APPENDIX

In the appendix, we provide additional details of the theories and the experiments. The contents of this appendix are as follows.

- In Appendix A, we discuss some details of PAGAR and $Minimax Regret$ that were omitted in Section 4. We briefly introduce some necessary preliminaries in Appendix A.1. Then we derive a Theorem 5 to support **Table 1** in Appendix A.2. The proves for **Theorem** 1 and 1, **Corollary** 2 and 3, **Theorem** 1 are in Appendix A.4. The details of **Example** 1 are in Appendix A.5.
- In Appendix B, we provide some details of Imitation Learning with PAGAR that were omitted in Section 5. we prove **Theorem 4** in Appendix B.1. Then we derive the objective functions in Appendix B.2. Some details of Algorithm 1 will be explained in Appendix B.3.
- In Appendix C, we provide some experimental details and additional results.

A REWARD DESIGN WITH PAGAR

This paper does not aim to resolve the ambiguity problem in IRL but provides a way to circumvent it so that reward ambiguity does not lead to reward misalignment in IRL-based IL. PAGAR, the semi-supervised reward design paradigm proposed in this paper, tackles this problem from the perspective of semi-supervised reward design. But the nature of PAGAR is distinct from IRL and IL: assume that a set of reward functions is available for some underlying task, where some of those reward functions align with the task while others are misaligned, PAGAR provides a solution for selecting reward functions to train a policy that successfully performs the task, without knowing which reward function aligns with the task. Our research demonstrates that policy training with PAGAR is equivalent to learning a policy to maximize an affine combination of utilities measured under a distribution of the reward functions in the reward function set. With this understanding of PAGAR, we integrate it with IL to illustrate its advantages.

A.1 SEMI-SUPERVISED REWARD DESIGN

Designing a reward function can be thought as deciding an ordering of policies. We adopt a concept, called *total domination*, from unsupervised environment design Dennis et al. (2020), and re-interpret this concept in the context of reward design. In this paper, we suppose that a function $U_r(\pi)$ is given to measure the performance of a policy. While the measurement of policy performance can vary depending on the free variable r, *total dominance* can be viewed as an invariance regardless of such dependency.

Definition 3 (Total Domination)**.** A policy, π_1, is totally dominated by some policy π_2 w.r.t a reward function set R, if for every pair of reward functions $r_1, r_2 \in R$, $U_{r_1}(\pi_1) < U_{r_2}(\pi_2)$.

If π_1 totally dominate π_2 w.r.t R, π_2 can be regarded as being unconditionally better than π_1. In other words, the two sets $\{U_r(\pi_1)|r \in R\}$ and $\{U_r(\pi_2)|r \in R\}$ are disjoint, such that $\sup\{U_r(\pi_1)|r \in R\} < \inf\{U_r(\pi_2)|r \in R\}$. Conversely, if a policy π is not totally dominated by any other policy, it indicates that for any other policy, say π_2, $\sup\{U_r(\pi_1)|r \in R\} \geq \inf\{U_r(\pi_2)|r \in R\}$.

Definition 4. A reward function set R aligns with an ordering $\prec_R$ among policies such that $\pi_1 \prec_R \pi_2$ if and only if π_1 is totally dominated by π_2 w.r.t. R.

Especially, designing a reward function r is to establish an ordering $\prec_{\{r\}}$ among policies. Total domination can be extended to policy-conditioned reward design, where the reward function r is selected by following a decision rule $\mathcal{R}(\pi)$ such that $\sum_{r\in R}\mathcal{R}(\pi)(r) = 1$. We let $\mathcal{U}_{\mathcal{R}}(\pi) = \sum_{r\in R_E}\mathcal{R}(\pi)(r)\cdot U_r(\pi)$ be an affine combination of $U_r(\pi)$'s with its coefficients specified by $\mathcal{R}(\pi)$.

Definition 5. A policy conditioned decision rule $\mathcal{R}$ is said to prefer a policy π_1 to another policy π_2, which is notated as $\pi_1 \prec^{\mathcal{R}} \pi_2$, if and only if $\mathcal{U}_{\mathcal{R}}(\pi_1) < \mathcal{U}_{\mathcal{R}}(\pi_2)$.

Making a decision rule for selecting reward functions from a reward function set to respect the total dominance w.r.t this reward function set is an unsupervised learning problem, where no additional

external supervision is provided. If considering expert demonstrations as a form of supervision and using it to constrain the set R_E of reward function via IRL, the reward design becomes semi-supervised.

A.2 SOLUTION TO THE MINIMAXREGRET

In Table 1, we mentioned that solving $MinimaxRegret(R_E)$ is equivalent to finding an optimal policy π^* to maximize a $\mathcal{U}_{\mathcal{R}_E}(\pi)$ under a decision rule $\mathcal{R}_E$. Without loss of generality, we use R instead of R_E in our subsequent analysis, because solving $MinimaxRegret(R)$ does not depend on whether there are constraints for R. In order to show such an equivalence, we follow the same routine as in Dennis et al. (2020), and start by introducing the concept of *weakly total domination*.

Definition 6 (Weakly Total Domination)**.** A policy π_1 is *weakly totally dominated* w.r.t a reward function set R by some policy π_2 if and only if for any pair of reward function $r_1, r_2 \in R$, $U_{r_1}(\pi_1) \leq U_{r_2}(\pi_2)$.

Note that a policy π being totally dominated by any other policy is a sufficient but not necessary condition for π being weakly totally dominated by some other policy. A policy π_1 being weakly totally dominated by a policy π_2 implies that $\sup\{U_r(\pi_1)|r \in R\} \leq \inf\{U_r(\pi_2)|r \in R\}$. We assume that there does not exist a policy π that weakly totally dominates itself, which could happen if and only if $U_r(\pi)$ is a constant. We formalize this assumption as the following.

Assumption 1. For the given reward set R and policy set Π, there does not exist a policy π such that for any two reward functions $r_1, r_2 \in R$, $U_{r_1}(\pi) = U_{r_2}(\pi)$.

This assumption makes weak total domination a non-reflexive relation. It is obvious that weak total domination is transitive and asymmetric. Now we show that successive weak total domination will lead to total domination.

