Appendix

Table of Contents

A Overview of Notation

In Table A.1, we provide a reference of the notation and symbols used in our paper.

Table A.1: Overview of our notation

Symbol	Name	Signature
$\mathcal{M}$	Markov decision process without reward (MDP\R)	$(\mathcal{S}, \mathcal{A}, P, H, s_0)$
$\mathcal{S}$	State space	
$\mathcal{A}$	Action space	
P	Transition model	$\mathcal{S} \times \mathcal{A} \to \Delta_{\mathcal{S}}$
H	Horizon	$H \in \mathbb{N}^+$
s_0	Initial state	$s_0 \in \mathcal{S}$
π	Policy	$\mathcal{S} \times [H] \to \Delta_{\mathcal{A}}$
r	Reward function	$\mathcal{S} \times \mathcal{A} \times [H] \to [0, R_{\max}]$, $R_{\max} \in \mathbb{R}^+$
$\mathcal{M} \cup r$	Markov decision process (MDP)	$(\mathcal{S}, \mathcal{A}, P, H, s_0, r)$
$Q^{\pi,h}_{\mathcal{M}\cup r}$	Q-function (of π in $\mathcal{M} \cup r$)	$\mathcal{S} \times \mathcal{A} \times [H] \to \mathbb{R}$
$V^{\pi,h}_{\mathcal{M}\cup r}$	Value function (of π in $\mathcal{M} \cup r$)	$\mathcal{S} \times [H] \to \mathbb{R}$
$A^{\pi,h}_{\mathcal{M}\cup r}$	Advantage function (of π in $\mathcal{M} \cup r$)	$\mathcal{S} \times \mathcal{A} \times [H] \to \mathbb{R}$
$\eta^{h,\cdot}_{\mathcal{M},\pi}(\cdot\|s_0)$	State-visitation frequency (conditioned on state)	$[H] \to \Delta_{\mathcal{S}}$
$\eta^{h,\cdot}_{\mathcal{M},\pi}(\cdot\|s_0, a_0)$	State-visitation frequency (conditioned on state-action)	$[H] \to \Delta_{\mathcal{S}}$
$\eta^{h,\cdot}_{\mathcal{M},\pi}(\cdot,\cdot\|s_0)$	State-action-visitation frequency (conditioned on state)	$[H] \times \mathcal{S} \to \Delta_{\mathcal{A}}$
$\eta^{h,\cdot}_{\mathcal{M},\pi}(\cdot,\cdot\|s_0, a_0)$	State-action-visitation frequency (conditioned on state)	$[H] \times \mathcal{S} \to \Delta_{\mathcal{A}}$
$\mathcal{R}_{\mathcal{M}\cup r}$	Feasible set of $\mathcal{M} \cup r$	
$\mathcal{R}_{\mathfrak{B}} = \mathcal{R}_{\mathcal{M}\cup\pi^E}$	Exact feasible set	
$\mathcal{R}_{\hat{\mathfrak{B}}} = \mathcal{R}_{\widehat{\mathcal{M}}\cup\hat{\pi}^E}$	Recovered feasible set	
ϵ	Target accuracy	$\epsilon \in \mathbb{R}^+$
δ	Significancy	$\delta \in (0, 1)$
N_E	Number of exploration episodes	$N_E \in \mathbb{N}^+$

B Proofs of Theoretical Results

B.1 Simulation Lemmas

In this section, we establish several simulation lemmas that we will use throughout our analysis. Some of the results were already derived in prior work for the infinite horizon setting, e.g., by Zanette et al. (2019) and Metelli et al. (2021). For completeness, we provide proofs for all results in the finite-horizon setting.

Definition B.1 (Occupancy measures)**.** *We define $\eta_{\mathcal{M},\pi}^{h,h'}(s|s_0)$ as the probability of being in state s at timestep $h' \geq h$ following a policy π in MDP\R $\mathcal{M}$ starting in state s_0 at timestep h. We can compute it recursively as:*

$$\eta_{\mathcal{M},\pi}^{h,h}(s'|s) \coloneqq \mathbb{1}_{\{s'=s\}}$$
$$\eta_{\mathcal{M},\pi}^{h,h'+1}(s'|s) \coloneqq \sum_{s'',\tilde{a}} P(s'|s'',\tilde{a})\pi_{h'}(\tilde{a}|s'')\eta_{\mathcal{M},\pi}^{h,h'}(s''|s)$$

We define the same probability for state-action pairs analogously:

$$\eta_{\mathcal{M},\pi}^{h,h'}(s',a'|s,a) \coloneqq \mathbb{1}_{\{s'=s,a'=a\}}$$
$$\eta_{\mathcal{M},\pi}^{h,h'+1}(s',a'|s,a) \coloneqq \sum_{\tilde{s},\tilde{a}} \pi_{h'}(a'|s')P(s'|\tilde{s},\tilde{a})\eta_{\mathcal{M},\pi}^{h,h'}(\tilde{s},\tilde{a}|s,a)$$

as well as

$$\eta_{\mathcal{M},\pi}^{h,h}(s',a'|s) \coloneqq \pi_h(a'|s')\mathbb{1}_{\{s'=s\}}$$
$$\eta_{\mathcal{M},\pi}^{h,h'+1}(s',a'|s) \coloneqq \sum_{\tilde{s},\tilde{a}} \pi_{h'}(a'|s')P(s'|\tilde{s},\tilde{a})\eta_{\mathcal{M},\pi}^{h,h'}(\tilde{s},\tilde{a}|s)$$

Because the environment is Markovian, it also holds for $h' > h$ that

$$\eta_{\mathcal{M},\pi}^{h,h'}(s'|s) = \sum_{\tilde{s},a} \eta_{\mathcal{M},\pi}^{h+1,h'}(s'|\tilde{s})P(\tilde{s}|s,a)\pi_h(a|s)$$

and equivalently for state-action pairs.

Lemma B.2. *The value function and Q-function of a policy π in an MDP $\mathcal{M} \cup r$ at timestep h can be expressed as:*

$$V_{\mathcal{M}\cup r}^{\pi,h}(s) = \sum_{h'=h}^{H} \sum_{s',a'} \eta_{\mathcal{M},\pi}^{h,h'}(s',a'|s)r_{h'}(s',a')$$

$$Q_{\mathcal{M}\cup r}^{\pi,h}(s,a) = \sum_{h'=h}^{H} \sum_{s',a'} \eta_{\mathcal{M},\pi}^{h,h'}(s',a'|s,a)r_{h'}(s',a')$$

Proof. We show the result for the value function; the derivation for the Q-function is analogous.

Note that for $h = H$ the statement holds because $V_{\mathcal{M}\cup r}^{\pi,H}(s) = 0$. The general result follows by induction. Assume that for $h+1$ the statement holds. Then:

$$\begin{aligned}
V_{\mathcal{M}\cup r}^{\pi,h}(s) &\overset{(a)}{=} \sum_a \pi_h(a|s)\left(r_h(s,a) + \sum_{s'} P(s'|s,a)V_{\mathcal{M}\cup r}^{\pi,h+1}(s')\right) \\
&\overset{(b)}{=} \sum_a \pi_h(a|s)\left(r_h(s,a) + \sum_{s'} P(s'|s,a)\left(\sum_{h'=h+1}^{H}\sum_{s'',a''} \eta_{\mathcal{M},\pi}^{h+1,h'}(s'',a''|s')r_{h'}(s'',a'')\right)\right) \\
&\overset{(c)}{=} \sum_a \pi_h(a|s)r_h(s,a) + \sum_{h'=h+1}^{H}\sum_{s',a'} \eta_{\mathcal{M},\pi}^{h,h'}(s'|s)\pi_{h'}(a'|s')r_{h'}(s',a') \\
&\overset{(d)}{=} \sum_{h'=h}^{H}\sum_{s',a'} \eta_{\mathcal{M},\pi}^{h,h'}(s'|s)\pi_{h'}(a'|s')r_{h'}(s',a')
\end{aligned}$$

where (a) uses the Bellman equation, (b) the induction step, (c) uses Definition B.1 and relabels $s'' \to s'$, $a'' \to a'$, and (d) uses Definition B.1 again and relabels $a \to a'$. □

Lemma B.3 (Simulation lemma 1 by Metelli et al. (2021))**.** *Let $\mathcal{M}$ be an MDP\R, and $r, \hat{r}$ two reward functions with corresponding optimal policies $\pi^*, \hat{\pi}^*$. Then,*

$$Q^{\pi^*,h}_{\mathcal{M}\cup r}(s,a) - Q^{\hat{\pi}^*,h}_{\mathcal{M}\cup \hat{r}}(s,a) \leq \sum_{h'=h}^{H} \sum_{s',a'} \eta^{h,h'}_{\mathcal{M},\pi^*}(s',a'|s,a) \left(r_{h'}(s',a') - \hat{r}_{h'}(s',a')\right)$$

$$V^{\pi^*,h}_{\mathcal{M}\cup r}(s) - V^{\hat{\pi}^*,h}_{\mathcal{M}\cup \hat{r}}(s) \leq \sum_{h'=h}^{H} \sum_{s',a'} \eta^{h,h'}_{\mathcal{M},\pi^*}(s',a'|s) \left(r_{h'}(s',a') - \hat{r}_{h'}(s',a')\right)$$

Proof. Note that $Q^{\hat{\pi}^*,h}_{\mathcal{M}\cup \hat{r}}(s,a) \geq Q^{\pi^*,h}_{\mathcal{M}\cup \hat{r}}(s,a)$ for all s, a because $\hat{\pi}^*$ is optimal for $\hat{r}$. Hence

$$\begin{aligned} Q^{\pi^*,h}_{\mathcal{M}\cup r}(s,a) - Q^{\hat{\pi}^*,h}_{\mathcal{M}\cup \hat{r}}(s,a) &\leq Q^{\pi^*,h}_{\mathcal{M}\cup r}(s,a) - Q^{\pi^*,h}_{\mathcal{M}\cup \hat{r}}(s,a) \\ &\overset{(a)}{=} \sum_{h'=h}^{H} \sum_{s',a'} \eta^{h,h'}_{\mathcal{M},\pi^*}(s',a'|s,a)(r_{h'}(s',a') - \hat{r}_{h'}(s',a')) \end{aligned}$$

where (a) uses Lemma B.2. After observing $V^{\hat{\pi}^*,h}_{\mathcal{M}\cup \hat{r}}(s) \geq V^{\pi^*,h}_{\mathcal{M}\cup \hat{r}}(s)$, the second result follows analogously. □

Lemma 6. *Let $\mathcal{M}$ be an MDP\R, $r, \hat{r}$ two reward functions with optimal policies $\pi^*, \hat{\pi}^*$. Then,*

$$Q^{\pi^*,h}_{\mathcal{M}\cup r}(s,a) - Q^{\hat{\pi}^*,h}_{\mathcal{M}\cup r}(s,a) \leq \sum_{h'=h}^{H} \sum_{s',a'} \left(\eta^{h,h'}_{\mathcal{M},\pi^*}(s',a'|s,a) - \eta^{h,h'}_{\mathcal{M},\hat{\pi}^*}(s',a'|s,a)\right) \left(r_{h'}(s',a') - \hat{r}_{h'}(s',a')\right)$$

Proof.

$$\begin{aligned} &Q^{\pi^*,h}_{\mathcal{M}\cup r}(s,a) - Q^{\hat{\pi}^*,h}_{\mathcal{M}\cup r}(s,a) = (Q^{\pi^*,h}_{\mathcal{M}\cup r}(s,a) - Q^{\hat{\pi}^*,h}_{\mathcal{M}\cup \hat{r}}(s,a)) + (Q^{\hat{\pi}^*,h}_{\mathcal{M}\cup \hat{r}}(s,a) - Q^{\hat{\pi}^*,h}_{\mathcal{M}\cup r}(s,a)) \\ &\overset{(a)}{\leq} \sum_{h'=h}^{H} \sum_{s',a'} \eta^{h,h'}_{\mathcal{M},\pi^*}(s',a'|s,a) \left(r_{h'}(s',a') - \hat{r}_{h'}(s',a')\right) + (Q^{\hat{\pi}^*,h}_{\mathcal{M}\cup \hat{r}}(s,a) - Q^{\hat{\pi}^*,h}_{\mathcal{M}\cup r}(s,a)) \\ &\overset{(b)}{=} \sum_{h'=h}^{H} \sum_{s',a'} \eta^{h,h'}_{\mathcal{M},\pi^*}(s',a'|s,a) \left(r_{h'}(s',a') - \hat{r}_{h'}(s',a')\right) + \sum_{h'=h}^{H} \sum_{s',a'} \eta^{h,h'}_{\mathcal{M},\hat{\pi}^*}(s',a'|s,a) \left(\hat{r}_{h'}(s',a') - r_{h'}(s',a')\right) \\ &= \sum_{h'=h}^{H} \sum_{s',a'} \left(\eta^{h,h'}_{\mathcal{M},\pi^*}(s',a'|s,a) - \eta^{h,h'}_{\mathcal{M},\hat{\pi}^*}(s',a'|s,a)\right) \left(r_{h'}(s',a') - \hat{r}_{h'}(s',a')\right) \end{aligned}$$

where (a) uses Lemma B.3 and (b) uses Lemma B.2. □

Lemma B.4. *Let $\mathcal{M}_1, \mathcal{M}_2$ be two MDP\R with transition dynamics P_1, P_2 respectively, r a reward function and π a policy. Then, for any state s and timestep h:*

$$V^{\pi,h}_{\mathcal{M}_2\cup r}(s) - V^{\pi,h}_{\mathcal{M}_1\cup r}(s) = \sum_{h'=h}^{H} \sum_{s',a',s''} \eta^{h,h'}_{\mathcal{M}_2,\pi}(s';s)\pi_{h'}(a'|s')(P_2(s''|s',a') - P_1(s''|s',a'))V^{\pi,h'+1}_{\mathcal{M}_1\cup r}(s'')$$

$$V^{\pi,h}_{\mathcal{M}_1\cup r}(s) - V^{\pi,h}_{\mathcal{M}_2\cup r}(s) = \sum_{h'=h}^{H} \sum_{s',a',s''} \eta^{h,h'}_{\mathcal{M}_2,\pi}(s';s)\pi_{h'}(a'|s')(P_1(s''|s',a') - P_2(s''|s',a'))V^{\pi,h'+1}_{\mathcal{M}_1\cup r}(s'')$$

