\phi}(x,\pi_h^{\mathrm{E},\phi}(x))-d_h^{\pi^{\mathrm{E}},\phi}(x,\pi_h^{\mathrm{E},\phi}(x))|\\
&\le\left|\frac{\sum_{\mathbf{tr}_h\in\mathcal{D}'_{\mathrm{env}}}\mathbb{I}\{\phi_h(\mathbf{tr}_h(\cdot))=x,\mathbf{tr}_h(a_h)=\pi_h^{\mathrm{E},\phi}(x),\mathbf{tr}_h\in\mathrm{Tr}_h^{\mathcal{D}_1,\phi}\}}{|\mathcal{D}'_{\mathrm{env}}|}\right.\\
&\left.-\sum_{\mathbf{tr}_h\in\mathrm{Tr}_h^{\mathcal{D}_1,\phi}}\mathbb{P}^{\pi^{\mathrm{E}}}(\mathbf{tr}_h)\mathbb{I}\{\phi_h(\mathbf{tr}_h(\cdot))=x,\mathbf{tr}_h(a_h)=\pi_h^{\mathrm{E},\phi}(x)\}\right|\\
&+\left|\frac{\sum_{\mathbf{tr}_h\in\mathcal{D}_1^c}\mathbb{I}\{\phi_h(\mathbf{tr}_h(\cdot))=x,\mathbf{tr}_h(a_h)=\pi_h^{\mathrm{E},\phi}(x),\mathbf{tr}_h\notin\mathrm{Tr}_h^{\mathcal{D}_1,\phi}\}}{|\mathcal{D}_1^c|}\right.\\
&\left.-\sum_{\mathbf{tr}_h\notin\mathrm{Tr}_h^{\mathcal{D}_1,\phi}}\mathbb{P}^{\pi^{\mathrm{E}}}(\mathbf{tr}_h)\mathbb{I}\{\phi_h(\mathbf{tr}_h(\cdot))=x,\mathbf{tr}_h(a_h)=\pi_h^{\mathrm{E},\phi}(x)\}\right|.
\end{aligned}$$

We denote the first term in RHS as $\mathrm{EA}_h(x)$ and the second term in RHS as $\mathrm{EB}_h(x)$. We have that

$$\sum_{h=1}^{H}\left\|d_h^{\pi^{\mathrm{E}},\phi}-\widetilde{d}_h^{\pi^{\mathrm{E}},\phi}\right\|_1\le\underbrace{\sum_{h=1}^{H}\sum_{x\in\Phi}\mathrm{EA}_h(x)}_{\text{Error A}}+\underbrace{\sum_{h=1}^{H}\sum_{x\in\Phi}\mathrm{EB}_h(x)}_{\text{Error B}}.$$

First, we analyze the term Error A. Let ${E'}_h^x$ be the event that $\mathbf{tr}_h$ agrees with expert policy at abstract state x in time step h and appears in $\mathrm{Tr}_h^{\mathcal{D}_1,\phi}$. Formally,

$${E'}_h^x=\mathbb{I}\{\phi_h(\mathbf{tr}_h(\cdot))=x\cap\mathbf{tr}_h(a_h)=\pi_h^{\mathrm{E},\phi}(x)\cap\mathbf{tr}_h\in\mathrm{Tr}_h^{\mathcal{D}_1,\phi}\}.$$

Then we leverage Chernoff's bound to upper bound $\mathrm{EA}_h(x)$. By Lemma 2, for each $x \in \mathcal{S}$ and $h \in [H]$, with probability at least $1 - \frac{\delta}{2|\Phi|H}$ over the randomness of $\mathcal{D}'$, we have

$$\mathrm{EA}_h(x) \leq \sqrt{\mathbb{P}^{\pi^{\mathrm{E}}}\left(E'^x_h\right) \frac{3\log\left(4|\Phi|H/\delta\right)}{n'}}.$$