Lemma 1. *for any three policies $\pi_1, \pi_2, \pi_3 \in \Pi$, if π_1 is weakly totally dominated by π_2, π_2 is weakly totally dominated by π_3, then π_3 totally dominates π_1.*

Proof. According to the definition of weak total domination, $\max_{r\in R} U_r(\pi_1) \leq \min_{r\in R} U_r(\pi_2)$ and $\max_{r\in R} U_r(\pi_2) \leq \min_{r\in R} U_r(\pi_3)$. If π_1 is weakly totally dominated but not totally dominated by π_3, then $\max_{r\in R} U_r(\pi_1) = \min_{r\in R} U_r(\pi_3)$ must be true. However, it implies $\min_{r\in R} U_r(\pi_2) = \max_{r\in R} U_r(\pi_2)$, which violates Assumption 1. We finish the proof. □

Lemma 2. *For the set $\Pi_{\neg wtd} \subseteq \Pi$ of policies that are not weakly totally dominated by any other policy in the whole set of policies w.r.t a reward function set R, there exists a range $\mathbb{U} \subseteq \mathbb{R}$ such that for any policy $\pi \in \Pi_{\neg wtd}$, $\mathbb{U} \subseteq [\min_{r\in R} U_r(\pi), \max_{r\in R} U_r(\pi)]$.*

Proof. For any two policies $\pi_1, \pi_2 \in \Pi_{\neg wtd}$, it cannot be true that $\max_{r\in R} U_r(\pi_1) = \min_{r\in R} U_r(\pi_2)$ nor $\min_{r\in R} U_r(\pi_1) = \max_{r\in R} U_r(\pi_2)$, because otherwise one of the policies weakly totally dominates the other. Without loss of generalization, we assume that $\max_{r\in R} U_r(\pi_1) > \min_{r\in R} U_r(\pi_2)$. In this case, $\max_{r\in R} U_r(\pi_2) > \min_{r\in R} U_r(\pi_1)$ must also be true, otherwise π_1 weakly totally dominates π_2. Inductively, $\min_{\pi\in\Pi_{\neg wtd}} \max_{r\in R} U_r(\pi) > \max_{\pi\in\Pi_{\neg wtd}} \min_{r\in R} U_r(\pi)$. Letting $ub = \min_{\pi\in\Pi_{\neg wtd}} \max_{r\in R} U_r(\pi)$ and $lb = \max_{\pi\in\Pi_{\neg wtd}} \min_{r\in R} U_r(\pi)$, any $\mathbb{U} \subseteq [lb, ub]$ shall support the assertion. We finish the proof. □

Lemma 3. *For a reward function set R, if a policy $\pi \in \Pi$ is weakly totally dominated by some other policy in Π and there exists a subset $\Pi_{\neg wtd} \subseteq \Pi$ of policies that are not weakly totally dominated by any other policy in π, then $\max_{r\in R} U_r(\pi) < \min_{\pi'\in\Pi_{\neg wtd}} \max_{r\in R} U_r(\pi')$*

Proof. If π_1 is weakly totally dominated by a policy $\pi_2 \in \Pi$, then $\min_{r\in R} U_r(\pi_2) = \max_{r\in R} U_r(\pi)$. If $\max_{r\in R} U_r(\pi) \geq \min_{\pi'\in\Pi_{\neg wtd}} \max_{r\in R} U_r(\pi')$, then $\min_{r\in R} U_r(\pi_2) \geq \min_{\pi'\in\Pi_{\neg wtd}} \max_{r\in R} U_r(\pi')$, making at

least one of the policies in $\Pi_{\neg wtd}$ being weakly totally dominated by π_2. Hence, $\max_{r\in R} U_r(\pi) < \min_{\pi'\in\Pi_{\neg wtd}} \max_{r\in R} U_r(\pi')$ must be true. □

Given a policy π and a reward function r, the regret is represented as Eq.8

$$Regret(\pi, r) \quad := \quad \max_{\pi'} U_r(\pi') - U_r(\pi) \tag{8}$$

Then we represent the $MinimaxRegret(R)$ problem in Eq.9.

$$MinimaxRegret(R) \quad := \quad \arg\min_{\pi\in\Pi} \left\{ \max_{r\in R} Regret(\pi, r) \right\} \tag{9}$$

We denote as $r^*_\pi \in R$ the reward function that maximizes $U_r(\pi)$ among all the r's that achieve the maximization in Eq.9. Formally,

$$r^*_\pi \quad \in \quad \arg\max_{r\in R} U_r(\pi) \qquad s.t.\ r \in \arg\max_{r'\in R} Regret(\pi, r') \tag{10}$$

Then $MinimaxRegret$ can be defined as minimizing the worst-case regret as in Eq.9. Next, we want to show that for some decision rule $\mathcal{R}$, the set of optimal policies which maximizes $\mathcal{U}_\mathcal{R}$ are the solutions to $MinimaxRegret(R)$. Formally,

$$MinimaxRegret(R) = \arg\max_{\pi\in\Pi} \mathcal{U}_\mathcal{R}(\pi) \tag{11}$$

We design $\mathcal{R}$ by letting $\mathcal{R}(\pi) := \overline{\mathcal{R}}(\pi) \cdot \delta_{r^*_\pi} + (1 - \overline{\mathcal{R}}(\pi)) \cdot \underline{\mathcal{R}}(\pi)$ where $\underline{\mathcal{R}} : \Pi \to \Delta(R)$ is a policy conditioned distribution over reward functions, $\delta_{r^*_\pi}$ be a delta distribution centered at r^*_π, and $\overline{\mathcal{R}}(\pi)$ is a coefficient. We show how to design $\underline{\mathcal{R}}$ by using the following lemma.

Lemma 4. *Given that the reward function set is R, there exists a decision rule $\underline{\mathcal{R}} : \Pi \to \Delta(R)$ which guarantees that: 1) for any policy π that is not weakly totally dominated by any other policy in Π, i.e., $\pi \in \Pi_{\neg wtd} \subseteq \Pi$, $\mathcal{U}_{\underline{\mathcal{R}}}(\pi) \equiv c$ where $c = \max_{\pi'\in\Pi_{\neg wtd}} \min_{r\in R} U_r(\pi')$; 2) for any π that is weakly totally dominated by some policy but not totally dominated by any policy, $\mathcal{U}_{\underline{\mathcal{R}}}(\pi) = \max_{r\in R} U_r(\pi)$; 3) if π is totally dominated by some other policy, $\overline{\mathcal{R}}(\pi)$ is a uniform distribution.*

Proof. Since the description of $\underline{\mathcal{R}}$ for the policies in condition 2) and 3) are self-explanatory, we omit the discussion on them. For the none weakly totally dominated policies in condition 1), having a constant $\mathcal{U}_{\underline{\mathcal{R}}}(\pi) \equiv c$ is possible if and only if for any policy $\pi \in \Pi_{\neg wed}$, $c \in [\min_{r\in R} U_r(\pi'), \max_{r\in R} U_r(\pi')]$. As mentioned in the proof of Lemma 2, c can exist within $[\min_{r\in R} U_r(\pi), \max_{r\in R} U_r(\pi)]$. Hence, $c = \max_{\pi'\in\Pi_{\neg wtd}} \min_{r\in R} U_r(\pi')$ is a valid assignment. □

Then by letting $\overline{\mathcal{R}}(\pi) := \frac{Regret(\pi, r^*_\pi)}{c - U_{r^*_\pi}(\pi)}$, we have the following theorem.