Moreover,

$$\left|V^{\pi,h}_{\mathcal{M}_2\cup r}(s) - V^{\pi,h}_{\mathcal{M}_1\cup r}(s)\right| \leq \sum_{h'=h}^{H} \sum_{s',a',s''} \eta^{h,h'}_{\mathcal{M}_2,\pi}(s';s)\pi_{h'}(a'|s')\left|P_2(s''|s',a') - P_1(s''|s',a')\right|V^{\pi,h'+1}_{\mathcal{M}_1\cup r}(s'')$$

Proof. We start by writing explicitly the value-functions:

$$\begin{aligned} &V^{\pi,h}_{\mathcal{M}_2\cup r}(s) - V^{\pi,h}_{\mathcal{M}_1\cup r}(s) = \sum_{a,s'} \pi_h(a|s) \left(P_2(s'|s,a)V^{\pi,h+1}_{\mathcal{M}_2\cup r}(s') - P_1(s'|s,a)V^{\pi,h+1}_{\mathcal{M}_1\cup r}(s') \pm P_2(s'|s,a)V^{\pi,h+1}_{\mathcal{M}_1\cup r}(s')\right) \\ &= \sum_{a,s'} \pi_h(a|s) \left((P_2(s'|s,a) - P_1(s'|s,a))V^{\pi,h+1}_{\mathcal{M}_1\cup r}(s') + P_2(s'|s,a)(V^{\pi,h+1}_{\mathcal{M}_2\cup r}(s') - V^{\pi,h+1}_{\mathcal{M}_1\cup r}(s'))\right) \end{aligned}$$

Unrolling the recursion gives the first result; the second result follows similarly:

$$V^{\pi,h}_{\mathcal{M}_1\cup r}(s) - V^{\pi,h}_{\mathcal{M}_2\cup r}(s) = \sum_{a,s'} \pi_h(a|s) \Big(P_1(s'|s,a) V^{\pi,h+1}_{\mathcal{M}_1\cup r}(s') - P_2(s'|s,a) V^{\pi,h+1}_{\mathcal{M}_2\cup r}(s') \pm P_2(s'|s,a) V^{\pi,h+1}_{\mathcal{M}_1\cup r}(s') \Big)$$

$$= \sum_{a,s'} \pi_h(a|s) \Big((P_1(s'|s,a) - P_2(s'|s,a)) V^{\pi,h+1}_{\mathcal{M}_1\cup r}(s') + P_2(s'|s,a)(V^{\pi,h+1}_{\mathcal{M}_1\cup r}(s') - V^{\pi,h+1}_{\mathcal{M}_2\cup r}(s')) \Big)$$

Together, the first two results imply the third one because all terms in the sums are non-negative. □

Lemma B.5. *Let $\mathcal{M}_1, \mathcal{M}_2$ be two MDP\R with transition dynamics P_1, P_2 respectively, r a reward function, and π_1^*, π_2^* optimal policy in $\mathcal{M}_1 \cup r$ and $\mathcal{M}_2 \cup r$, respectively. Then, for any state s and timestep h:*

$$V^{*,h}_{\mathcal{M}_1\cup r}(s) - V^{*,h}_{\mathcal{M}_2\cup r}(s) \le \sum_{h'=h} \sum_{s',a',s''} \eta^{h,h'}_{\mathcal{M}_2,\pi_1^*}(s';s) \pi^*_{1,h}(a'|s') (P_1(s''|s',a') - P_2(s''|s',a')) V^{*,h}_{\mathcal{M}_1\cup r}(s'')$$

$$V^{*,h}_{\mathcal{M}_2\cup r}(s) - V^{*,h}_{\mathcal{M}_1\cup r}(s) \le \sum_{h'=h} \sum_{s',a',s''} \eta^{h,h'}_{\mathcal{M}_2,\pi_2^*}(s';s) \pi^*_{2,h}(a'|s') (P_2(s''|s',a') - P_1(s''|s',a')) V^{*,h}_{\mathcal{M}_2\cup r}(s'')$$

Proof.

$$\begin{aligned}
V^{*,h}_{\mathcal{M}_1\cup r}(s) - V^{*,h}_{\mathcal{M}_2\cup r}(s) &= \sum_{a,s'} \Big(\pi^*_{1,h}(a|s) P_1(s'|s,a) V^{\pi_1^*,h+1}_{\mathcal{M}_1\cup r}(s') - \pi^*_{2,h}(a|s) P_2(s'|s,a) V^{\pi_2^*,h+1}_{\mathcal{M}_2\cup r}(s') \\
&\qquad \pm \pi^*_{1,h}(a|s) P_2(s'|s,a) V^{\pi_1^*,h+1}_{\mathcal{M}_1\cup r}(s') \pm \pi^*_{1,h}(a|s) P_2(s'|s,a) V^{\pi_2^*,h+1}_{\mathcal{M}_2\cup r}(s') \Big) \\
&= \sum_{a,s'} \Big(\pi^*_{1,h}(a|s) P_2(s'|s,a) (V^{\pi_1^*,h+1}_{\mathcal{M}_1\cup r}(s') - V^{\pi_2^*,h+1}_{\mathcal{M}_2\cup r}(s')) \\
&\qquad + \pi^*_{1,h}(a|s) (P_1(s'|s,a) - P_2(s'|s,a)) V^{\pi_1^*,h+1}_{\mathcal{M}_1\cup r}(s') \\
&\qquad + (\pi^*_{1,h}(a|s) - \pi^*_{2,h}(a|s)) P_2(s'|s,a) V^{\pi_2^*,h+1}_{\mathcal{M}_2\cup r}(s') \Big) \\
&\le \sum_{a,s'} \Big(\pi^*_{1,h}(a|s) P_2(s'|s,a) (V^{\pi_1^*,h+1}_{\mathcal{M}_1\cup r}(s') - V^{\pi_2^*,h+1}_{\mathcal{M}_2\cup r}(s')) \\
&\qquad + \pi^*_{1,h}(a|s) (P_1(s'|s,a) - P_2(s'|s,a)) V^{\pi_1^*,h+1}_{\mathcal{M}_1\cup r}(s') \Big)
\end{aligned}$$

where the last inequality uses that π^* is optimal for $\mathcal{M}_2 \cup r$. Unrolling the recursion gives the first result. A similar argument yields the second results:

$$\begin{aligned}
V^{*,h}_{\mathcal{M}_2\cup r}(s) - V^{*,h}_{\mathcal{M}_1\cup r}(s) &= \sum_{a,s'} \Big(\pi^*_{2,h}(a|s) P_2(s'|s,a) V^{\pi_2^*,h+1}_{\mathcal{M}_2\cup r}(s') - \pi^*_{1,h}(a|s) P_1(s'|s,a) V^{\pi_1^*,h+1}_{\mathcal{M}_1\cup r}(s') \\
&\qquad \pm \pi^*_{2,h}(a|s) P_2(s'|s,a) V^{\pi_1^*,h+1}_{\mathcal{M}_1\cup r}(s') \Big) \\
&= \sum_{a,s'} \Big(\pi^*_{2,h}(a|s) P_2(s'|s,a) (V^{\pi_2^*,h+1}_{\mathcal{M}_2\cup r}(s') - V^{\pi_1^*,h+1}_{\mathcal{M}_1\cup r}(s')) \\
&\qquad + \pi^*_{2,h}(a|s) P_2(s'|s,a) V^{\pi_1^*,h+1}_{\mathcal{M}_1\cup r}(s') - \pi^*_{1,h}(a|s) P_1(s'|s,a) V^{\pi_1^*,h+1}_{\mathcal{M}_1\cup r}(s') \\
&\le \sum_{a,s'} \Big(\pi^*_{2,h}(a|s) P_2(s'|s,a) (V^{\pi_2^*,h+1}_{\mathcal{M}_2\cup r}(s') - V^{\pi_1^*,h+1}_{\mathcal{M}_1\cup r}(s')) \\
&\qquad + \pi^*_{2,h}(a|s) (P_2(s'|s,a) - P_1(s'|s,a)) V^{\pi_1^*,h+1}_{\mathcal{M}_1\cup r}(s') \\
&\qquad \Big)
\end{aligned}$$

□

B.2 Feasible Reward Set

In this section, we characterize the feasible reward set first implicitly, then explicitly, and prove a result about error propagation. Metelli et al. (2021) provide a similar analysis in the infinite horizon setting.

Lemma 3 (Feasible Reward Set Implicit)**.** *A reward function r is feasible if and only if for all s, a, h it holds that: $A^{\pi,h}_{\mathcal{M}\cup r}(s,a) = 0$ if $\pi^E_h(a|s) \ge 0$ and $A^{\pi,h}_{\mathcal{M}\cup r}(s,a) \le 0$ if $\pi^E_h(a|s) = 0$. Moreover, if the second inequality is strict, π^E is uniquely optimal, i.e., $\Pi^*_{\mathcal{M}\cup r} = \{\pi^E\}$.*

Proof. The result follows directly from Definition 1. □

Lemma B.6. *A Q-function satisfies the conditions of Lemma 3 if and only if there exists an* $\{A_h \in \mathbb{R}_{\geq 0}^{\mathcal{S}\times\mathcal{A}}\}_{h\in H}$ *and* $\{V_h \in \mathbb{R}^S\}$ *such that for every* $h, s, a \in [H] \times \mathcal{S} \times \mathcal{A}$*:*

$$Q^{\pi^E,h}_{\mathcal{M}\cup r}(s,a) = -A_h(s,a)\mathbb{1}_{\{\pi_h^E(a|s)=0\}} + V_h(s)$$

Proof. We first show that if $Q^{\pi^E,h}_{\mathcal{M}\cup r}(s,a)$ has this form, the conditions of Lemma 3 are satisfied, and then the converse. Assume $Q^{\pi^E,h}_{\mathcal{M}\cup r}(s,a) = -A_h(s,a)\mathbb{1}_{\{\pi_h^E(a|s)=0\}} + V_h(s)$. Then,

$$V^{\pi^E,h}_{\mathcal{M}\cup r}(s) = \sum_a \pi_h^E(a|s) Q^{\pi^E,h}_{\mathcal{M}\cup r}(s,a) = V_h(s).$$

If $\pi_h^E(a|s) > 0$, then $Q^{\pi^E,h}_{\mathcal{M}\cup r}(s,a) = V^{\pi^E,h}_{\mathcal{M}\cup r}(s)$, which is the first condition of Lemma 3. If $\pi_h^E(a|s) = 0$, $Q^{\pi^E,h}_{\mathcal{M}\cup r}(s,a) = V^{\pi^E,h}_{\mathcal{M}\cup r}(s) - A_h(s,a) \leq V^{\pi^E,h}_{\mathcal{M}\cup r}(s)$, which is the second condition of Lemma 3.

For the converse, assume that the conditions of Lemma 3 hold, and let $V_h(s) = V^{\pi^E,h}_{\mathcal{M}\cup r}(s)$ and $A_h(s,a) = V^{\pi^E,h}_{\mathcal{M}\cup r}(s) - Q^{\pi^E,h}_{\mathcal{M}\cup r}(s,a)$. □

Lemma 4 (Feasible Reward Set Explicit)**.** *A reward function* r *is feasible if and only if there exists an* $\{A_h \in \mathbb{R}_{\geq 0}^{\mathcal{S}\times\mathcal{A}}\}_{h\in[H]}$ *and* $\{V_h \in \mathbb{R}^{\mathcal{S}}\}_{h\in[H]}$ *such that for all* s, a, h *it holds that:*

$$r_h(s,a) = -A_h(s,a)\mathbb{1}_{\{\pi_h^E(a|s)=0\}} + V_h(s) + \sum_{s'} P(s'|s,a)V_{h+1}(s')$$

Proof. Since $Q^{\pi^E,h}_{\mathcal{M}\cup r}(s,a) = r_h(s,a) + \sum_{s'} P(s'|s,a)V_{h+1}(s')$, using Lemma B.6, we have:

$$\begin{aligned} r_h(s,a) &= Q^{\pi^E,h}_{\mathcal{M}\cup r}(s,a) - \sum_{s'} P(s'|s,a)V_{h+1}(s') \\ &= -A_h(s,a)\mathbb{1}_{\{\pi_h^E(a|s)=0\}} + V_h(s) + \sum_{s'} P(s'|s,a)V_{h+1}(s') \end{aligned}$$