By union bound, with probability at least $1 - \frac{\delta}{2}$ over the randomness of $\mathcal{D}'_{\mathrm{env}}$, we have

$$\begin{aligned}\sum_{h=1}^{H}\sum_{x\in\Phi}\mathrm{EA}_h(x) &\leq \sum_{h=1}^{H}\sum_{x\in\Phi}\sqrt{\mathbb{P}^{\pi^{\mathrm{E}}}\left(E'^x_h\right) \frac{3\log\left(4|\Phi|H/\delta\right)}{n'}} \\ &\leq \sum_{h=1}^{H}\sqrt{|\Phi|}\sqrt{\sum_{x\in\Phi}\mathbb{P}^{\pi^{\mathrm{E}}}\left(E'^x_h\right) \frac{3\log\left(4|\Phi|H/\delta\right)}{n'}}\end{aligned}$$

The last inequality follows the Cauchy-Schwartz inequality. It remains to upper bound $\sum_{x\in\Phi}\mathbb{P}^{\pi^{\mathrm{E}}}(E^x_h)$ for all $h \in [H]$. To this end, we define the event $G'^{\mathcal{D}_1}_h$ that expert policy π^{E} visits abstract states covered in $\mathcal{D}_1$ up to time step h. Formally, $G'^{\mathcal{D}_1}_h = \mathbb{I}\{\forall h' \leq h, \phi_{h'}(s_{h'}) \in \Phi_{h'}(\mathcal{D}_1)\}$, where $\Phi_h(\mathcal{D}_1)$ is the set of abstract states in $\mathcal{D}_1$ at time step h. Then, for all $h \in [H]$, we have

$$\sum_{x\in\Phi}\mathbb{P}^{\pi^{\mathrm{E}}}\left(E'^x_h\right) = \mathbb{P}^{\pi^{\mathrm{E}}}(G'^{\mathcal{D}_1}_h) \leq \mathbb{P}(G'^{\mathcal{D}_1}_1).$$

The last inequality holds since $G'^{\mathcal{D}_1}_h \subseteq G'^{\mathcal{D}_1}_1$ for all $h \in [H]$. Then we have that

$$\sum_{h=1}^{H}\sum_{x\in\Phi}\mathrm{EA}_h(x) \leq H\sqrt{\frac{3|\Phi|\log\left(4|\Phi|H/\delta\right)}{n'}}.$$

When the interaction complexity satisfies that $n' \gtrsim \frac{|\Phi|H^2}{\varepsilon^2}\log\left(\frac{|\Phi|H}{\delta}\right)$, with probability at least $1 - \frac{\delta}{2}$ over the randomness of $\mathcal{D}'$, we have $\sum_{h=1}^{H}\sum_{x\in\Phi}\mathrm{EA}_h(x) \leq \frac{\varepsilon}{2}$.

Second, we upper bound the term Error B. Similarly, we can leverage Chernoff's bound to characterize its concentration rate. For a trajectory $\mathbf{tr}_h$, let E^x_h be the event that $\mathbf{tr}_h$ agrees with expert policy at abstract state x at time step h but is not in $\mathrm{Tr}^{\mathcal{D}_1,\phi}_h$, that is,

$$E^x_h = \{\phi_h(\mathbf{tr}_h(\cdot)) = x \cap \mathbf{tr}_h(a_h) = \pi^{\mathrm{E},\phi}_h(x) \cap \mathbf{tr}_h \notin \mathrm{Tr}^{\mathcal{D}_1,\phi}_h\}.$$

We consider E^x_h is measured by the stochastic process induced by the expert policy π^{E}. Accordingly, its probability is denoted as $\mathbb{P}^{\pi^{\mathrm{E}}}(E^x_h)$. We see that $\mathbb{P}^{\pi^{\mathrm{E}}}(E^x_h)$ is equal to the second term in $\mathrm{EB}_h(x)$. Moreover, the first term in $\mathrm{EB}_h(x)$ is an empirical estimation for $\mathbb{P}^{\pi^{\mathrm{E}}}(E^x_h)$. After applying Chernoff's bound, with probability at least $1 - \delta/(2|\Phi|H)$ with $\delta \in (0,1)$ (over the randomness of the expert demonstrations $\mathcal{D}^c_1$), for each $h \in [H], x \in \Phi$, we have

$$\mathrm{EB}_h(x) \leq \sqrt{\mathbb{P}^{\pi^{\mathrm{E}}}\left(E^x_h\right) \frac{3\log(4|\Phi|H/\delta)}{m}}.$$