Theorem 5. *By letting $\mathcal{R}(\pi) := \overline{\mathcal{R}}(\pi) \cdot \delta_{r^*_\pi} + (1 - \overline{\mathcal{R}}(\pi)) \cdot \underline{\mathcal{R}}(\pi)$ with $\overline{\mathcal{R}}(\pi) := \frac{Regret(\pi, r^*_\pi)}{c - U_{r^*_\pi}(\pi)}$ and any $\underline{\mathcal{R}}$ that satisfies Lemma 4,*

$$MinimaxRegret(R) = \arg\max_{\pi\in\Pi} \mathcal{U}_\mathcal{R}(\pi) \tag{12}$$

Proof. If a policy $\pi \in \Pi$ is totally dominated by some other policy, since there exists another policy with larger $\mathcal{U}_\mathcal{R}$, π cannot be a solution to $\arg\max_{\pi\in\Pi} \mathcal{U}_\mathcal{R}(\pi)$. Hence, there is no need for further discussion on totally dominated policies. We discuss the none weakly totally dominated policies

and the weakly totally dominated but not totally dominated policies (shortened to "weakly totally dominated" from now on) respectively. First we expand $\arg\max_{\pi\in\Pi}\mathcal{U}_{\mathcal{R}}(\pi)$ as in Eq.13.

$$
\begin{aligned}
& \arg\max_{\pi\in\Pi}\mathcal{U}_{\mathcal{R}}(\pi) \\
= \ & \arg\max_{\pi\in\Pi}\sum_{r\in R}\mathcal{R}(\pi)(r)\cdot U_r(\pi) \\
= \ & \arg\max_{\pi\in\Pi}\frac{Regret(\pi,r^*_\pi)\cdot U_{r^*_\pi}(\pi)+(\mathcal{U}_{\underline{\mathcal{R}}}(\pi)-U_{r^*_\pi}(\pi)-Regret(\pi,r^*_\pi))\cdot\mathcal{U}_{\underline{\mathcal{R}}}(\pi)}{c-U_{r^*_\pi}(\pi)} \\
= \ & \arg\max_{\pi\in\Pi}\frac{(\mathcal{U}_{\underline{\mathcal{R}}}(\pi)-U_{r^*_\pi}(\pi))\cdot\mathcal{U}_{\underline{\mathcal{R}}}(\pi)-(\mathcal{U}_{\underline{\mathcal{R}}}(\pi)-U_{r^*_\pi}(\pi))\cdot Regret(\pi,r^*_\pi))}{c-U_{r^*_\pi}(\pi)} \\
= \ & \arg\max_{\pi\in\Pi}\frac{\mathcal{U}_{\underline{\mathcal{R}}}(\pi)-U_{r^*_\pi}(\pi)}{c-U_{r^*_\pi}(\pi)}\cdot\mathcal{U}_{\underline{\mathcal{R}}}(\pi)-Regret(\pi,r^*_\pi)
\end{aligned}
\tag{13}
$$

1) For the none weakly totally dominated policies, since by design $\mathcal{U}_{\underline{\mathcal{R}}}\equiv c$, Eq.13 is equivalent to $\arg\max_{\pi\in\Pi_1}\ -Regret(\pi,r^*_\pi)$ which exactly equals $MinimaxRegret(R)$. Hence, the equivalence holds among the none weakly totally dominated policies. Furthermore, if a none weakly totally dominated policy $\pi\in\Pi_{\neg wtd}$ achieves optimality in $MinimaxRegret(R)$, its $\mathcal{U}_{\mathcal{R}}(\pi)$ is also no less than any weakly totally dominated policy. Because according to Lemma 3, for any weakly totally dominated policy π_1, its $\mathcal{U}_{\underline{\mathcal{R}}}(\pi_1)\le c$, hence $\frac{\mathcal{U}_{\underline{\mathcal{R}}}(\pi)-U_{r^*_\pi}(\pi)}{c-U_{r^*_\pi}(\pi)}\cdot\mathcal{U}_{\underline{\mathcal{R}}}(\pi_1)\le c$. Since $Regret(\pi,r^*_\pi)\le Regret(\pi_1,r^*_{\pi_1})$, $\mathcal{U}_{\mathcal{R}}(\pi)\ge\mathcal{U}_{\mathcal{R}}(\pi_1)$. Therefore, we can assert that if a none weakly totally dominated policy π is a solution to $MinimaxRegret(R)$, it is also a solution to $\arg\max_{\pi\in\Pi}\mathcal{U}_{\mathcal{R}}(\pi)$. Additionally, to prove that if a none weakly totally dominated policy π is a solution to $\arg\max_{\pi'\in\Pi}\mathcal{U}_{\mathcal{R}}(\pi')$, it is also a solution to $MinimaxRegret(R)$, it is only necessary to prove that π achieve no larger regret than all the weakly totally dominated policies. But we delay the proof to 2).

2) If a policy π is weakly totally dominated and is a solution to $MinimaxRegret(R)$, we show that it is also a solution to $\arg\max_{\pi\in\Pi}\mathcal{U}_{\mathcal{R}}(\pi)$, i.e., its $\mathcal{U}_{\mathcal{R}}(\pi)$ is no less than that of any other policy.