□

Theorem 5 (Error Propagation)**.** *Let* $(\mathcal{M}, \pi^E)$ *and* $(\widehat{\mathcal{M}}, \widehat{\pi}^E)$ *be two IRL problems. Then, for any* $r \in \mathcal{R}_{(\mathcal{M},\pi^E)}$ *there exists* $\widehat{r} \in \widehat{\mathcal{R}}_{(\widehat{\mathcal{M}},\widehat{\pi}^E)}$ *such that:*

$$|r_h(s,a) - \widehat{r}_h(s,a)| \leq A_h(s,a)|\pi_h^E(a|s) - \widehat{\pi}_h^E(a|s)| + \sum_{s'} V_{h+1}(s')|P(s'|s,a) - \widehat{P}(s'|s,a)|$$

and we can bound $V_h \leq (H-h)R_{\max}$ *and* $A_h \leq (H-h)R_{\max}$*.*

Proof. We start by rewriting r and $\hat{r}$ using Lemma 4:

$$\begin{aligned} r_h(s,a) &= -A_h(s,a)\mathbb{1}_{\{\pi_h^E(a|s)=0\}} + V_h(s) + \sum_{s'} P(s'|s,a)V_{h+1}(s') \\ \hat{r}_h(s,a) &= -\hat{A}_h(s,a)\mathbb{1}_{\{\hat{\pi}_h^E(a|s)=0\}} + \hat{V}_h(s) + \sum_{s'} \widehat{P}(s'|s,a)\hat{V}_{h+1}(s') \end{aligned}$$

We can choose (w.l.o.g.) $V_h = \hat{V}_h$ and $\hat{A}_h = \mathbb{1}_{\{\pi_h^E(a|s)=0\}}A_h$:

$$\begin{aligned} r_h(s,a) - \hat{r}_h(s,a) =& - A_h(s,a)\mathbb{1}_{\{\pi_h^E(a|s)=0\}} + V_h(s) + \sum_{s'} P(s'|s,a)V_{h+1}(s') \\ &+ A_h(s,a)\mathbb{1}_{\{\hat{\pi}_h^E(a|s)=0\}}\mathbb{1}_{\{\pi_h^E(a|s)=0\}} - V_h(s) - \sum_{s'} \widehat{P}(s'|s,a)V_{h+1}(s') \\ =& A_h(s,a)\mathbb{1}_{\{\pi_h^E(a|s)=0\}}(\mathbb{1}_{\{\hat{\pi}_h^E(a|s)=0\}} - 1) + \sum_{s'} V_{h+1}(s')(P(s'|s,a) - \widehat{P}(s'|s,a)) \\ =& - A_h(s,a)\mathbb{1}_{\{\pi_h^E(a|s)=0\}}\mathbb{1}_{\{\hat{\pi}_h(a|s)\geq 0\}} + \sum_{s'} V_{h+1}(s')(P(s'|s,a) - \widehat{P}(s'|s,a)) \end{aligned}$$

The result follows by taking the absolute value and applying the triangle inequality. □

Algorithm 2 Uniform sampling IRL with a generative model.

Input: significance $\delta \in (0,1)$, target accuracy ϵ, maximum number of samples per iter. $n_{\max}$
Initialize $k \leftarrow 0$, $\epsilon_0 \leftarrow H$
while $\epsilon_k > \epsilon/2$ **do**
    Uniformly sample $\lceil \frac{n_{\max}}{SAH} \rceil$ samples from all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$
    For all samples, observe sample from transition dynamics and expert policy
    $k \leftarrow k+1$
    Update $\widehat{P}_k$, $\hat{\pi}_k$, and C_k^h
    Update accuracy $\epsilon_k \leftarrow H \max_{s,a,h} C_k^h(s,a)$
end while

B.3 Uniform Sampling IRL with a Generative Model

In this section, we derive sample complexity results for uniform sampling with a generative model (Algorithm 2). Metelli et al. (2021) proved an analogous result for the infinite horizon setting focusing on transferable rewards. In contrast, our focus is on the finite horizon setting. Moreover, Metelli et al. (2021) considers to learn a reward that is transferable to a known target environment. In our setting, instead, we suppose to use the recovered reward function in the unknown source environment.

Definition 2 (Optimality Criterion)**.** *Let $\mathcal{R}_{\mathfrak{B}}$ be the exact feasible set and $\mathcal{R}_{\hat{\mathfrak{B}}}$ be the feasible set recovered after observing $n \geq 0$ samples collected from $\mathcal{M}$ and π^E. We say that an algorithm for Active IRL is (ϵ, δ, n)-*correct *if after n iterations with probability at least $1-\delta$ it holds that:*

$$\inf_{\hat{r} \in \mathcal{R}_{\hat{\mathfrak{B}}}} \sup_{\hat{\pi}^* \in \Pi^*_{\widehat{\mathcal{M}} \cup \hat{r}}} \max_{s,a,h} \left| Q^{\pi^*, h}_{\mathcal{M} \cup r}(s,a) - Q^{\hat{\pi}^*, h}_{\mathcal{M} \cup r}(s,a) \right| \leq \epsilon \quad \textit{for each } r \in \mathcal{R}_{\mathfrak{B}},$$

$$\inf_{r \in \mathcal{R}_{\mathfrak{B}}} \sup_{\pi^* \in \Pi^*_{\mathcal{M} \cup r}} \max_{s,a,h} \left| Q^{\pi^*, h}_{\mathcal{M} \cup r}(s,a) - Q^{\hat{\pi}^*, h}_{\mathcal{M} \cup r}(s,a) \right| \leq \epsilon \quad \textit{for each } \hat{r} \in \mathcal{R}_{\hat{\mathfrak{B}}},$$

where π^* is an optimal policy in $\mathcal{M} \cup r$ and $\hat{\pi}^*$ is an optimal policy in $\widehat{\mathcal{M}} \cup \hat{r}$.

Lemma B.7 (Good Event)**.** *Let π^E be a (possibly stochastic) expert policy. We estimate the expert policy with $\hat{\pi}^E$ and the transition model P with an estimate $\widehat{P}_k$ from k episodic interactions. Let $n_k^h(s,a)$ and $n_k^h(s)$ be the number of times state action pairs and states have been observed at time h within the first k episodes, and $n_k^{h+}(s,a) = \max\{1, n_k^h(s,a)\}$. Then,*

$$\mathbb{1}_{\{\pi_h^E(a|s)=0\}} \mathbb{1}_{\{\hat{\pi}_h^E(a|s) \geq 0\}} A_h(s,a) \leq (H-h) R_{\max} \sqrt{\frac{\ell_k^h(s,a)}{n_k^{h+}(s,a)}}$$

$$\mathbb{1}_{\{\hat{\pi}_h^E(a|s)=0\}} \mathbb{1}_{\{\pi_h^E(a|s) \geq 0\}} \hat{A}_h(s,a) \leq (H-h) R_{\max} \sqrt{\frac{\ell_k^h(s,a)}{n_k^{h+}(s,a)}}$$

$$\sum_{s'} |(P(s'|s,a) - \hat{P}_k(s'|s,a)) V_r^{\pi,h}(s')| \leq (H-h) R_{\max} \sqrt{\frac{2\ell_k^h(s,a)}{n_k^{h+}(s,a)}}$$

$$\sum_{s'} |(P(s'|s,a) - \hat{P}_k(s'|s,a)) \hat{V}_r^{\pi,h}(s')| \leq (H-h) R_{\max} \sqrt{\frac{2\ell_k^h(s,a)}{n_k^{h+}(s,a)}}$$

where $\ell_k^h(s,a) = \log\left(24SAH(n_k^{h+}(s,a))^2/\delta\right)$, holds simultaneously for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$ and $k \geq 1$ with probability at least $1-\delta$. We call the event that these equations hold the good event *$\mathscr{E}$ and write $P(\mathscr{E}) \geq 1-\delta$.*

Proof. We show that each statement individually does not hold with probability less than $\delta/4$, which implies the result via a union bound. Let us denote $\beta_1(s,a,h) \coloneqq (H-h) R_{\max} \sqrt{\frac{2\ell_k^h(s,a)}{n_k^{h+}(s,a)}}$. First, consider the last two inequalities. The probability that either of them does not hold is:

$$
\begin{aligned}
&\Pr\left(\exists k \geq 1, (s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H] : \sum_{s'} |(P(s'|s,a) - \widehat{P}_k(s'|s,a))V_r^{\pi,h}(s')| > \beta_1(s,a,h)\right) \\
&\overset{(a)}{\leq} \Pr\left(\exists m \geq 0, (s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H] : \sum_{s'} |(P(s'|s,a) - \widehat{P}_k(s'|s,a))V_r^{\pi,h}(s')| > \beta_1(s,a,h)\right) \\
&\overset{(b)}{\leq} \sum_{m\geq 0}\sum_{s,a}\sum_{h=0}^{H} \Pr\left(\sum_{s'} |(P(s'|s,a) - \widehat{P}_k(s'|s,a))V_r^{\pi,h}(s')| > \beta_1(s,a,h)\right) \\
&\overset{(c)}{\leq} \sum_{m\geq 0}\sum_{s,a}\sum_{h=0}^{H} 2\exp\left(-\frac{2\beta_1(s,a,h)^2 m^2}{4m(H-h)^2 R_{\max}^2}\right) \leq \sum_{m\geq 0}\sum_{s,a}\sum_{h=0}^{H} 2\exp\left(-\ell_k(s,a)\right) \\
&= \sum_{m\geq 0}\sum_{s,a}\sum_{h=0}^{H} \frac{2\delta}{24SAH(m^+)^2} = \frac{\delta}{12}\left(1 + \sum_{m\geq 0}\frac{1}{m^2}\right) = \frac{\delta}{12}\left(1+\frac{\pi^2}{6}\right) \leq \frac{\delta}{4}
\end{aligned}
$$

Step (a) assumes that we visit a state action pair m times, and focuses on these m times the transition model for the given state-action pair is updated. Step (b) uses a union bound over m and (s,a). Step (c) applies Hoeffding's inequality using that we estimate P with an average of samples, and $V_r^{\pi,h} \leq (H-h)R_{\max}$. The factor m^2 in the numerator results from dividing by $1/m$ to average over samples, and the factor $4m$ in the denominator results from the sum over m in the denominator of Hoeffding's bound.

We show the first two inequalities similarly, with $\beta_2(s,a,h) \coloneqq (H-h)R_{\max}\sqrt{\frac{\ell_k^h(s,a)}{n_k^{h+}(s,a)}}$

$$
\begin{aligned}
&\Pr\left(\exists k \geq 1, (s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H] : |(\pi_k^E(a|s) - \hat{\pi}_k^E(a|s))V_r^{\pi,h}(s')| > \beta_2(s,a,h)\right) \\
&\overset{(a)}{\leq} \Pr\left(\exists m \geq 0, (s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H] : |(\pi_k^E(a|s) - \hat{\pi}_k^E(a|s))V_r^{\pi,h}(s')| > \beta_2(s,a,h)\right) \\
&\overset{(b)}{\leq} \sum_{m\geq 0}\sum_{s,a}\sum_{h=0}^{H} \Pr\left(|(\pi_k^E(a|s) - \hat{\pi}_k^E(a|s))V_r^{\pi,h}(s')| > \beta_2(s,a,h)\right) \\
&\overset{(c)}{\leq} \sum_{m\geq 0}\sum_{s,a}\sum_{h=0}^{H} 2\exp\left(-\frac{2\beta_2(s,a,h)^2 m^2}{m(H-h)^2 R_{\max}^2}\right) \leq \sum_{m\geq 0}\sum_{s,a}\sum_{h=0}^{H} 2\exp\left(-\ell_k(s,a)\right) \\
&= \sum_{m\geq 0}\sum_{s,a}\sum_{h=0}^{H} \frac{2\delta}{24SAH(m^+)^2} = \frac{\delta}{12}\left(1 + \sum_{m\geq 0}\frac{1}{m^2}\right) = \frac{\delta}{12}\left(1+\frac{\pi^2}{6}\right) \leq \frac{\delta}{4}
\end{aligned}
$$

A union bound over all equations results in $P(\mathscr{E}) \geq 1-\delta$. □

Definition B.8. *We define the reward uncertainty as*

$$
C_k^h(s,a) = (H-h)R_{\max}\min\left(1, 2\sqrt{\frac{2\ell_k^h(s,a)}{n_k^h(s,a)}}\right)
$$

Corollary B.9. *Under the good event $\mathscr{E}$, in each iteration k it holds for all $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$ that:*

$$
|r_h(s,a) - \hat{r}_h^k(s,a)| \leq C_k^h(s,a)
$$

Proof.

$$
\begin{aligned}
|r_h(s,a) - \hat{r}_h^k(s,a)| &\overset{(a)}{\leq} A_h(s,a)\mathbb{1}_{\{\pi_h^E(a|s)=0\}}\mathbb{1}_{\{\hat{\pi}_h^E(a|s)\geq 0\}} + \sum_{s'} V_{h+1}(s')|P(s'|s,a) - \widehat{P}(s'|s,a)| \\
&\overset{(b)}{\leq} (H-h)R_{\max}\left(2\sqrt{\frac{2\ell_k^h(s,a)}{n_k^{h+}(s,a)}}\right) = C_k^h(s,a)
\end{aligned}
$$

where (a) uses Theorem 5 and (b) uses Lemma B.7. □

Corollary B.10. *Let $\mathscr{S}$ be a sampling strategy. Let $\mathcal{R}_{\mathfrak{B}}$ be the exact feasible set and $\mathcal{R}_{\hat{\mathfrak{B}}_k}$ be the feasible set recovered after k iterations. If*

$$
H\max_{s,a,h} C_k^h(s,a) \leq \frac{\epsilon}{2},
$$

then the conditions of Definition 2 are satisfied.