Therefore, with probability at least $1 - \delta/2$, we have

$$\begin{aligned}\sum_{h=1}^{H}\sum_{x\in\Phi}\mathrm{EB}_h(x) &\leq \sum_{h=1}^{H}\sum_{x\in\Phi}\sqrt{\mathbb{P}^{\pi^{\mathrm{E}}}\left(E^x_h\right) \frac{3\log(4|\Phi|H/\delta)}{m}} \\ &\leq \sum_{h=1}^{H}\sqrt{\sum_{x\in\Phi}\mathbb{P}^{\pi^{\mathrm{E}}}\left(E^x_h\right) \frac{3|\Phi|\log(4|\Phi|H/\delta)}{m}},\end{aligned}$$

where the last step follows the Cauchy–Schwarz inequality. It remains to upper bound $\sum_{x\in\Phi}\mathbb{P}^{\pi^{\mathrm{E}}}\left(E^x_h\right)$ for all $h \in [H]$. To this end, we define the event $G^{\mathcal{D}_1}_h$: the expert policy visits certain abstract states uncovered in $\mathcal{D}_1$ up to time step h. Formally,

$G_h^{\mathcal{D}_1} = \{\exists h' \le h, \phi_{h'}(s_{h'}) \notin \Phi_{h'}(\mathcal{D}_1)\}$, where $\Phi_{h'}(\mathcal{D}_1)$ is the set of abstract states in $\mathcal{D}_1$ at time step h. Then, for all $h \in [H]$, we have

$$\sum_{x\in\Phi} \mathbb{P}^{\pi^{\mathrm{E}}}(E_h^x) = \mathbb{P}^{\pi^{\mathrm{E}}}\left(G_h^{\mathcal{D}_1}\right) \le \mathbb{P}^{\pi^{\mathrm{E}}}\left(G_H^{\mathcal{D}_1}\right),$$

where the first equality is true because $\cup_{x\in\Phi} E_h^x$ corresponds to the event that π^{E} has visited some state uncovered in $\mathcal{D}_1$, and the last inequality holds since $G_h^{\mathcal{D}_1} \subseteq G_H^{\mathcal{D}_1}$ for all $h \in [H]$. Conditioned on $\mathcal{D}_1$, we further have

$$\mathbb{P}(G_H^{\mathcal{D}_1}) \le \sum_{h=1}^{H}\sum_{x\in\Phi} d_h^{\pi^{\mathrm{E}},\phi}(x)\mathbb{I}\{x \notin \Phi_h(\mathcal{D}_1)\}.$$

We first consider the expectation $\mathbb{E}[\sum_{h=1}^{H}\sum_{x\in\Phi} d_h^{\pi^{\mathrm{E}},\phi}(x)\mathbb{I}\{x \notin \Phi_h(\mathcal{D}_1)\}]$, where the expectation is taken over the expert dataset $\mathcal{D}_1$.

$$\mathbb{E}\left[\sum_{h=1}^{H}\sum_{x\in\Phi} d_h^{\pi^{\mathrm{E}},\phi}(x)\mathbb{I}\{x \notin \Phi_h(\mathcal{D}_1)\}\right] \le \sum_{h=1}^{H}\sum_{x\in\Phi} d_h^{\pi^{\mathrm{E}},\phi}(x)\left(1 - d_h^{\pi^{\mathrm{E}},\phi}(x)\right)^{m/2} \le \frac{8|\Phi|H}{9m},$$

where the last step uses the numerical inequality[1] $\max_{x\in[0,1]} x(1-x)^m \le 1/(1+m)\cdot(1-1/m)^m \le 4/(9m)$. With [Rajaraman et al., 2020, Lemma A.3], with probability at least $1-\delta$ with $\delta \in (0, \min\{1, H/5\})$, we have

$$\sum_{h=1}^{H}\sum_{x\in\Phi} d_h^{\pi^{\mathrm{E}},\Phi}(x)\mathbb{I}\{x \notin \Phi_h(\mathcal{D}_1)\} \le \frac{8|\Phi|H}{9m} + \frac{6\sqrt{|\Phi|}H\log(H/\delta)}{m}.$$