We start by comparing with non weakly totally dominated policy. for any weakly totally dominated policy $\pi_1\in MinimaxRegret(R)$, it must hold true that $Regret(\pi_1,r^*_{\pi_1})\le Regret(\pi_2,r^*_{\pi_2})$ for any $\pi_2\in\Pi$ that weakly totally dominates π_1. However, it also holds that $Regret(\pi_2,r^*_{\pi_2})\le Regret(\pi_1,r^*_{\pi_2})$ due to the weak total domination. Therefore, $Regret(\pi_1,r^*_{\pi_1})=Regret(\pi_2,r^*_{\pi_2})=Regret(\pi_1,r^*_{\pi_2})$, implying that π_2 is also a solution to $MinimaxRegret(R)$. It also implies that $U_{r^*_{\pi_2}}(\pi_1)=U_{r^*_{\pi_2}}(\pi_2)\ge U_{r^*_{\pi_1}}(\pi_1)$ due to the weak total domination. However, by definition $U_{r^*_{\pi_1}}(\pi_1)\ge U_{r^*_{\pi_2}}(\pi_1)$. Hence, $U_{r^*_{\pi_1}}(\pi_1)=U_{r^*_{\pi_2}}(\pi_1)=U_{r^*_{\pi_2}}(\pi_2)$ must hold. Now we discuss two possibilities: a) there exists another policy π_3 that weakly totally dominates π_2; b) there does not exist any other policy that weakly totally dominates π_2. First, condition a) cannot hold. Because inductively it can be derived $U_{r^*_{\pi_1}}(\pi_1)=U_{r^*_{\pi_2}}(\pi_1)=U_{r^*_{\pi_2}}(\pi_2)=U_{r^*_{\pi_3}}(\pi_3)$, while Lemma 1 indicates that π_3 totally dominates π_1, which is a contradiction. Hence, there does not exist any policy that weakly totally dominates π_2, meaning that condition b) is certain. We note that $U_{r^*_{\pi_1}}(\pi_1)=U_{r^*_{\pi_2}}(\pi_1)=U_{r^*_{\pi_2}}(\pi_2)$ and the weak total domination between π_1,π_2 imply that $r^*_{\pi_1},r^*_{\pi_2}\in\arg\max_{r\in R}U_r(\pi_1)$, $r^*_{\pi_2}\in\arg\min_{r\in R}U_r(\pi_2)$, and thus $\min_{r\in R}U_r(\pi_2)\le\max_{\pi\in\Pi_{\neg wtd}}\min_{r\in R}U_r(\pi)=c$. Again, $\pi_1\in MinimaxRegret(R)$ makes $Regret(\pi_1,r^*_\pi)\le Regret(\pi_1,r^*_{\pi_1})\le Regret(\pi,r^*_\pi)$ not only hold for $\pi=\pi_2$ but also for any other policy $\pi\in\Pi_{\neg wtd}$, then for any policy $\pi\in\Pi_{\neg wtd}$, $U_{r^*_\pi}(\pi_1)\ge U_{r^*_\pi}(\pi)\ge\min_{r\in R}U_r(\pi)$. Hence, $U_{r^*_\pi}(\pi_1)\ge\max_{\pi\in\Pi_{\neg wtd}}\min_{r\in R}U_r(\pi)=c$. Since $U_{r^*_\pi}(\pi_1)=\min_{r\in R}U_r(\pi_2)$ as aforementioned, $\min_{r\in R}U_r(\pi_2)>\max_{\pi\in\Pi_{\neg wtd}}\min_{r\in R}U_r(\pi)$ will cause a contradiction. Hence, $\min_{r\in R}U_r(\pi_2)=\max_{\pi\in\Pi_{\neg wtd}}\min_{r\in R}U_r(\pi)=c$. As a result, $\mathcal{U}_{\underline{\mathcal{R}}}(\pi)=U_{r^*_\pi}(\pi)=\max_{\pi'\in\Pi_{\neg wtd}}\min_{r\in R}U_r(\pi')=c$, and $\mathcal{U}_{\mathcal{R}}(\pi)=c-Regret(\pi,r^*_\pi)\ge\max_{\pi'\in\Pi_{\neg wtd}}c-Regret(\pi',r^*_{\pi'})=\max_{\pi'\in\Pi_{\neg wtd}}\mathcal{U}_{\mathcal{R}}(\pi')$. In other words, if a weakly

totally dominated policy π is a solution to $MinimaxRegret(R)$, then its $\mathcal{U}_{\mathcal{R}}(\pi)$ is no less than that of any non weakly totally dominated policy. This also complete the proof at the end of 1), because if a none weakly totally dominated policy π_1 is a solution to $\arg\max_{\pi\in\Pi}\mathcal{U}_{\mathcal{R}}(\pi)$ but not a solution to $MinimaxRegret(R)$, then $Regret(\pi_1, r^*_{\pi_1}) > 0$ and a weakly totally dominated policy π_2 must be the solution to $MinimaxRegret(R)$. Then, $\mathcal{U}_{\mathcal{R}}(\pi_2) = c > c - Regret(\pi_1, r^*_{\pi_1}) = \mathcal{U}_{\mathcal{R}}(\pi_1)$, which, however, contradicts $\pi_1 \in \arg\max_{\pi\in\Pi}\mathcal{U}_{\mathcal{R}}(\pi)$.

It is obvious that a weakly totally dominated policy $\pi \in MinimaxRegret(R)$ has a $\mathcal{U}_{\mathcal{R}}(\pi)$ no less than any other weakly totally dominated policy. Because for any other weakly totally dominated policy $\pi_1, \mathcal{U}_{\underline{\mathcal{R}}}(\pi_1) \leq c$ and $Regret(\pi_1, r^*_{\pi_1}) \leq Regret(\pi, r^*_{\pi})$, hence $\mathcal{U}_{\mathcal{R}}(\pi_1) \leq \mathcal{U}_{\mathcal{R}}(\pi)$ according to Eq.13.

So far we have shown that if a weakly totally dominated policy π is a solution to $MinimaxRegret(R)$, it is also a solution to $\arg\max_{\pi'\in\Pi}\mathcal{U}_{\mathcal{R}}(\pi')$. Next, we need to show that the reverse is also true, i.e., if a weakly totally dominated policy π is a solution to $\arg\max_{\pi\in\Pi}\mathcal{U}_{\mathcal{R}}(\pi)$, it must also be a solution to $MinimaxRegret(R)$. In order to prove its truthfulness, we need to show that if $\pi \notin MinimaxRegret(R)$, whether there exists: a) a none weakly totally dominated policy π_1, or b) another weakly totally dominated policy π_1, such that $\pi_1 \in MinimaxRegret(R)$ and $\mathcal{U}_{\mathcal{R}}(\pi_1) \leq \mathcal{U}_{\mathcal{R}}(\pi)$. If neither of the two policies exists, we can complete our proof. Since it has been proved in 1) that if a none weakly totally dominated policy achieves $MinimaxRegret(R)$, it also achieves $\arg\max_{\pi'\in\Pi}\mathcal{U}_{\mathcal{R}}(\pi')$, the policy described in condition a) does not exist. Hence, it is only necessary to prove that the policy in condition b) also does not exist.