Proof. For the first condition of Definition 2, observe:

$$\begin{aligned}
&\inf_{\hat{r}\in\mathcal{R}_{\hat{\mathfrak{B}}_k}} \sup_{\hat{\pi}^*\in\Pi^*_{\widehat{\mathcal{M}}\cup\hat{r}}} \max_{s,a,h}(Q^{\pi^*,h}_{\mathcal{M}\cup r}(s,a) - Q^{\hat{\pi}^*,h}_{\mathcal{M}\cup r}(s,a)) \\
\overset{(a)}{\leq} &\inf_{\hat{r}\in\mathcal{R}_{\hat{\mathfrak{B}}_k}} \sup_{\hat{\pi}^*\in\Pi^*_{\widehat{\mathcal{M}}\cup\hat{r}}} \max_{s,a,h} \sum_{h'=h}^{H}\sum_{s',a'} \left(\eta^{h,h'}_{\mathcal{M},\pi^*}(s',a'|s,a) - \eta^{h,h'}_{\mathcal{M},\hat{\pi}^*}(s',a'|s,a)\right)\left(r_{h'}(s',a') - \hat{r}_{h'}(s',a')\right) \\
\overset{(b)}{\leq} &\inf_{\hat{r}\in\mathcal{R}_{\hat{\mathfrak{B}}_k}} \sup_{\hat{\pi}^*\in\Pi^*_{\widehat{\mathcal{M}}\cup\hat{r}}} \max_{s,a,h} \Big|\sum_{h'=h}^{H}\sum_{s',a'} \left(\eta^{h,h'}_{\mathcal{M},\pi^*}(s',a'|s,a) - \eta^{h,h'}_{\mathcal{M},\hat{\pi}^*}(s',a'|s,a)\right) C_k^{h'}(s',a')\Big| \\
\leq &2H \max_{s,a,h} C_k^h(s,a)
\end{aligned}$$

where (a) uses Lemma 6 and (b) uses Corollary B.9.

For the second condition of Definition 2, it follows similarly that:

$$\inf_{r\in\mathcal{R}_{\mathfrak{B}}} \sup_{\pi^*\in\Pi^*_{\mathcal{M}\cup r}} \max_{s,a,h}(Q^{\pi^*,h}_{\mathcal{M}\cup r}(s,a) - Q^{\hat{\pi}^*,h}_{\mathcal{M}\cup r}(s,a)) \leq 2H \max_{s,a,h} C_k^h(s,a)$$

Hence, if $H \max_{s,a,h} C_k^h(s,a) \leq \epsilon/2$, both conditions of Definition 2 are satisfied. □

Theorem B.11 (Sample Complexity of Uniform Sampling IRL)**.** *With probability at least $1-\delta$, Algorithm 2 stops at iteration τ fulfilling Definition 2 with a number of samples upper bounded by:*

$$n \leq \tilde{\mathcal{O}}\left(\frac{H^5 R_{\max}^2 SA}{\epsilon^2}\right)$$

Proof. First, note

$$H \max_{s,a,h} C_k^h(s,a) = H^2 R_{\max} \max_{s,a,h}\left(2\sqrt{\frac{2\ell_k^h(s,a)}{n_k^{h+}(s,a)}}\right)$$

After τ iterations, we have collected $\tau \cdot n_{\max}$ samples and for each s,a,h, we have: $n_\tau^{h+}(s,a) \geq \frac{\tau n_{\max}}{SAH} \geq 1$

To terminate at iteration τ, we need to have for all s,a,h:

$$2H^2 R_{\max}\sqrt{\frac{2\ell_\tau^h(s,a)}{n_\tau^h(s,a)}} \leq \frac{\epsilon}{2}$$

which implies

$$n_\tau^h(s,a) \geq \frac{32H^4 R_{\max}^2 \ell_\tau^h(s,a)}{\epsilon^2}$$

By using Lemma B.8 by Metelli et al. (2021), we can conclude that the number of samples necessary to ensure accuracy ε is:

$$n \leq \tilde{\mathcal{O}}\left(\frac{H^5 R_{\max}^2 SA}{\epsilon^2}\right)$$

□

Corollary B.12. *If the true reward function does not depend on the timestep h, i.e., $r_h(s,a) = r(s,a)$, then we can modify Algorithm 2 to only need $n \leq \tilde{\mathcal{O}}\left(\frac{H^4 R_{\max}^2 SA}{\epsilon^2}\right)$ samples.*

Proof. If we know that the reward function does not depend on h we can choose $C_k(s,a) = \min_h C_k^h(s,a)$ as a confidence interval of the reward. Consequently, we can sample all states for a fixed h.

We still need for all s,a:

$$2H^2 R_{\max}\sqrt{\frac{2\ell_\tau^h(s,a)}{n_\tau^h(s,a)}} \leq \frac{\epsilon}{2} \;\Rightarrow\; n_\tau^h(s,a) \geq \frac{32H^4 R_{\max}^2 \ell_\tau^h(s,a)}{\epsilon^2}$$

Again, we use Lemma B.8 by Metelli et al. (2021), but we can eliminate one sum over H, ending up with:

$$n \leq \tilde{\mathcal{O}}\left(\frac{H^4 R_{\max}^2 SA}{\epsilon^2}\right)$$

□

B.4 Sample Complexity of AceIRL in Unknown Environments (Problem Independent)

We are now ready to analyze the sample complexity of AceIRL (Algorithm 1). We first consider the simple version of the algorithm: AceIRL Greedy. Then, we consider the full version of the algorithm after introducing a few additional lemma about the policy confidence set. We start by defining the error upper bound and deriving two lemmas that will help us to show that it is indeed an upper bound on the error we want to reduce.

Definition B.13. *We define recursively:*

$$E_k^H(s,a) = 0; \quad E_k^h(s,a) = \min\Big((H-h)R_{\max}, C_k^h(s,a) + \sum_{s'} \widehat{P}(s'|s,a) \max_{a' \in \mathcal{A}} E_k^{h+1}(s',a')\Big)$$

where $\widehat{P}$ is the estimated transition model of the environment.

The first lemma shows that the error upper bound can upper bound the error due to estimating the transition model.

Lemma B.14. *Under the good event $\mathscr{E}$, for all policies π and reward functions r and all s, a, h:*

$$|Q^{\pi,h}_{\widehat{\mathcal{M}}\cup r}(s,a) - Q^{\pi,h}_{\mathcal{M}\cup r}(s,a)| \le E_k^h(s,a)$$

Proof.

$$\begin{aligned}
&|Q^{\pi,h}_{\widehat{\mathcal{M}}\cup r}(s,a) - Q^{\pi,h}_{\mathcal{M}\cup r}(s,a)| = \Big|\sum_{s'} \widehat{P}(s'|s,a) \sum_{a'} \pi(a'|s') Q^{\pi,h+1}_{\widehat{\mathcal{M}}\cup r}(s',a') \\
&- \sum_{s'} P(s'|s,a) \sum_{a'} \pi(a'|s') Q^{\pi,h+1}_{\mathcal{M}\cup r}(s',a') \pm \sum_{s'} \widehat{P}(s'|s,a) \sum_{a'} \pi(a'|s') Q^{\pi,h+1}_{\mathcal{M}\cup r}(s',a')\Big| \\
&\le \Big|\sum_{s'} \big(\widehat{P}(s'|s,a) - P(s'|s,a)\big) \sum_{a'} \pi(a'|s') Q^{\pi,h+1}_{\mathcal{M}\cup r}(s',a')\Big| \\
&+ \sum_{s'} \widehat{P}(s'|s,a) \sum_{a'} \pi(a'|s') \Big|Q^{\pi,h+1}_{\widehat{\mathcal{M}}\cup r}(s,a) - Q^{\pi,h+1}_{\mathcal{M}\cup r}(s,a)\Big| \\
&\le C_k^h(s,a) + \sum_{s'} \widehat{P}(s'|s,a) \sum_{a'} \pi(a'|s') \Big|Q^{\pi,h+1}_{\widehat{\mathcal{M}}\cup r}(s,a) - Q^{\pi,h+1}_{\mathcal{M}\cup r}(s,a)\Big|
\end{aligned}$$

For $h = H$ the result holds trivially. Now assuming it holds for $h+1$, we consider step h:

$$\begin{aligned}
&|Q^{\pi,h}_{\widehat{\mathcal{M}}\cup r}(s,a) - Q^{\pi,h}_{\mathcal{M}\cup r}(s,a)| \le C_k^h(s,a) + \sum_{s'} \widehat{P}(s'|s,a) \sum_{a'} \pi(a'|s') \Big|Q^{\pi,h+1}_{\widehat{\mathcal{M}}\cup r}(s,a) - Q^{\pi,h+1}_{\mathcal{M}\cup r}(s,a)\Big| \\
&\le C_k^h(s,a) + \sum_{s'} \widehat{P}(s'|s,a) \max_{a'} \Big|Q^{\pi,h+1}_{\widehat{\mathcal{M}}\cup r}(s,a) - Q^{\pi,h+1}_{\mathcal{M}\cup r}(s,a)\Big| \\
&\le C_k^h(s,a) + \sum_{s'} \widehat{P}(s'|s,a) \max_{a'} E_k^{h+1}(s',a') = E_k^h(s,a)
\end{aligned}$$

□

The next lemma shows that the error upper bound can also upper bound the error in estimating the reward function, which is due to estimating the transition model and the expert policy.

Lemma B.15. *Under the good event $\mathscr{E}$, for all reward function r, all policies π, and all $s, a \in \mathcal{S} \times \mathcal{A}$:*

$$|Q^{\pi,h}_{\widehat{\mathcal{M}}\cup \hat{r}}(s,a) - Q^{\pi,h}_{\widehat{\mathcal{M}}\cup r}(s,a)| \le E_k^h(s,a)$$

Proof. For $h = H$ the result holds trivially. Now assuming it holds for $h+1$, we consider step h:

$$\begin{aligned}
&|Q^{\pi,h}_{\widehat{\mathcal{M}}\cup \hat{r}}(s,a) - Q^{\pi,h}_{\widehat{\mathcal{M}}\cup r}(s,a)| \\
\le &|\hat{r}(s,a) - r(s,a)| + \sum_{s'} \widehat{P}(s'|s,a) \sum_{a'} \pi(a'|s') |Q^{\pi,h+1}_{\widehat{\mathcal{M}}\cup \hat{r}}(s',a') - Q^{\pi,h+1}_{\widehat{\mathcal{M}}\cup r}(s',a')| \\
\le &|\hat{r}(s,a) - r(s,a)| + \sum_{s'} \widehat{P}(s'|s,a) \max_{a'} |Q^{\pi,h+1}_{\widehat{\mathcal{M}}\cup \hat{r}}(s',a') - Q^{\pi,h+1}_{\widehat{\mathcal{M}}\cup r}(s',a')| \\
\le &|\hat{r}(s,a) - r(s,a)| + \sum_{s'} \widehat{P}(s'|s,a) \max_{a'} E_k^{h+1}(s',a') = E_k^h(s,a)
\end{aligned}$$

□

We can now combine the previous two lemmas to show that E is indeed an upper bound on the error we want to reduce. This implies correctness of AceIRL Greedy, which the following lemma formalizes.

Lemma B.16 (Correctness of AceIRL Greedy)**.** *If AceIRL Greedy stops in episode k, after sampling n samples, i.e., $E_k^0(s_0, \pi_{k+1}(s_0)) \leq \frac{\epsilon}{4}$, then it fulfills Definition 2.*

Proof. Let us define the error

$$e_k^h(s,a) \coloneqq |Q^{\pi^*,h}_{\mathcal{M}\cup r}(s,a) - Q^{\hat{\pi}^*,h}_{\mathcal{M}\cup r}(s,a)|$$

where π^* is the true optimal policy in $\mathcal{M} \cup r$, and $\hat{\pi}^*$ is the optimal policy in $\widehat{\mathcal{M}} \cup \hat{r}$, i.e., in the estimated MDP using the inferred reward function. Then,

$$\begin{aligned}
e_k^h(s,a) &= |Q^{\pi^*,h}_{\mathcal{M}\cup r}(s,a) - Q^{\hat{\pi}^*,h}_{\mathcal{M}\cup r}(s,a) \pm Q^{\pi^*,h}_{\widehat{\mathcal{M}}\cup r}(s,a) \pm Q^{\hat{\pi}^*,h}_{\widehat{\mathcal{M}}\cup r}(s,a)| \\
&\leq \underbrace{|Q^{\pi^*,h}_{\mathcal{M}\cup r}(s,a) - Q^{\pi^*,h}_{\widehat{\mathcal{M}}\cup r}(s,a)|}_{\leq E_k^h(s,a)} + |Q^{\pi^*,h}_{\widehat{\mathcal{M}}\cup r}(s,a) - Q^{\hat{\pi}^*,h}_{\widehat{\mathcal{M}}\cup r}(s,a)| + \underbrace{|Q^{\hat{\pi}^*,h}_{\widehat{\mathcal{M}}\cup r}(s,a) - Q^{\hat{\pi}^*,h}_{\mathcal{M}\cup r}(s,a)|}_{\leq E_k^h(s,a)} \\
&\leq 2E_k^h(s,a) + |Q^{\pi^*,h}_{\widehat{\mathcal{M}}\cup r}(s,a) - Q^{\hat{\pi}^*,h}_{\widehat{\mathcal{M}}\cup r}(s,a)|
\end{aligned}$$

where, we used Lemma B.14.