Then we have

$$\begin{aligned}\sum_{h=1}^{H}\sum_{x\in\Phi} \mathrm{EB}_h(x) &\le \sum_{h=1}^{H}\sqrt{\left(\frac{8|\Phi|H}{9m} + \frac{6\sqrt{|\Phi|}H\log(2H/\delta)}{m}\right)\frac{3|\Phi|\log(4|\Phi|H/\delta)}{m}} \\ &\le \frac{H^{3/2}|\Phi|}{m}\log^{1/2}\left(\frac{4|\Phi|H}{\delta}\right)\sqrt{\frac{8}{3} + 18\log(2H/\delta)}.\end{aligned}$$

When the expert sample complexity satisfies that $m \gtrsim \frac{H^{3/2}|\Phi|}{\varepsilon}\log\left(\frac{|\Phi|H}{\delta}\right)$, with probability at least $1-\frac{\delta}{2}$ over the randomness of $\mathcal{D}$, we have $\sum_{h=1}^{H}\sum_{x\in\Phi}\mathrm{EB}_h(x) \le \frac{\varepsilon}{2}$. Then, with union bound, with probability at least $1-\delta$, we can obtain

$$\sum_{h=1}^{H}\left\|\widetilde{d}_h^{\pi^{\mathrm{E}},\phi} - d_h^{\pi^{\mathrm{E}},\phi}\right\|_1 \le \sum_{h=1}^{H}\sum_{x\in\Phi}\mathrm{EA}_h(x) + \sum_{h=1}^{H}\sum_{x\in\Phi}\mathrm{EB}_h(x) \le \varepsilon,$$

which completes the proof. □

6 EXPERIMENT DETAILS

Environment. The Reset Cliff MDP is from [Rajaraman et al., 2020, Xu et al., 2021]. The state space $\mathcal{S} = \{1, 2, \cdots, |\mathcal{S}|-1, b\}$ and action space $|\mathcal{A}| = \{1, 2, \ldots, |\mathcal{A}|-1, a^{\mathrm{E}}\}$, where b is a unique absorbing state and a^{E} is the expert action. An example with there states and two actions are shown in Figure 1, where the expert action is shown in green. Only the expert action has a reward $+1$. All non-expert actions have the same transitions and rewards. The initial state distribution $\rho = (1/m, 1/m, 1-|\mathcal{S}|/m+2/m, 0)$.

In our experiments, we implement the Reset Cliff MDP with 20 states and 5 actions. The planning horizon is 20. All algorithms are provided with 100 expert trajectories. All experiments run with 20 random seeds.

[1]The first inequality is based on the basic calculus and the second inequality is based on the fact that $(1-1/x)^x \le 1/e \le 4/9$ while $x \ge 1$.

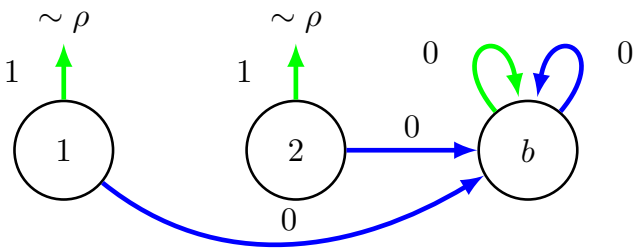

Figure 1: An example of Reset Cliff with three states and two actions. Arrows indicate the transitions and digits indicate the reward values.