If such weakly totally dominated policy π_1 exists, $\pi \notin MinimaxRegret(R)$ and $\pi_1 \in MinimaxRegret(R)$ indicates $Regret(\pi, r^*_{\pi}) > Regret(\pi_1, r^*_{\pi_1})$. Since $\mathcal{U}_{\mathcal{R}}(\pi_1) \geq \mathcal{U}_{\mathcal{R}}(\pi)$, according to Eq.13, $\mathcal{U}_{\mathcal{R}}(\pi_1) = c - Regret(\pi_1, r^*_{\pi_1}) \leq \mathcal{U}_{\mathcal{R}}(\pi) = \frac{\mathcal{U}_{\mathcal{R}}(\pi) - U_{r^*_{\pi}}(\pi)}{c - U_{r^*_{\pi}}(\pi)} \cdot \mathcal{U}_{\underline{\mathcal{R}}}(\pi) - Regret(\pi, r^*_{\pi})$. Thus $\frac{\mathcal{U}_{\mathcal{R}}(\pi) - U_{r^*_{\pi}}(\pi)}{c - U_{r^*_{\pi}}(\pi)}(\pi) \cdot \mathcal{U}_{\underline{\mathcal{R}}} \geq c + Regret(\pi, r^*_{\pi}) - Regret(\pi_1, r^*_{\pi_1}) > c$, which is impossible due to $\mathcal{U}_{\underline{\mathcal{R}}} \leq c$. Therefore, such π_1 also does not exist. In fact, this can be reasoned from another perspective. If there exists a weakly totally dominated policy π_1 with $U_{r^*_{\pi_1}}(\pi_1) = c = U_{r^*_{\pi}}(\pi)$ but $\pi_1 \notin MinimaxRegret(R)$, then $Regret(\pi, r^*_{\pi}) > Regret(\pi_1, r^*_{\pi_1})$. It also indicates $\max_{\pi'\in\Pi} U_{r^*_{\pi}}(\pi') > \max_{\pi'\in\Pi} U_{r^*_{\pi_1}}(\pi')$. Meanwhile, $Regret(\pi_1, r^*_{\pi}) := \max_{\pi'\in\Pi} U_{r^*_{\pi}}(\pi') - U_{r^*_{\pi}}(\pi_1) \leq Regret(\pi_1, r^*_{\pi_1}) := \max_{\pi'\in\Pi} U_{r^*_{\pi_1}}(\pi') - U_{r^*_{\pi_1}}(\pi_1) := \max_{r\in R}\max_{\pi'\in\Pi} U_r(\pi') - U_r(\pi_1)$ indicates $\max_{\pi'\in\Pi} U_{r^*_{\pi}}(\pi') - \max_{\pi'\in\Pi} U_{r^*_{\pi_1}}(\pi') \leq U_{r^*_{\pi}}(\pi_1) - U_{r^*_{\pi_1}}(\pi_1)$. However, we have proved that, for a weakly totally dominated policy, $\pi_1 \in MinimaxRegret(R)$ indicates $U_{r^*_{\pi_1}}(\pi_1) = \max_{r\in R} U_r(\pi_1)$. Hence, $\max_{\pi'\in\Pi} U_{r^*_{\pi}}(\pi') - \max_{\pi'\in\Pi} U_{r^*_{\pi_1}}(\pi') \leq U_{r^*_{\pi}}(\pi_1) - U_{r^*_{\pi_1}}(\pi_1) \leq 0$ and it contradicts $\max_{\pi'\in\Pi} U_{r^*_{\pi}}(\pi') > \max_{\pi'\in\Pi} U_{r^*_{\pi_1}}(\pi')$. Therefore, such π_1 does not exist. In summary, we have exhausted all conditions and can assert that for any policies, being a solution to $MinimaxRegret(R)$ is equivalent to a solution to $\arg\max_{\pi\in\Pi}\mathcal{U}_{\mathcal{R}}(\pi)$. We complete our proof.

□

A.3 Measuring Policy Performance

Recall that the function $U_r(\pi)$ is used to measure the performance of a policy π under a reward function r. In Dennis et al. (2020), $U_r(\pi) = \eta_r(\pi)$. In this section, we discuss the validity of letting $U_r(\pi)$ be the loss function of a generic IRL objective, e.g., $U_r(\pi) = \eta_r(\pi) - \eta_r(\pi_E)$ where $\eta_r(\pi_E)$ measures the expected return of the expert policy π_E and can be estimated if an expert demonstration set E instead of π_E is provided. If further letting $R_E = \{r | r \in \arg\min_{r'\in R}\max_{\pi\in\Pi} U_{r'}(\pi) - U_{r'}(\pi_E)\}$, $\max_{\pi\in\Pi} U_r(\pi)$ is a constant for any $r \in R_E$, notated as $u := \max_{\pi\in\Pi} U_r(\pi)$. Because by definition $R_E = \{r | r \in \arg\min_{r\in R}\max_{\pi\in\Pi} U_r(\pi)\}$. If there exists $r_1, r_2 \in R_E$ such that $\max_{\pi\in\Pi} U_{r_1}(\pi) < \max_{\pi\in\Pi} U_{r_2}(\pi)$, r_2 will not be a member of R_E. Furthermore, $\{U_r(\pi) | \pi \in \Pi, r \in R_E\}$ will be upper-bounded by a

constant $u = \max_{\pi \in \Pi} U_r(\pi)$. Because if there exists a policy $\pi \in \Pi$ and a reward function $r \in R_E$ with $U_r(\pi) > u$, it contradicts the fact that $u = \max_{\pi' \in \Pi} U_r(\pi')$. In this case, $MinimaxRegret(R_E) = \arg\min_{\pi \in \Pi} \max_{r \in R_E} Regret(\pi, r) = \arg\min_{\pi \in \Pi} \max_{r \in R_E} u - U_r(\pi) = \arg\max_{\pi \in \Pi} \min_{r \in R_E} U_r(\pi)$. Note that before making any other assumption on R_E, Π and $U_r(\cdot)$, $\max_{\pi \in \Pi} \min_{r \in R_E} U_r(\pi)$ cannot be regarded as the same as IRL itself $\min_{r \in R} \max_{\pi \in \Pi} U_r(\pi)$. The solution to $\arg\max_{\pi \in \Pi} \min_{r \in R_E} U_r(\pi)$ is the policy with the highest worst case $U_r(\pi)$ for $r \in R_E$. The IRL problem however may induce a policy that maximizes $U_r(\cdot)$ for some $r \in R_E$ while minimizing $U_{r'}(\cdot)$ for some other $r' \in R_E$. While $\min_{r \in R} \max_{\pi \in \Pi} U_r(\pi) = u$, it is possible that $\max_{\pi \in \Pi} \min_{r \in R_E} U_r(\pi) < u$. In fact, it is easily observable that the solutions to $MinimaxRegret$ with some $U_r(\pi)$ will be the same as that of letting $U_r(\pi) := U_r(\pi) - U_r(\pi_E)$. Hence, in this paper we simply use $\eta_r(\pi)$ as $U_r(\pi)(R)$.