Let us consider the remaining term $|Q^{\pi^*,h}_{\widehat{\mathcal{M}}\cup r}(s,a) - Q^{\hat{\pi}^*,h}_{\widehat{\mathcal{M}}\cup r}(s,a)|$ in two steps. First, we have:

$$\begin{aligned}
Q^{\pi^*,h}_{\widehat{\mathcal{M}}\cup r}(s,a) - Q^{\hat{\pi}^*,h}_{\widehat{\mathcal{M}}\cup r}(s,a) \leq \underbrace{Q^{\pi^*,h}_{\widehat{\mathcal{M}}\cup r}(s,a) - Q^{\pi^*,h}_{\widehat{\mathcal{M}}\cup \hat{r}}(s,a)}_{\leq E_k^h(s,a)} + \underbrace{Q^{\pi^*,h}_{\widehat{\mathcal{M}}\cup \hat{r}}(s,a) - Q^{\hat{\pi}^*,h}_{\widehat{\mathcal{M}}\cup \hat{r}}(s,a)}_{\leq 0} + \\
+ \underbrace{Q^{\hat{\pi}^*,h}_{\widehat{\mathcal{M}}\cup \hat{r}}(s,a) - Q^{\hat{\pi}^*,h}_{\widehat{\mathcal{M}}\cup r}(s,a)}_{\leq E_k^h(s,a)} \leq 2E_k^h(s,a),
\end{aligned}$$

where we used Lemma B.15 and the fact that $\hat{\pi}^*$ is optimal in the MDP $\widehat{\mathcal{M}} \cup \hat{r}$. Second, we have:

$$\begin{aligned}
Q^{\hat{\pi}^*,h}_{\widehat{\mathcal{M}}\cup r}(s,a) - Q^{\pi^*,h}_{\widehat{\mathcal{M}}\cup r}(s,a) \leq \underbrace{Q^{\hat{\pi}^*,h}_{\widehat{\mathcal{M}}\cup r}(s,a) - Q^{\hat{\pi}^*,h}_{\mathcal{M}\cup r}(s,a)}_{\leq E_k^h(s,a)} + \underbrace{Q^{\hat{\pi}^*,h}_{\mathcal{M}\cup r}(s,a) - Q^{\pi^*,h}_{\mathcal{M}\cup r}(s,a)}_{\leq 0} + \\
+ \underbrace{Q^{\pi^*,h}_{\mathcal{M}\cup r}(s,a) - Q^{\pi^*,h}_{\widehat{\mathcal{M}}\cup r}(s,a)}_{\leq E_k^h(s,a)} \leq 2E_k^h(s,a),
\end{aligned}$$

where we used Lemma B.14 and the fact that π^* is optimal in the MDP $\mathcal{M} \cup r$. Overall, we find that

$$|Q^{\pi^*,h}_{\widehat{\mathcal{M}}\cup r}(s,a) - Q^{\hat{\pi}^*,h}_{\widehat{\mathcal{M}}\cup r}(s,a)| \leq 2E_k^h(s,a),$$

and consequently,

$$e_k^h(s,a) \leq 4E_k^h(s,a).$$

Note that, $E_k^h(s,a)$ only sums positive terms, hence:

$$\max_{s,a,h} E_k^h(s,a) \leq \max_a E_k^0(s_0,a) = E_k^0(s_0, \pi_{k+1}(s_0))$$

Hence, if $E_k^0(s_0, \pi_{k+1}(s_0)) \leq \frac{\epsilon}{4}$, we have for all $s, a, h \in \mathcal{S} \times \mathcal{A} \times [H]$:

$$e_k^h(s,a) \leq \epsilon$$

which implies correctness according to Definition 2. □

Next, we will analyze the sample complexity of AceIRL Greedy. Let us first define pseudo-counts that will be crucial to deal with the uncertainty of the transition dynamics in our analysis. This is similar to the analysis of UCRL for reward-free exploration by Kaufmann et al. (2021).

Definition B.17. *We define the* pseudo-counts *of visiting a specific state action pair at timestep h within the first k iterations as*

$$\bar{n}_k^h(s,a) \coloneqq \sum_{i=1}^{k} \eta^{0,h}_{\mathcal{M},\pi_i}(s,a|s_0),$$

where π_i is the exploration policy in episode i.

The following lemma allows us to introduce the pseudo-counts when considering the contraction of the reward confidence intervals.

Lemma B.18. *With probability at least $1 - \frac{\delta}{2}$ for all $s, a, h, k \in \mathcal{S} \times \mathcal{A} \times [H] \times \mathbb{N}^+$, we have:*

$$\min\left(\frac{2\ell_k^h(s,a)}{n_k^h(s,a)}, 1\right) \leq \frac{8\bar{\ell}_k^h(s,a)}{\max\left(\bar{n}_k^h(s,a), 1\right)}$$

where $\bar{\ell}_k^h(s,a) = \log\left(24SAH(\bar{n}_k^h(s,a))^2/\delta\right)$.

Proof. This result adapts Lemma 7 by Kaufmann et al. (2021) to our setting.

By Lemma 10 in Kaufmann et al. (2021), we have with probability at least $1 - \frac{\delta}{2}$:

$$n_k^h(s,a) \geq \frac{1}{2}\bar{n}_k^h(s,a) - \beta_{\text{cnt}}(\delta),$$

where $\beta_{\text{cnt}}(\delta) = \log(2SAH/\delta)$.

We distinguish two cases. First let $\beta_{\text{cnt}}(\delta) \leq \frac{1}{4}\bar{n}_k^h(s,a)$. Then $n_k^h(s,a) \geq \frac{1}{4}\bar{n}_k^h(s,a)$, and

$$\begin{aligned}
&\min\left(\frac{2\ell_k^h(s,a)}{n_k^h(s,a)}, 1\right) \leq \frac{2\ell_k^h(s,a)}{\max(n_k^h(s,a), 1)} = \frac{2\log(24SAH(n_k^h(s,a))^2/\delta)}{\max(n_k^h(s,a), 1)} \\
&\leq \frac{2\log(24SAH(\bar{n}_k^h(s,a)/4)^2/\delta)}{(\bar{n}_k^h(s,a)/4)} \leq \frac{8\bar{\ell}_k^h(s,a)}{\max(\bar{n}_k^h(s,a), 1)}
\end{aligned}$$

where we use that $\log(24SAHx^2/\delta)/x$ is non-increasing for $x > 1$, and $\log(24SAHx^2/\delta)$ is non-decreasing and $\beta_{\text{cnt}}(\delta) \geq 1$.

Now consider let $\beta_{\text{cnt}}(\delta) > \frac{1}{4}\bar{n}_k^h(s,a)$. Then,

$$\min\left(\frac{2\ell_k^h(s,a)}{n_k^h(s,a)}, 1\right) \leq 1 < 4\frac{\beta_{\text{cnt}}(\delta)}{\max(\bar{n}_k^h(s,a), 1)} \leq \frac{4\bar{\ell}_k^h(s,a)}{\max(\bar{n}_k^h(s,a), 1)}$$

where we used that $\ell_k^h(s,a) = \log\left(24SAH(n_k^h(s,a))^2/\delta\right) = \beta_{\text{cnt}}(\delta) + \log\left(6n_k^h(s,a))^2\right) \geq \beta_{\text{cnt}}(\delta)$. □

The final lemma we need shows relates the error upper bound which is defined using our estimated transition model to a similar quantity defined using the (unknown) real transitions.

Lemma B.19. *Under the good event $\mathscr{E}$, we have for any s, a, h :*

$$E_k^h(s,a) \leq 2C_k^h(s,a) + \sum_{s'} P(s'|s,a) \max_{a'} E_k^{h+1}(s', a')$$

where P is the true transition model that we do not know.

Proof. First note that $E_k^h(s,a) \leq H$ by definition. Now, consider:

$$\begin{aligned}
E_k^h(s,a) &\leq C_k^h(s,a) + \sum_{s'} \widehat{P}(s'|s,a) \max_{a'} E_k^{h+1}(s', a') \\
&= C_k^h(s,a) + \sum_{s'} (\widehat{P}(s'|s,a) - P(s'|s,a) + P(s'|s,a)) \max_{a'} E_k^{h+1}(s', a') \\
&= C_k^h(s,a) + \sum_{s'} \underbrace{(\widehat{P}(s'|s,a) - P(s'|s,a)) \max_{a'} E_k^{h+1}(s', a')}_{\leq C_k^h(s,a)} + \sum_{s'} P(s'|s,a)) \max_{a'} E_k^{h+1}(s', a') \\
&\leq 2C_k^h(s,a) + \sum_{s'} P(s'|s,a) \max_{a'} E_k^{h+1}(s', a')
\end{aligned}$$

where we used the good event and the fact that C_k^h can only shrink over episodes. □

Finally, we can analyze the sample complexity of AceIRL Greedy.

Theorem B.20 (AceIRL Greedy Sample Complexity (problem independent))**.** *AceIRL Greedy terminates with an (ϵ, δ, n)-correct solution, with*

$$n \leq \tilde{\mathcal{O}}\left(\frac{H^5 R_{\max}^2 SA}{\epsilon^2}\right).$$

Proof. Lemma B.16 shows that if AceIRL Greedy terminates, then it returns a (ϵ, δ, n)-correct solution. So, we need to show that it terminates within τ iterations and bound τ.

Let us consider the average error, defined by

$$\begin{aligned}
q_k^h &\coloneqq \sum_{s,a} \eta^{0,h}_{\mathcal{M},\pi_{k+1}}(s,a|s_0) E_k^h(s,a) \\
&\overset{(a)}{\leq} \sum_{s,a} \eta^{0,h}_{\mathcal{M},\pi_{k+1}}(s,a|s_0)\big(2C_k^h(s,a) + \sum_{s'} P(s'|s,a) \max_{a'} E_k^{h+1}(s',a')\big) \\
&= \sum_{s,a} \eta^{0,h}_{\mathcal{M},\pi_{k+1}}(s,a|s_0)\big(2C_k^h(s,a) + \sum_{s'} P(s'|s,a) \sum_{a'} \pi_{k+1}(a'|s') E_k^{h+1}(s',a')\big) \\
&= 2\sum_{s,a} \eta^{0,h}_{\mathcal{M},\pi_{k+1}}(s,a|s_0) C_k^h(s,a) + q_k^{h+1}
\end{aligned}$$

where we used Lemma B.19 in step (a). Unrolling the recursion, results in:

$$q_k^h \leq 2\sum_{h'=h}^{H} \sum_{s,a} \eta^{0,h'}_{\mathcal{M},\pi_{k+1}}(s,a|s_0) C_k^{h'}(s,a)$$

If the algorithm terminates at τ, we have for each $k < \tau$, and $s,a,h \in \mathcal{S} \times \mathcal{A} \times [H]$: $\epsilon < 4E_k^0(s_0,\pi_{k+1}(s_0))$. We have $q_k^0 = E_k^0(s_0,\pi_{k+1}(s_0))$; therefore, as long we haven't stopped, we have $\epsilon \leq 4q_k^0$. Writing out this inequality, yields:

$$\epsilon \leq 4q_k^0 \leq 8\sum_{h=0}^{H}\sum_{s,a} \eta^{0,h}_{\mathcal{M},\pi_{k+1}}(s,a|s_0) C_k^h(s,a) \leq 4HR_{\max}\sum_{h=0}^{H}\sum_{s,a} \eta^{0,h}_{\mathcal{M},\pi_{k+1}}(s,a|s_0)\sqrt{\frac{8\log(12SAH(n_k^h(s,a))^2/\delta)}{\max(n_k^h(s,a),1)}}$$

Using Lemma B.18, we can relate this to the pseudo-counts

$$\begin{aligned}
\epsilon &< 4HR_{\max}\sum_{h=0}^{H}\sum_{s,a} \eta^{0,h}_{\mathcal{M},\pi_{k+1}}(s,a|s_0)\sqrt{\frac{8\log(12SAH(\bar{n}_k^h(s,a))^2/\delta)}{\max(\bar{n}_k^h(s,a),1)}} \\
&\leq 4HR_{\max}\sum_{h=0}^{H}\sum_{s,a} \eta^{0,h}_{\mathcal{M},\pi_{k+1}}(s,a|s_0)\sqrt{\frac{8\log(12SAHk^2/\delta)}{\max(\bar{n}_k^h(s,a),1)}}
\end{aligned}$$

Summing the inequality over $k = 0,\dots T$ with $T < \tau$, we obtain

$$\begin{aligned}
\epsilon(T+1) &\leq 4HR_{\max}\sqrt{8\log(12SAHT^2/\delta)}\sum_{h=0}^{H}\sum_{s,a}\sum_{k=1}^{T} \eta^{0,h}_{\mathcal{M},\pi_{k+1}}(s,a|s_0)\frac{1}{\sqrt{\max(\bar{n}_k^h(s,a),1)}} \\
&= 4HR_{\max}\sqrt{8\log(12SAHT^2/\delta)}\sum_{h=0}^{H}\sum_{s,a}\sum_{k=1}^{T} \frac{\bar{n}_h^{k+1}(s,a) - \bar{n}_h^k(s,a)}{\sqrt{\max(\bar{n}_k^h(s,a),1)}}
\end{aligned}$$

where we used the definition of the pseudo-counts in the last equality. Using Lemma 19 by Jaksch et al. (2010), we can further bound the sum in k:

$$\begin{aligned}
\epsilon(T+1) &= 4HR_{\max}\sqrt{8\log(12SAHT^2/\delta)}\sum_{h=0}^{H}\sum_{s,a}\sqrt{\bar{n}_h^{T+1}(s,a)} \\
&\leq 4HR_{\max}\sqrt{8\log(12SAHT^2/\delta)}\sqrt{SA}\sum_{h=0}^{H}\sqrt{\sum_{s,a}\bar{n}_h^{T+1}(s,a)} \\
&= 4H^2R_{\max}\sqrt{8\log(12SAHT^2/\delta)}\sqrt{SA}\sqrt{T+1}
\end{aligned}$$

It follows that

$$\begin{aligned}
\epsilon\sqrt{T+1} &\leq 4H^2R_{\max}\sqrt{8SA\log(12SAHT^2/\delta)} \\
\epsilon^2\tau &\leq 128H^4R_{\max}^2 SA\log(12SAH(\tau-1)^2/\delta)
\end{aligned}$$

setting $\tau = T+1$.