Algorithm Implementation. BC directly estimates the expert policy from expert demonstrations. Since the expert policy is deterministic, BC copies the expert action on visited states and takes a uniform policy on non-visited states. The implementation of FEM and GTAL follows the description in [Abbeel and Ng, 2004] and [Syed and Schapire, 2007], respectively.

MB-TAIL first establishes the estimator in Equation (7) with 20% of the environment interactions and learns an empirical transition model by invoking RF-Express [Ménard et al., 2021] to collect the remaining 80% trajectories. Subsequently, MB-TAIL performs policy and reward optimization with the recovered transition model. In particular, MB-TAIL utilizes value iteration to obtain the optimal policy (Line 2 of Algorithm 2). Besides, MB-TAIL utilizes online gradient descent to update the reward function. To utilize the optimization structure, we implement an adaptive step size [Orabona, 2019] rather than the constant step size:

$$\eta_t = \frac{D}{\sqrt{\sum_{i=1}^{t} \left\| \nabla_w f^{(i)}\left(w^{(i)}\right) \right\|_2^2}},$$

where $D = \sqrt{2H|\mathcal{S}||\mathcal{A}|}$ is the diameter of the set $\mathcal{W}$. Conclusions about the sample complexity and computational complexity do not change by this adaptive step size. The number of iterations T of MB-TAIL is 500.

To encourage exploration, OAL adds a bonus function to the Q-function. The bonus function used in the theoretical analysis of [Shani et al., 2022] is too big and impractical. Therefore, we simplify their bonus function from $b_h^k(s,a) = \sqrt{\frac{4H^2|\mathcal{S}|\log(3H^2|\mathcal{S}||\mathcal{A}|n/\delta)}{n_h^k(s,a)\vee 1}}$ to $b_h^k(s,a) = \sqrt{\frac{\log(H|\mathcal{S}||\mathcal{A}|n/\delta)}{n_h^k(s,a)\vee 1}}$, where n is the total number of interactions, δ is the failure probability, $n_h^k(s,a)$ is the number of times visiting (s,a) at time step h until episode k, and $n_h^k(s,a)\vee 1 = \max\{n_h^k(s,a), 1\}$. With the learned transition model and Q-function, OAL uses mirror descent (MD) to optimize the policy and reward function. The step sizes of MD are set by the results in the theoretical analysis of [Shani et al., 2022]. The number of iterations T of OAL is also 500.

References

P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the 21st International Conference on Machine Learning*, pages 1–8, 2004.

D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 2016.

J. Ho and S. Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems 29*, pages 4565–4573, 2016.

N. Jiang, A. Kulesza, and S. Singh. Abstraction selection in model-based reinforcement learning. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 179–188, 2015.

C. Jin, Z. Allen-Zhu, S. Bubeck, and M. I. Jordan. Is q-learning provably efficient? In *Advances in Neural Information Processing Systems 31*, pages 4868–4878, 2018.

L. Li, T. J. Walsh, and M. L. Littman. Towards a unified theory of state abstraction for mdps. *ISAIM*, 4:5, 2006.

P. Ménard, O. D. Domingues, A. Jonsson, E. Kaufmann, E. Leurent, and M. Valko. Fast active learning for pure exploration in reinforcement learning. In *Proceedings of the 38th International Conference on Machine Learning*, pages 7599–7608, 2021.

F. Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

N. Rajaraman, L. F. Yang, J. Jiao, and K. Ramchandran. Toward the fundamental limits of imitation learning. In *Advances in Neural Information Processing Systems 33*, pages 2914–2924, 2020.

S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2): 107–194, 2012.

L. Shani, T. Zahavy, and S. Mannor. Online apprenticeship learning. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, pages 8240–8248, 2022.

U. Syed and R. E. Schapire. A game-theoretic approach to apprenticeship learning. In *Advances in Neural Information Processing Systems 20*, pages 1449–1456, 2007.

R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

T. Xu, Z. Li, and Y. Yu. Error bounds of imitating policies and environments for reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6968–6980, 2021.