Lemma 5. *If a policy $\pi \in MinimaxRegret(R_E)$ when the policy performance is measured with some U_r, then $\pi \in MinimaxRegret(R_E)$ when letting $U_r(\pi) := U_r(\pi) - U_r(\pi_E)$.*

Proof. When using $U_r(\pi) := U_r(\pi) - U_r(\pi_E)$ to measure the policy performance, solving $MinimaxRegret(R)$ is to solve Eq. 14, which is the same as Eq.9.

$$\begin{aligned} MimimaxRegret(R_E) &= \arg\max_{\pi \in \Pi} \min_{r \in R_E} Regret(\pi, r) \\ &= \arg\max_{\pi \in \Pi} \min_{r \in R_E} \max_{\pi' \in \Pi} \{U_r(\pi') - U_r(\pi_E)\} - (U_r(\pi) - U_r(\pi_E) \\ &= \arg\max_{\pi \in \Pi} \min_{r \in R_E} \max_{\pi' \in \Pi} U_r(\pi') - U_r(\pi) \end{aligned} \tag{14}$$

□

A.4 Criterion for Successful Policy Learning

Theorem 1.(Task-Failure Avoidance) If the following conditions (1) (2) hold for R, then the optimal protagonist policy $\pi_P := MinimaxRegret(R)$ satisfies that $\forall r_{al} \in R, U_{r_{al}}(\pi_P) \notin F_{r_{al}}$.

(1) There exists $r_{al} \in R$, and $\max_{r_{al} \in R} \{\sup F_{r_{al}} - \inf F_{r_{al}}\} < \min_{r_{al} \in R} \{\inf S_{r_{al}} - \sup F_{r_{al}}\} \wedge \max_{r_{al} \in R} \{\sup S_{r_{al}} - \inf S_{r_{al}}\} < \min_{r_{al} \in R} \{\inf S_{r_{al}} - \sup F_{r_{al}}\}$;

(2) There exists a policy π^* such that $\forall r_{al} \in R$, $U_{r_{al}}(\pi^*) \in S_{r_{al}}$, and $\forall r_{mis} \in R$, $\max_{\pi \in \Pi} U_{r_{mis}}(\pi) - U_{r_{mis}}(\pi^*) < \min_{r_{al} \in R} \{\inf S_{r_{al}} - \sup F_{r_{al}}\}$.

Proof. Suppose the conditions are met, and a policy π_1 satisfies the property described in conditions 2). Then for any policy $\pi_2 \in MinimaxRegret(R)$, if π_2 does not satisfy the mentioned property, there exists a task-aligned reward function $r_{al} \in R_E$ such that $U_{r_{al}}(\pi_2) \in F_{r_{al}}$. In this case $Regret(\pi_2, r_{al}) = \max_{\pi \in \Pi} U_{r_{al}}(\pi) - U_{r_{al}}(\pi_2) \geq \inf S_{r_{al}} - \sup F_{r_{al}} \geq \min_{r'_{al} \in R_E} \left\{\inf S_{r'_{al}} - \sup F_{r'_{al}}\right\}$. However, for π_1, it holds for any task-aligned reward function $\hat{r}_{al} \in R_E$ that $Regret(\pi_2, \hat{r}_{al}) \leq \sup S_{\hat{r}_{al}} - \inf S_{\hat{r}_{al}} < \min_{r'_{al} \in R_E} \left\{\inf S_{r'_{al}} - \sup F_{r'_{al}}\right\}$, and it also holds for any misaligned reward function $r_{mis} \in R_E$ that $Regret(\pi_2, r_{mis}) = \max_{\pi \in \Pi} U_{r_{mis}}(\pi) - U_{r_{mis}}(\pi_2) < \min_{r'_{al} \in R_E} \{\inf S_{r_{al}} - \sup F_{r_{al}}\}$. Hence, $Regret(\pi_2, r_{al}) < Regret(\pi_1, r_{al})$, contradicting $\pi_1 \in MiniRegret$. We complete the proof. □

Corollary 1.(Task-Success Guarantee) Assume that Condition (1) in Theorem 1 is satisfied. If there exists a policy π^* such that $\forall r \in R$, $\max_{\pi} U_r(\pi) - U_r(\pi^*) < \min_{r_{al} \in R} \{\sup S_{r_{al}} - \inf S_{r_{al}}\}$, then the optimal protagonist policy $\pi_P := MinimaxRegret(R)$ satisfies that $\forall r_{al} \in R$, $U_{r_{al}}(\pi) \in S_{r_{al}}$.

Proof. Since $\max_{r\in R} \max_{\pi} U_r(\pi) - U_r(\pi_P) \leq \max_{r\in R} \max_{\pi} U_r(\pi) - U_r(\pi^*) < \min_{r_{al}\in R} \{\sup S_{r_{al}} - \inf S_{r_{al}}\}$, we can conclude that for any $r_{al} \in R$, $\max_{\pi} U_{r_{al}}(\pi) - U_{r_{al}}(\pi_P) \leq \{\sup S_{r_{al}} - \inf S_{r_{al}}$, in other words, $U_{r_{al}}(\pi_P) \in S_{r_{al}}$. The proof is complete. □

Theorem 2.)(Task-Failure Avoidance) If the following conditions (1) (2) hold for $R_{E,\delta}$, then the optimal protagonist policy $\pi_P := MinimaxRegret(R_{E,\delta})$ satisfies that $\forall r_{al} \in R_{E,\delta}$, $U_{r_{al}}(\pi) \notin F_{r_{al}}$.