For large enough τ, this inequality cannot hold because $\sqrt{T+1}$ on the l.h.s grows faster than $\log(\tau)$ on the r.h.s. Hence, the stopping time τ is finite. Further, we can apply Lemma 15 by Kaufmann et al. (2021), and follow that

$$\tau \leq \tilde{\mathcal{O}}\left(\frac{H^4R_{\max}^2 SA}{\epsilon^2}\right)$$

If we observe H samples in each iteration, i.e., $N_E = 1$, we get a sample complexity of

$$n \leq \tilde{\mathcal{O}}\left(\frac{H^5 R_{\max}^2 SA}{\epsilon^2}\right)$$

□

B.5 Sample Complexity of AceIRL in Unknown Environments (Problem Dependent)

For the problem dependent analysis, we will need this additional lemma also used by Kakade and Langford (2002).

Lemma B.21 (Lemma 6.1 by Kakade and Langford (2002))**.** *For any policy π:*

$$V_{\mathcal{M}\cup r}^{\pi^*,h}(s) - V_{\mathcal{M}\cup r}^{\pi,h}(s) = -\sum_{s',a'}\sum_{h'=h}^{H} \eta_{\mathcal{M},\pi}^{h,h'}(s',a';s) A_{\mathcal{M}\cup r}^{*,h'}(s',a')$$

Proof.

$$\begin{aligned}
&V_{\mathcal{M}\cup r}^{*,h}(s) - V_{\mathcal{M}\cup r}^{\pi,h}(s) \\
=&\sum_a \pi_h^*(a|s)\left(r_h(s,a) + \sum_{s'} P(s'|s,a) V_{\mathcal{M}\cup r}^{*,h+1}(s')\right) \\
&- \sum_a \pi_h(a|s)\left(r_h(s,a) + \sum_{s'} P(s'|s,a) V_{\mathcal{M}\cup r}^{\pi,h+1}(s')\right) \pm \sum_{a,s'} \pi_h(a|s) P(s'|s,a) V_{\mathcal{M}\cup r}^{*,h+1}(s') \\
=&\sum_a (\pi_h^*(a|s) - \pi_h(a|s)) r(s,a) + \sum_{a,s'} (\pi_h^*(a|s) - \pi_h(a|s)) P(s'|s,a) V_{\mathcal{M}\cup r}^{*,h+1}(s') \\
&+ \sum_{a,s'} \pi_h(a|s) P(s'|s,a)(V_{\mathcal{M}\cup r}^{*,h+1}(s) - V_{\mathcal{M}\cup r}^{\pi,h+1}(s)) \\
=&-\sum_a \pi(a|s) A_{\mathcal{M}\cup r}^{*,h}(s,a) + \sum_{a,s'} \pi_h(a|s) P(s'|s,a)(V_{\mathcal{M}\cup r}^{*,h+1}(s) - V_{\mathcal{M}\cup r}^{\pi,h+1}(s))
\end{aligned}$$

Unrolling the recursion yields the result. □

We can now start with the analysis. First, we define the policy confidence set, and show that it indeed contains the relevant policies under the good event.

Definition B.22. *We define the* policy confidence set *as*

$$\hat{\Pi}_k = \{\pi | V_{\widehat{\mathcal{M}}\cup\hat{r}}^{*,}(s_0) - V_{\widehat{\mathcal{M}}\cup\hat{r}}^{\pi,}(s_0) \leq 10\epsilon_k\}$$

where $\hat{r} = \mathscr{A}(\mathcal{R}_{\hat{\mathfrak{B}}})$ is the reward estimated using an IRL algorithm $\mathscr{A}$. We choose ϵ_k recursively by solving the optimization problem

$$\epsilon_k = \max_{\pi\in\hat{\Pi}_{k-1}} \sum_{h=0}^{H}\sum_{s',a'} \eta_{\widehat{\mathcal{M}},\pi}^{0,h}(s',a';s_0) C_k^h(s',a')$$

starting with $\epsilon_0 = \frac{1}{10}H$.

The following lemma will help us to deal with uncertainty about the transition dynamics.

Lemma B.23. *Under the good event $\mathscr{E}$, if $\pi \in \hat{\Pi}_k$, then:*

$$\begin{aligned}
|V_{\widehat{\mathcal{M}}\cup\hat{r}}^{\pi,h}(s) - V_{\mathcal{M}\cup\hat{r}}^{\pi,h}(s)| &\leq \epsilon_k \\
|V_{\mathcal{M}\cup\hat{r}}^{*,h}(s) - V_{\widehat{\mathcal{M}}\cup\hat{r}}^{*,h}(s)| &\leq \epsilon_k
\end{aligned}$$

Proof. First by Lemma B.4:

$$\begin{aligned}
|V_{\widehat{\mathcal{M}}\cup r}^{\pi,h}(s) - V_{\mathcal{M}\cup r}^{\pi,h}(s)| &\leq \sum_{h'=h}^{H}\sum_{s',a',s''} \eta_{\widehat{\mathcal{M}},\pi}^{h,h'}(s';s)\pi_{h'}(a'|s')|\widehat{P}(s''|s',a') - P(s''|s',a')| V_{\mathcal{M}\cup r}^{\pi,h'+1}(s'') \\
&\leq \sum_{h'=h}^{H}\sum_{s',a'} \eta_{\widehat{\mathcal{M}},\pi}^{h,h'}(s';s)\pi_{h'}(a'|s') C_k(s',a') \leq \epsilon_k
\end{aligned}$$

Then, by Lemma B.5:

$$V^{*,h}_{\mathcal{M}\cup r}(s) - V^{*,h}_{\widehat{\mathcal{M}}\cup r}(s) \le \sum_{h'=h}\sum_{s',a',s''} \eta^{h,h'}_{\widehat{\mathcal{M}},\pi^*}(s';s)\pi^*_{h'}(a'|s')(P(s''|s',a') - \widehat{P}(s''|s',a'))V^{*,h}_{\mathcal{M}\cup r}(s'')$$
$$\le \sum_{h'=h}\sum_{s',a'} \eta^{h,h'}_{\widehat{\mathcal{M}},\pi^*}(s';s)\pi^*_{h'}(a'|s')C_k(s',a') \le \epsilon_k$$

And, similarly

$$V^{*,h}_{\widehat{\mathcal{M}}\cup r}(s) - V^{*,h}_{\mathcal{M}\cup r}(s) \le \sum_{h'=h}\sum_{s',a',s''} \eta^{h,h'}_{\widehat{\mathcal{M}},\hat{\pi}^*}(s';s)\hat{\pi}^*_{h'}(a'|s')(\widehat{P}(s''|s',a') - P(s''|s',a'))V^{*,h}_{\widehat{\mathcal{M}}\cup r}(s'')$$
$$\le \sum_{h'=h}\sum_{s',a'} \eta^{h,h'}_{\widehat{\mathcal{M}},\hat{\pi}^*}(s';s)\hat{\pi}^*_{h'}(a'|s')C_k(s',a') \le \epsilon_k$$

□

Now we show that the relevant policies are always in the policy confidence set, conditioned on the good event.

Lemma B.24. *Conditioned the good event $\mathscr{E}$, if $\pi^*, \hat{\pi}^* \in \hat{\Pi}_{k-1}$, then $\pi^* \in \hat{\Pi}_k$.*

Proof. Let $r \in \mathcal{R}_{\mathfrak{B}}$. Then

$$V^{*,h}_{\widehat{\mathcal{M}}\cup\hat{r}_k}(s) - V^{\pi^*,h}_{\widehat{\mathcal{M}}\cup\hat{r}_k}(s) = V^{*,h}_{\widehat{\mathcal{M}}\cup\hat{r}_k}(s) - V^{*,h}_{\widehat{\mathcal{M}}\cup r}(s) + V^{*,h}_{\widehat{\mathcal{M}}\cup r}(s) - V^{\pi^*,h}_{\widehat{\mathcal{M}}\cup\hat{r}_k}(s)$$
$$\overset{(a)}{\le} \sum_{h'=h}^{H}\sum_{s',a'} \eta^{h,h'}_{\widehat{\mathcal{M}},\pi^*}(s',a'|s)C^{h'}_k(s',a') + \sum_{h'=h}^{H}\sum_{s',a'} \eta^{h,h'}_{\widehat{\mathcal{M}},\pi^*}(s',a'|s)C^{h'}_k(s',a') \overset{(b)}{\le} 2\epsilon_k$$

where (a) uses Lemma B.2, Lemma B.3 and Corollary B.9, (b) uses that $\pi^* \in \hat{\Pi}_{k-1}$ and the definition of ϵ_k. Hence,

$$\max_s \left(V^{*,h}_{\widehat{\mathcal{M}}\cup\hat{r}_k}(s) - V^{\pi^*,h}_{\widehat{\mathcal{M}}\cup\hat{r}_k}(s)\right) \le 2\epsilon_k \le 10\epsilon_k$$

and therefore $\pi^* \in \hat{\Pi}_k$. □

Lemma B.25. *Conditioned on the good event $\mathscr{E}$, for every policy π and episodes $k' > k$, there exists $\hat{r}_{k'} \in \mathcal{R}_{\hat{\mathfrak{B}}_{k'}}$, such that:*

$$\max_s \left(V^{\pi,h}_{\mathcal{M}\cup\hat{r}_{k'}}(s) - V^{\pi,h}_{\mathcal{M}\cup\hat{r}_k}(s)\right) \le 4\epsilon_k$$

Proof. Similarly to the proof of the previous lemma, we have

$$V^{\pi,h}_{\widehat{\mathcal{M}}\cup\hat{r}_{k'}}(s) - V^{\pi,h}_{\widehat{\mathcal{M}}\cup\hat{r}_k}(s) = V^{\pi,h}_{\widehat{\mathcal{M}}\cup\hat{r}_{k'}}(s) - V^{\pi,h}_{\widehat{\mathcal{M}}\cup r}(s) + V^{\pi,h}_{\widehat{\mathcal{M}}\cup r}(s) - V^{\pi,h}_{\widehat{\mathcal{M}}\cup\hat{r}_k}(s)$$
$$\le \sum_{h'=h}^{H}\sum_{s',a'} \eta^{h,h'}_{\widehat{\mathcal{M}},\pi}(s',a'|s)C^{h'}_{k'}(s',a') + \sum_{h'=h}^{H}\sum_{s',a'} \eta^{h,h'}_{\widehat{\mathcal{M}},\pi}(s',a'|s)C^{h'}_k(s',a') \le 2\epsilon_k$$

where we use that the confidence intervals are shrinking with increasing episode number, i.e., $\epsilon_{k'} \le \epsilon_k$.

By combining this with Lemma B.23, we get the result:

$$\max_s \left(V^{\pi,h}_{\mathcal{M}\cup\hat{r}_{k'}}(s) - V^{\pi,h}_{\mathcal{M}\cup\hat{r}_k}(s)\right)$$
$$= \max_s \Big(\underbrace{V^{\pi,h}_{\mathcal{M}\cup\hat{r}_{k'}}(s) - V^{\pi,h}_{\widehat{\mathcal{M}}\cup\hat{r}_{k'}}(s)}_{\le\epsilon_k} + \underbrace{V^{\pi,h}_{\widehat{\mathcal{M}}\cup\hat{r}_{k'}}(s) - V^{\pi,h}_{\widehat{\mathcal{M}}\cup\hat{r}_k}(s)}_{\le 2\epsilon_k} + \underbrace{V^{\pi,h}_{\widehat{\mathcal{M}}\cup\hat{r}_k}(s) - V^{\pi,h}_{\mathcal{M}\cup\hat{r}_k}(s)}_{\le\epsilon_k}\Big) \le 4\epsilon_k$$

□

Lemma B.26. *Under the good event $\mathscr{E}$, if $\hat{\pi}^*_k, \pi \in \hat{\Pi}_{k-1}$ and $\pi \notin \hat{\Pi}_k$, then the policy π is suboptimal for some reward $\hat{r}_{k'} \in \mathcal{R}_{\hat{\mathfrak{B}}_{k'}}$ for all $k' \ge k$.*

Proof. We can observe that

$$
\begin{aligned}
&V^{\pi,h}_{\mathcal{M}\cup\hat{r}_{k'}}(s_0) - V^{*,h}_{\mathcal{M}\cup\hat{r}_{k'}}(s_0) = V^{\pi,h}_{\mathcal{M}\cup\hat{r}_{k'}}(s_0) - V^{\hat{\pi}^*_k,h}_{\mathcal{M}\cup\hat{r}_{k'}}(s_0) \\
&= \underbrace{V^{\pi,h}_{\mathcal{M}\cup\hat{r}_{k'}}(s_0) - V^{\pi,h}_{\mathcal{M}\cup\hat{r}_k}(s_0)}_{\overset{(a)}{\leq} 4\epsilon_k} + \underbrace{V^{\pi,h}_{\mathcal{M}\cup\hat{r}_k}(s_0) - V^{\pi,h}_{\widehat{\mathcal{M}}\cup\hat{r}_k}(s_0)}_{\overset{(b)}{\leq} \epsilon_k} \\
&\quad + \underbrace{V^{\pi,h}_{\widehat{\mathcal{M}}\cup\hat{r}_k}(s_0) - V^{\hat{\pi}^*_k,h}_{\widehat{\mathcal{M}}\cup\hat{r}_k}(s_0)}_{\overset{(c)}{>} 10\epsilon_k} + \underbrace{V^{\hat{\pi}^*_k,h}_{\widehat{\mathcal{M}}\cup\hat{r}_k}(s_0) - V^{\hat{\pi}^*_k,h}_{\mathcal{M}\cup\hat{r}_k}(s_0)}_{\overset{(b)}{\leq} \epsilon_k} \\
&\quad + \underbrace{V^{\hat{\pi}^*_k,h}_{\mathcal{M}\cup\hat{r}_k}(s_0) - V^{\hat{\pi}^*_k,h}_{\mathcal{M}\cup\hat{r}_{k'}}(s_0)}_{\overset{(a)}{\leq} 4\epsilon_k} > 0
\end{aligned}
$$

where we applied (a) Lemma B.23, (b) Lemma B.25, and (c) the definition of $\hat{\Pi}_k$ and the fact that $\pi \notin \hat{\Pi}_k$. Consequently, π is suboptimal for at least some reward function $\hat{r}_{k'} \in \mathcal{R}_{\hat{\mathfrak{B}}_{k'}}$. □