(1) The condition (1) in Theorem 1 holds

(2) $\forall r_{al} \in R_{E,\delta}$, $L_{r_{al}} \cdot W_E - \delta \leq \sup S_{r_{al}} - \inf S_{r_{al}}$ and $\forall r_{mis} \in R_{E,\delta}$, $L_{r_{mis}} \cdot W_E - \delta < \min_{r_{al}\in R_{E,\delta}} \{\inf S_{r_{al}} - \sup F_{r_{al}}\}$.

Proof. Without loss of generality, we consider $U_r(\pi) = \mathbb{E}_{\tau\sim\pi}[r(\tau)]$. Since $W_E \triangleq \min_{\pi\in\Pi} W_1(\pi, E) = \frac{1}{K} \sup_{\|r\|_L \leq K} U_r(E) - U_r(\pi)$ for any $K > 0$, let π^* be the policy that achieves the minimality in W_E. Then for any $r_{al} \in R$, the term $L_{r_{al}} \cdot W_E - \delta \geq L_{r_{al}} \cdot \frac{1}{L_{r_{al}}} \sup_{\|r\|_L \leq L_{r_{al}}} U_r(E) - U_r(\pi) \geq U_{r_{al}}(E) - U_{r_{al}}(\pi)$. Hence, for all $r_{al} \in R$, $U_{r_{al}}(E) - U_{r_{al}}(\pi) < \sup S_{r_{al}} - \inf S_{r_{al}}$, i.e., $U_{r_{al}}(\pi^*) \in S_{r_{al}}$. Likewise, $L_{r_{mis}} \cdot W_E - \delta < \min_{r_{al}\in R_{E,\delta}} \{\inf S_{r_{al}} - \sup F_{r_{al}}\}$ indicates that for all $r_{mis} \in R$, $U_{r_{mis}}(E) - U_{r_{mis}}(\pi) < \min_{r_{al}\in R_{E,\delta}} \{\inf S_{r_{al}} - \sup F_{r_{al}}\}$. Then, we have recovered the condition (2) in Theorem 1. As a result, we deliver the same guarantees in Theorem 1. □

Corollary 2.(Task-Success Guarantee) Assume that the condition (1) in Theorem 1 holds for $R_{E,\delta}$. If for any $r \in R_{E,\delta}$, $L_r \cdot W_E - \delta \leq \min_{r_{al}\in R} \{\sup S_{r_{al}} - \inf S_{r_{al}}\}$, then the optimal protagonist policy $\pi_P = MinimaxRegret(R_{E,\delta})$ satisfies $\forall r_{al} \in R_{E,\delta}$, $U_{r_{al}}(\pi) \in S_{r_{al}}$.

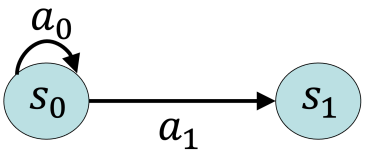
Proof. Again, we let π^* be the policy that achieves the minimality in W_E. Then, we have $L_r \cdot W_E - \delta \geq L_r \cdot \frac{1}{L_r} \sup_{\|r\|_L \leq L_r} U_r(E) - U_r(\pi^*) \geq U_r(E) - U_r(\pi^*)$ for any $r \in R$. We have recovered the condition in Corollary 1. The proof is complete. □

Theorem 3. Let the IRL loss be in the form of $J_{IRL}(\pi, r) := U_r(\pi) - U_r(\pi_E)$ for some $U_r(\pi)$. If $\arg\min_{r\in R} \max_{\pi\in\Pi} J_{IRL}(r_E, \pi)$ can reach Nash Equilibrium with a reward function set R_E and a policy set Π_E, then Π_E equals the set of solutions to $MinimiaxRegret$.

Proof. The reward function set R_E and the policy set Π_E achieving Nash Equilibrium for $\arg\min_{r\in R} \max_{\pi\in\Pi} J_{IRL}(r_E, \pi)$ indicates that for any $r \in R_E$, $\pi \in \Pi_E$, $\pi \in \arg\max_{\pi\in\Pi} U_r(\pi) - U_r(\pi_E)$. Then Π_E will be the solution to $\arg\max_{\pi_P\in\Pi} \min_{r\in R_E} \left\{ \max_{\pi_A\in\Pi} U_r(\pi_A) - U_r(\pi_E) \right\} - (U_r(\pi_P) - U_r(\pi_E))$ because the policies in Π_E achieve zero regret. Then Lemma 5 states that Π_E will also be the solution to $\arg\max_{\pi_P\in\Pi} \min_{r\in R_E} \left\{ \max_{\pi_A\in\Pi} U_r(\pi_A) \right\} - U_r(\pi_P)$. We finish the proof. □

A.5 Running Example

Example 1 In a two-state transition system where s_0 is the initial state, and s_1 is an absorbing state, an agent can choose action a_0 at state s_0 to stay at s_0 or choose a_1 to reach s_1. Any agent can start from s_0 and choose actions for 5 timesteps. The task is to learn a stochastic policy to reach s_1. Expert only demonstrates one trajectory $E = \{\tau_E = (s_0, a_1, s_1, s_1, s_1, s_1)\}$, i.e., choose a_1 at s_0 and then stay in s_1 for the rest 4 steps. The convex combinations of two basis functions r_1 and r_2 constitute the hypothesis set of reward functions $R = \{r | r(s, a) = \alpha \cdot r_1(s, a) + (1 - \alpha) \cdot r_2(s, a), \alpha \in [0, 1]\}$ where $r_1(s, a) \equiv 4$ constantly, and $r_2(s_0, a_0) = r_2(s_0, a_1) = 0, r_2(s_1, \cdot) = 5$. We now prove that when applying IRL to learn a reward function from E, any convex combination of r_1 and r_2 is an



Expert: $s_1 \xrightarrow{a_1} s_1 \rightarrow s_1 \rightarrow s_1 \rightarrow s_1$

	(s_0, a_0)	(s_0, a_1)	$(s_1, \cdot)$
r_1	4	4	4
r_2	0	0	5

Figure 5: ss

optimal solution, i.e., $R_E = R$. Furthermore, we prove that $MinimaxRegret(R)$ will produce $\pi_P(a_1|s_0) = 1$ which is the desired solution.