Corollary B.27. *For $\epsilon_0 = \frac{H}{10}$, for every $k \geq 0$ it holds that both $\pi^*, \hat{\pi}^*_{k+1} \in \hat{\Pi}_k$.*

Proof. We show the statement by induction over k. For $k = 0$, we have $10\epsilon_0 = H$ and therefore $\hat{\Pi}_0$ contains all policies. Assume that for $k-1$ the statement holds, i.e., $\pi^*, \hat{\pi}^*_k \in \hat{\Pi}_{k-1}$, and consider k. By Lemma B.24, $\pi^* \in \hat{\Pi}_k$. Note, that $\hat{\pi}^*_{k+1} \in \hat{\Pi}_{k-1}$. Hence, by Lemma B.25, it follows that $\hat{\pi}^*_{k+1} \in \hat{\Pi}_k$ because it would be suboptimal otherwise which is a contradiction. □

The last result we need, is quantifying the size of the policy confidence set.

Lemma B.28. *Under the good event $\mathscr{E}$, let $\tilde{r} \in \operatorname{argmin}_{r\in\mathcal{R}_{\mathfrak{B}}} \max_{s,a}(r(s,a) - \hat{r}_k(s,a))$, where $\hat{r}_k = \mathscr{A}(\mathcal{R}_{\hat{\mathfrak{B}}_k})$. If $\pi \in \hat{\Pi}_k$, then $\max_s(V^{*,h}_{\widehat{\mathcal{M}}\cup\tilde{r}}(s) - V^{\pi,h}_{\widehat{\mathcal{M}}\cup\tilde{r}}(s)) \leq 12\epsilon_k$.*

Proof.

$$
V^{*,h}_{\widehat{\mathcal{M}}\cup\tilde{r}}(s) - V^{\pi,h}_{\widehat{\mathcal{M}}\cup\tilde{r}}(s) = \underbrace{V^{*,h}_{\widehat{\mathcal{M}}\cup\tilde{r}}(s) - V^{*,h}_{\widehat{\mathcal{M}}\cup\hat{r}_k}(s)}_{\leq\epsilon_k} + \underbrace{V^{*,h}_{\widehat{\mathcal{M}}\cup\hat{r}_k}(s) - V^{\pi,h}_{\widehat{\mathcal{M}}\cup\hat{r}_k}(s)}_{\leq 10\epsilon_k} + \underbrace{V^{\pi,h}_{\widehat{\mathcal{M}}\cup\hat{r}_k}(s) - V^{\pi,h}_{\widehat{\mathcal{M}}\cup\tilde{r}}(s)}_{\leq\epsilon_k} \epsilon_k \leq 14\epsilon_k
$$

□

Next, we define the error upper bound based on the policy confidence set.

Definition B.29. *Using $\hat{\Pi}_k$, we define recursively:*

$$
\begin{aligned}
&\hat{E}^H_k(s,a) = 0 \\
&\hat{E}^h_k(s,a) = \min\Big((H-h)R_{\max}, C^h_k(s,a) + \sum_{s'} \widehat{P}(s'|s,a) \max_{\pi\in\hat{\Pi}_{k-1}} \pi(a'|s')\hat{E}^{h+1}_k(s',a')\Big)
\end{aligned}
$$

where $\widehat{P}$ is the estimated transition model of the environment. In contrast to Definition B.13, the maximization is over policies in $\hat{\Pi}_k$ rather than all actions.

This definition allows us to derive results that are analogous to the problem independent case.

Lemma B.30. *Under the good event $\mathscr{E}$, for all policies $\pi \in \hat{\Pi}_k$ and reward functions r and all $s, a \in \mathcal{S}\times\mathcal{A}$:*

$$
|Q^{\pi,h}_{\widehat{\mathcal{M}}\cup r}(s,a) - Q^{\pi,h}_{\mathcal{M}\cup r}(s,a)| \leq \hat{E}^h_k(s,a)
$$

Proof. The proof is the same as for Lemma B.14, restricting the set of policies to $\hat{\Pi}_k$. □

Lemma B.31. *Under the good event $\mathscr{E}$, for all reward function r, all policies $\pi \in \hat{\Pi}_k$, and all $s, a \in \mathcal{S}\times\mathcal{A}$:*

$$
|Q^{\pi,h}_{\widehat{\mathcal{M}}\cup\hat{r}}(s,a) - Q^{\pi,h}_{\widehat{\mathcal{M}}\cup r}(s,a)| \leq \hat{E}^h_k(s,a)
$$

Proof. The proof is the same as for Lemma B.15, restricting the set of policies to $\hat{\Pi}_k$. □

Lemma B.32. *Under the good event $\mathscr{E}$, we have for any s, a, h :*

$$\hat{E}_k^h(s,a) \leq 2C_k^h(s,a) + \sum_{s'} P(s'|s,a) \max_{\pi \in \hat{\Pi}_{k-1}} \pi(a'|s') \hat{E}_k^{h+1}(s',a')$$

Proof. The proof is the same as for Lemma B.32. □

Finally, we can combine these results to analyze the algorithm's sample complexity.

Theorem 8. *[AceIRL Sample Complexity] AceIRL returns a (ϵ, δ, n)-correct solution with*

$$n \leq \tilde{\mathcal{O}}\left(\min\left[\frac{H^5 R_{\max}^2 SA}{\epsilon^2}, \frac{H^4 R_{\max}^2 SA \epsilon_{\tau-1}^2}{\min_{s,a,h}(A^{*,h}_{\mathcal{M}\cup r}(s,a))^2 \epsilon^2}\right]\right)$$

where $\epsilon_{\tau-1}$ depends on the choice of N_E, the number of episodes of exploration in each iteration. $A^{*,h}_{\mathcal{M}\cup r}(s,a)$ is the advantage function of $r \in \operatorname{argmin}_{r\in\mathcal{R}_{\mathfrak{B}}} \max_{h,s,a}(r_h(s,a) - \hat{r}_{k,h}(s,a))$, the reward function from the feasible set $\mathcal{R}_{\mathfrak{B}}$ closest to the estimated reward function $\hat{r}_k$.

Proof. First note that the analysis of Theorem B.20 still applies; so, in the worst case we get the same sample complexity. The key difference is that we no longer use the overall greedy policy w.r.t E_k^h, but restrict ourselves to policies in $\hat{\Pi}_k$.

Again, we consider the error

$$e_k^{\pi,h}(s,a) \coloneqq |Q^{\pi^*,h}_{\mathcal{M}\cup r}(s,a) - Q^{\hat{\pi}^*,h}_{\mathcal{M}\cup r}(s,a)|$$

where π^* is the true optimal policy in $\mathcal{M} \cup r$, and $\hat{\pi}^*$ is the optimal policy in $\widehat{\mathcal{M}} \cup \hat{r}$, i.e., in the estimated MDP using the inferred reward function.

Similar, to the proof of Lemma B.16, we can use Lemma B.30 and Lemma B.31 to show for all policies $\pi \in \hat{\Pi}_k^h$, that:

$$e_k^{\pi,h}(s,a) \leq 4\hat{E}_k^h(s,a)$$

which implies the correctness of the algorithm according to Corollary B.10 when stopping at

$$\hat{E}_k^0(s_0, \pi_{k+1}(s_0)) \leq \frac{\epsilon}{4} \tag{2}$$

Now, consider the following condition for all s, a, h:

$$C_k^h(s,a) \leq -A^{*,h}_{\mathcal{M}\cup\tilde{r}}(s,a)\frac{\epsilon}{48\epsilon_{k-1}}, \tag{3}$$

where $\tilde{r} \in \operatorname{argmin}_{r\in\mathcal{R}_{\mathfrak{B}}} \max_{h,s,a}(r_h(s,a) - \hat{r}_{k,h}(s,a))$. We will (a) show that when this condition holds the previous stopping condition also holds, and (b) analyze after how many iterations this condition will certainly hold. Together this will yield the result.

To show that Equation (3) implies Equation (2), we assume that Equation (3) holds. Then, we get by applying Lemma B.32 recursively:

$$\begin{aligned}
\hat{E}_k^0(s_0, \pi_{k+1}(s_0)) &\leq 2 \max_{\pi\in\hat{\Pi}_{k-1}} \max_a \sum_{h=0}^{H} \sum_{s',a'} \eta^{0,h}_{\mathcal{M},\pi}(s',a';s_0,a) C_k^h(s',a') \\
&\leq 2 \max_{\pi\in\hat{\Pi}_{k-1}} \max_a \sum_{h=0}^{H} \sum_{s',a'} \eta^{0,h}_{\mathcal{M},\pi}(s',a';s_0,a) \left(-A^{*,h}_{\mathcal{M}\cup\tilde{r}}(s',a')\frac{\epsilon}{48\epsilon_{k-1}}\right) \\
&\overset{(a)}{\leq} 2 \max_{\pi\in\hat{\Pi}_{k-1}} (V^{*,0}_{\mathcal{M}\cup r}(s_0) - V^{\pi,0}_{\mathcal{M}\cup r}(s_0))\frac{\epsilon}{48\epsilon_{k-1}} \overset{(b)}{\leq} \frac{\epsilon}{4}
\end{aligned}$$

where (a) uses Lemma B.21 and (b) uses Lemma B.28.

Next, we analyze after how many iterations Equation (3) holds, which will give a lower bound on the sample complexity result. The argument proceeds similar to the proof of Theorem B.20.

Before the algorithm terminates at τ, we have for all $k < \tau$:

$$\min_{s,a,h}(-A^{*,h}_{\widehat{\mathcal{M}}\cup\tilde{r}}(s,a))\frac{\epsilon}{48\epsilon_{k-1}} < \max_{s,a,h} C_k^h(s,a) \leq HR_{\max}\sqrt{\frac{2\ell_k^h(s,a)}{\max(N_k^h(s,a),)}}$$

Using similar argument to the proof of Theorem B.20, using the same pseudo-counts, we arrive at:

$$\min_{s,a,h}(-A^{*,h}_{\mathcal{M}\cup\tilde{r}}(s,a))\frac{\epsilon}{48\epsilon_{\tau-1}}\sqrt{\tau+1} \leq HR_{\max}\sqrt{8SA\log(12SAH\tau^2/\delta)}$$

Again, we can use Lemma 15 by Kaufmann et al. (2021) to find that

$$\tau \leq \tilde{\mathcal{O}}\left(\frac{H^3R^2_{\max}SA\epsilon^2_{\tau-1}}{\min_{s,a,h}(A^{*,h}_{\mathcal{M}\cup\tilde{r}}(s,a))^2\epsilon^2}\right)$$

□

B.6 Computing the Exploration Policy

To run AceIRL, we need to solve the optimization problem:

$$\pi^h_k = \min_{\pi}\max_{\hat{\pi}\in\hat{\Pi}_{k-1}}\sum_{h=0}^{H}\sum_{s',a'}\eta^{0,h}_{\widehat{\mathcal{M}},\hat{\pi}}(s',a';s_0)\hat{C}^h_k(s',a'|\pi)$$

For simplicity let us denote the state visitation frequencies by

$$\mu_h(s,a) \coloneqq \eta^{0,h}_{\widehat{\mathcal{M}},\pi}(s,a;s_0)$$
$$\hat{\mu}_h(s,a) \coloneqq \eta^{0,h}_{\widehat{\mathcal{M}},\hat{\pi}}(s,a;s_0)$$

Let us introduce the following matrix notation

$$\tilde{A} = \begin{bmatrix} I & 0 & 0 & 0 & \dots & 0 \\ \widehat{P} & -I & 0 & 0 & \dots & 0 \\ 0 & \widehat{P} & -I & 0 & \dots & 0 \\ & & \dots & & & \\ 0 & 0 & \dots & 0 & \widehat{P} & -I \\ I & 0 & 0 & \dots & 0 & 0 \\ 0 & I & 0 & \dots & 0 & 0 \\ & & \dots & & & \\ 0 & 0 & 0 & \dots & I & 0 \\ 0 & 0 & 0 & \dots & 0 & I \end{bmatrix}, \qquad a = \begin{bmatrix} \hat{r}^0_{k-1} \\ \hat{r}^1_{k-1} \\ \dots \\ \hat{r}^H_{k-1} \end{bmatrix}, \qquad A = \begin{bmatrix} A & 0 \\ a^T & -1 \end{bmatrix},$$

$$x = \begin{bmatrix} \mu_0 \\ \mu_1 \\ \dots \\ \mu_H \\ t \end{bmatrix}, \qquad \hat{x} = \begin{bmatrix} \hat{\mu}_0 \\ \hat{\mu}_1 \\ \dots \\ \hat{\mu}_H \end{bmatrix}, \qquad b = \begin{bmatrix} \bar{\mu}_0 \\ 0 \\ \dots \\ 0 \\ 1 \\ \dots \\ 1 \\ -10\epsilon_{k-1} \end{bmatrix}, \qquad c = \begin{bmatrix} C_0 \\ C_1 \\ \dots \\ C_H \\ 1 \end{bmatrix},$$

where $\bar{\mu}_0$ is the actual initial state distribution of the environment (which we assume to know). We can now write the inner maximization problem above as a linear program:

$$\begin{aligned} \max_x \ & c^T x \\ & Ax = b \\ & x \geq 0 \end{aligned}$$

The corresponding dual problem is:

$$\begin{aligned} \min_y \ & b^T y \\ & A^T y \geq c \end{aligned}$$

Using this we can write the full min-max problem as:

$$\begin{aligned} \min_{\hat{x},y} \ & b^T y \\ & A^T y \geq c(x) \\ & \tilde{A}x = b \\ & x \geq 0 \end{aligned}$$

which is a convex optimization problem, if we use:

$$C_h(s,a) = 2(H-h)R_{\max}\sqrt{\frac{2\log\left(24SAH(\max(1,n_k^h(s,a)))^2/\delta\right)}{\max(1,\hat{n}_{k+1}^h(s,a)}})$$

where $\hat{n}_{k+1}^h(s,a) = n_k^h(s,a) + \mu_h(s,a) * N_E$ is the number of times we expect h, s, a to be visited at the next iteration.

Solving this optimization problem yields the state-visitation frequencies $\hat{\mu}_k(s,a)$. We can then find the exploration policy that induces these state-visitations simply as:

$$\pi_{k,h}(a|s) \coloneqq \frac{\hat{\mu}_k^h(s,a)}{\sum_{a'}\hat{\mu}_k^h(s,a')}.$$

C Experimental Details

In this section, we provide more details on our experiments. We discuss the environments in detail (Appendix C.1), provide some information on the implementation and the libraries and computational resources we used (Appendix C.2), and we provide more full plots of all experiments we discussed in the main paper (Appendix C.3).

C.1 Details on the Environments

Four Paths. The four paths environment has 41 states and 4 actions:

$$\mathcal{S} = \{c, l_1, \dots, l_{10}, u_1, \dots, u_{10}, r_1, \dots, r_{10}, d_1, \dots, d_{10}\}, \qquad \mathcal{A} = \{a_1, a_2, a_3, a_4\},$$

and a time horizon of $H = 20$. The agent starts in the center state c, from which can move in four directions: left (a_1), up (a_2), right (a_3), or down (a_4). Each action a_i has a probability p_i of failing. If an action fails it moves in the opposite direction. $p_1, \dots, p_4$ are sampled uniformly from $(0, 0.3)$. One of the states $(l_{10}, u_{10}, r_{10}, d_{10})$ is chosen as the goal state at random. The reward in the goal state is 1, all other rewards are 0.

Double Chain. The *Double Chain* MDP, proposed by Kaufmann et al. (2021), consists of L states $\mathcal{S} = \{s_0, \dots, s_{L-1}\}$, and two actions $\mathcal{A} = \{\text{left}, \text{right}\}$, which correspond to a transition to the left or to the right. When the agent takes an action, there is a 0.1 probability of moving to the other direction. The state s_{L-1} has reward 1, all other states have reward 0, and the agent starts in the center of the chain at $s_{(L-1)/2}$. We choose $L = 31$, similar to Kaufmann et al. (2021). The environment has horizon $H = 20$.

Chain. The *Chain* MDP, proposed by Metelli et al. (2021) has 6 states $\mathcal{S} = \{s_1, s_2, s_3, s_4, s_5, s_u\}$ and 10 actions $\mathcal{A} = \{a_1, \dots, a_{10}\}$. The agent starts in a random initial state. Taking action a_{10} moves it right along the chain with probability 0.7 and to state s_u with probability 0.3. Any other action moves the agent right with probability 0.3 and to state s_u with probability 0.7. If the agent is in state s_u, action a_{10} moves it back to state s_1 with probability 0.05. Any other action moves it to s_1 with probability 0.01. The reward is 1 in all states except s_u where the reward is 0. Metelli et al. (2021) provide an illustration of the environment in Figure 3. We choose $H = 10$ for the chain.

Gridworld. The *Gridworld*, proposed by Metelli et al. (2021), is a 3×3 gridworld with an obstacle in the center cell $(2, 2)$ and a goal cell at the right center cell $(2, 1)$. The agent starts in a random non-goal cell, and it has 4 action one to move in each direction. If the agent takes an action with probability 0.3 the action fails and the agent moves in a random direction instead. If the agent is in the center cell $(2, 2)$ which has the obstacle, if the agent would move right it instead stays in the center cell with probability 0.8. The reward in the goal cell is 1, all other rewards are 0. Metelli et al. (2021) provide an illustration of the gridworld in Figure 6. We choose $H = 10$ for the gridworld.

Random MDPs. We generate random MDPs by uniformly sampling an initial state distribution and transition matrix and normalizing them. The rewards are sampled uniformly between 0 and 1. Our random MDPs have 9 states, 4 actions and horizon 10.

C.2 Implementation Details

We provide a full implementation of AceIRL in Python, using multiple open sources libraries, including `cvxpy` and the SCS optimizer (Diamond and Boyd, 2016; O'Donoghue et al., 2016) for solving the optimization problem in Appendix B.6, and standard libraries for numerical computing, including `numpy`, and `scipy`. We choose Maximum Entropy IRL (Ziebart et al., 2008) as an IRL algorithm, but AceIRL is agnostic to this choice.

We ran experiments in parallel on a server with two 64 Core AMD EPYC 7742 2.25GHz processors. We estimate a total wall-clock time of less than 48 hours for running all experiments presented in this paper, including 50 random seeds each.

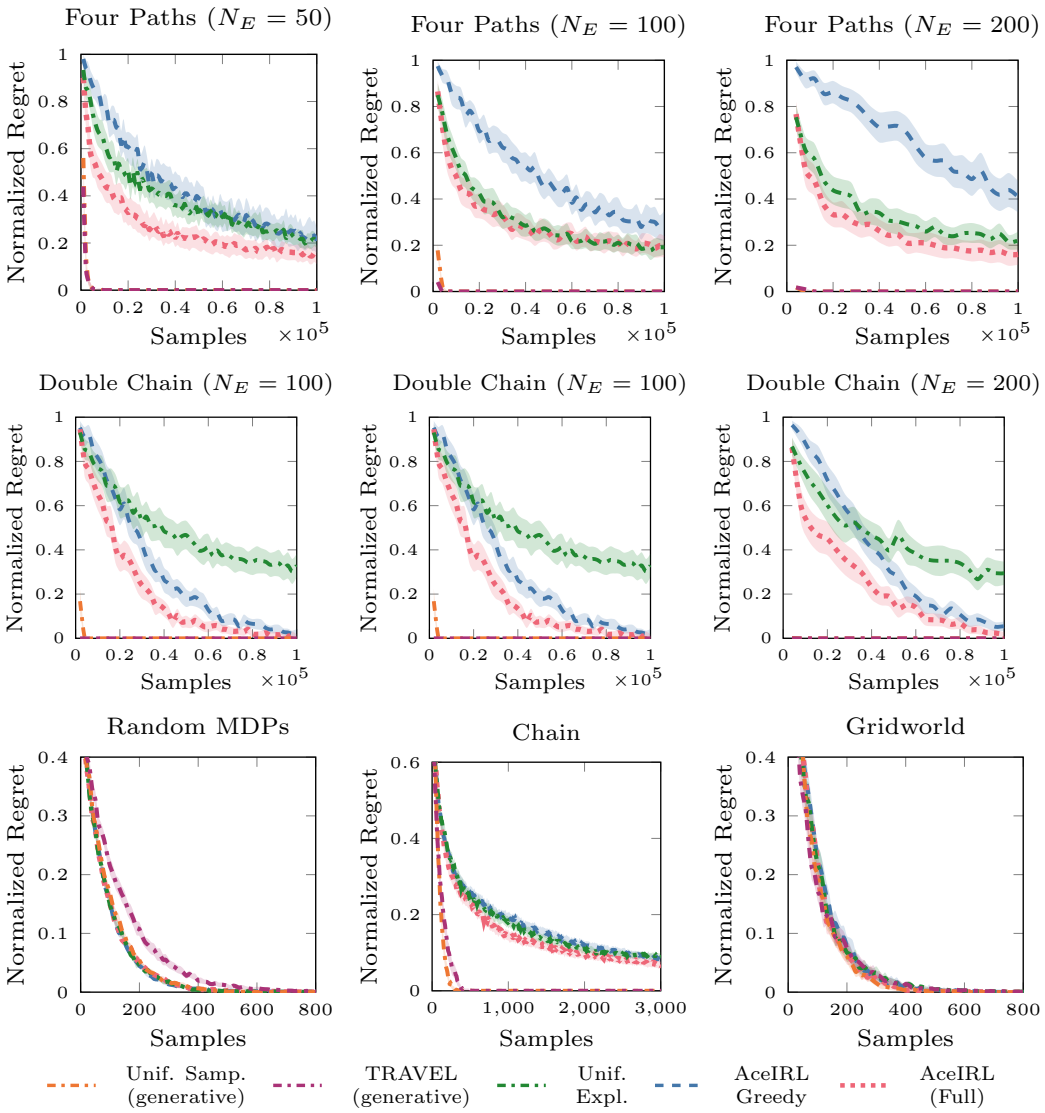

Figure C.1: Full learning curves for all experiments shown in Table 1. Similar to Figure 2, we show the mean and 95% confidence intervals computed over 50 random seeds. In addition to the exploration algorithms, we also show uniform sampling and TRAVEL which are much faster in most cases because they have access to a generative model.

C.3 Additional Results

We provide full learning curves for all experiments discussed in the main paper in Figure C.1.

D Connection to Reward-free Exploration

In the *reward-free exploration* problem, introduced by Jin et al. (2020), the agent explores an MDP\R to learn a transition model. In each iteration it chooses a new exploration policy based on previous data. The goal is to ensure that if the agent is given a reward function r after the exploration phase it can find a good policy using its transition model. Jin et al. (2020) formalize this goal as reducing the error:

$$V^{\pi^*,0}_{\mathcal{M}\cup r}(s_0) - V^{\hat{\pi}^*,0}_{\mathcal{M}\cup r},$$

where $\hat{\pi}^*$ is the optimal policy in the estimated MDP $\widehat{\mathcal{M}} \cup r$. Note the striking similarity between this problem, and the active IRL problem, we study in this paper. We want to reduce a similar error (cf. Definition 2), but we have additional information about the reward in form of the expert policy.

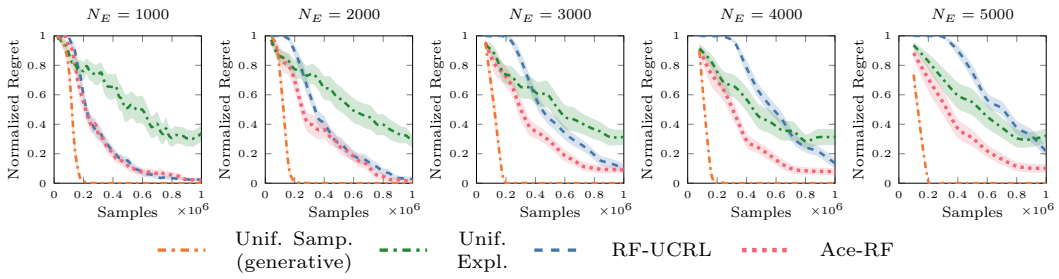

Figure D.2: Illustrative experiments for reward-free exploration in the *Double Chain* environment proposed by Kaufmann et al. (2021). The difference to our Active IRL setting is that the agent does not have access to the expert policy during exploration, but still tries to learn a good model of the environment. During testing it then gets access to the reward function, and the regret measures the suboptimality of the policy trained in the agent's transition model. We find that the ideas used in AceIRL are also useful for batched reward-free exploration with larget N_E.

The *Reward-free UCRL* algorithm, proposed by Kaufmann et al. (2021), is essentially analogous to AceIRL Greedy (Section 6.1). Reward-free UCRL explores greedily with respect to an upper bound on the value function error. However, the exploration policy needs to be updated after each episode to adapt to the new uncertainty estimates. This might be expensive or not possible in practice. Instead, we could consider a *batched* version of reward-free exploration, where in each iteration the agent explores for N_E episodes, similar to our Active IRL problem. In this setting, a greedy policy w.r.t. uncertainty is suboptimal because it does not adapt to the reduced uncertainty over the N_E episodes.

Instead, we can consider reducing the expected uncertainty at the next iteration, similar to our discussion in Section 6.2. If our error estimate is denoted by $E_k(s, a)$, we do no longer act greedily w.r.t. E_k. Instead we try to estimate the error at the next iteration $\hat{E}_{k+1}(s, a|\pi)$ as a function of the policy and try to select the policy that reduces this error. In the tabular case, we can formulate this as a convex optimization problem, analogous to Appendix B.6. We call this adaptation of AceIRL to the reward-free exploration problem *Ace-RF*.

Figure D.2 shows illustrative results of this algorithm in the batched reward-free exploration setting in the *Double Chain* environment. We find that for larger batch sizes, choosing an exploration policy that reduces future uncertainty is significantly better than reward-free UCRL.