Proof. For any policy π and reward function $r \in R$, the policy performance $U_r(\pi)$ can be represented as follows.

$$\begin{aligned} U_r(\pi) &= 4\alpha\pi(a_0|s_0)^t + \sum_{t=0}^{t}(4\alpha t + 4\alpha + (4\alpha + 5(1-\alpha))(4-t)) \cdot \pi(a_0|s_0)^t(1-\pi(a_0|s_0)) \\ &= 20 - 5(1-\alpha)\sum_{t=1}^{4}\pi(a_0|s_0)^4 \end{aligned} \tag{15}$$

Then for any $\alpha \in [0, 1)$, $\max_\pi U_r(\pi) = 20$, and the optimal policy is $\pi^*(a_0|s_0) = 0, \pi^*(a_1|s_0) = 1$. For $\alpha = 1, U_r(\pi) = 20$ constantly for any π. For expert demonstration $E, U_r(E) = 4\alpha{\cdot}5{+}5(1{-}\alpha){\cdot}4 \equiv 20$. Therefore, any $r = \alpha r_1 + (1-\alpha)r_2$ with $\alpha \in [0, 1]$ can be the optimal solution to IRL since they all induce the same performance gap $U_r(E) - \max_\pi U_r(\pi) = 0$. This ambiguity in choosing reward functions can cause reward misalignment because the reward function $r = r_1$ induced by $\alpha = 1$ violates the criterion of task-reward alignment in Definition 1: achieving optimality under r_1 does not guarantee task success since all policies are optimal under r_1. On the other hand, it is obvious that $MinimaxRegret(R)$ induces a policy $\pi_P(a_1|s_0) = 1$. Because this policy has the lowest worst-case regret $Regret(\pi_P, r) = \max_{r \in R} \max_{\pi_A} U_r(\pi_A) - U_r(\pi_P)) = 0$. □

B Approach to Solving MinimaxRegret

In this section, we develop a series of theories that lead to two bounds of the Protagonist Antagonist Induced Regret. By using those bounds, we formulate objective functions for solving Imitation Learning problems with PAGAR.

B.1 Protagonist Antagonist Induced Regret Bounds

Our theories are inspired by the on-policy policy improvement methods in Schulman et al. (2015). The theories in Schulman et al. (2015) are under the setting where entropy regularizer is not considered. In our implementation, we always consider entropy regularized RL of which the objective is to learn a policy that maximizes $J_{RL}(\pi; r) = \eta_r(\pi) + \mathcal{H}(\pi)$. Also, since we use GAN-based IRL algorithms, the learned reward function r as proved by Fu et al. (2018) is a distribution. Moreover, it is also proved in Fu et al. (2018) that a policy π being optimal under r indicates that $\log \pi \equiv r \equiv \mathcal{A}_\pi$. We omit the proof and let the reader refer to Fu et al. (2018) for details. Although all our theories are about the relationship between the Protagonist Antagonist Induced Regret and the soft advantage function $\mathcal{A}_\pi$, the equivalence between $\mathcal{A}_\pi$ and r allows us to use the theories to formulate our

reward optimization objective functions. To start off, we denote the reward function to be optimized as r. Given the intermediate learned reward function r, we study the Protagonist Antagonist Induced Regret between two policies π_1 and π_2.

Lemma 6. *Given a reward function r and a pair of policies π_1 and π_2,*

$$\eta_r(\pi_1) - \eta_r(\pi_2) = \mathop{\mathbb{E}}_{\tau\sim\pi_1}\left[\sum_{t=0}^{\infty}\gamma^t \mathcal{A}_{\pi_2}(s^{(t)}, a^{(t)})\right] + \mathop{\mathbb{E}}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t \mathcal{H}\left(\pi_2(\cdot|s^{(t)})\right)\right] \tag{16}$$

Proof. This proof follows the proof of Lemma 1 in Schulman et al. (2015) where RL is not entropy-regularized. For entropy-regularized RL, since $\mathcal{A}_\pi(s, a^{(t)}) = \mathop{\mathbb{E}}_{s'\sim\mathcal{P}(\cdot|s,a^{(t)})}\big[r(s, a^{(t)}) + \gamma\mathcal{V}_\pi(s') - \mathcal{V}_\pi(s)\big]$,

$$\begin{aligned}
& \mathop{\mathbb{E}}_{\tau\sim\pi_1}\left[\sum_{t=0}^{\infty}\gamma^t \mathcal{A}_{\pi_2}(s^{(t)}, a^{(t)})\right] \\
= & \mathop{\mathbb{E}}_{\tau\sim\pi_1}\left[\sum_{t=0}^{\infty}\gamma^t \left(r(s^{(t+1)}, a^{(t+1)}) + \gamma\mathcal{V}_{\pi_2}(s^{(t+1)}) - \mathcal{V}_{\pi_2}(s^{(t)})\right)\right] \\
= & \mathop{\mathbb{E}}_{\tau\sim\pi_1}\left[\sum_{t=0}^{\infty}\gamma^t r(s^{(t)}, a^{(t)}) - \mathcal{V}_{\pi_2}(s^{(0)})\right] \\
= & \mathop{\mathbb{E}}_{\tau\sim\pi_1}\left[\sum_{t=0}^{\infty}\gamma^t r(s^{(t)}, a^{(t)})\right] - \mathop{\mathbb{E}}_{s^{(0)}\sim d_0}\left[\mathcal{V}_{\pi_2}(s^{(0)})\right] \\
= & \mathop{\mathbb{E}}_{\tau\sim\pi_1}\left[\sum_{t=0}^{\infty}\gamma^t r(s^{(t)}, a^{(t)})\right] - \mathop{\mathbb{E}}_{\tau\sim\pi_2}\left[\sum_{t=0}^{\infty}\gamma^t r(s^{(t)}, a^{(t)}) + \mathcal{H}\left(\pi_2(\cdot|s^{(t)})\right)\right] \\
= & \eta_r(\pi_1) - \eta_r(\pi_2) - \mathop{\mathbb{E}}_{\tau\sim\pi_2}\left[\sum_{t=0}^{\infty}\gamma^t \mathcal{H}\left(\pi_2(\cdot|s^{(t)})\right)\right] \\
= & \eta_r(\pi_1) - \eta_r(\pi_2) - \mathcal{H}(\pi_2)
\end{aligned}$$

□

Remark 1. Lemma 6 confirms that $\mathop{\mathbb{E}}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t \mathcal{A}_\pi(s^{(t)}, a^{(t)})\right] = \eta_r(\pi) - \eta_r(\pi) + \mathcal{H}(\pi) = \mathcal{H}(\pi)$.

We follow Schulman et al. (2015) and denote $\Delta\mathcal{A}(s) = \mathop{\mathbb{E}}_{a\sim\pi_1(\cdot|s)}[\mathcal{A}_{\pi_2}(s, a)] - \mathop{\mathbb{E}}_{a\sim\pi_2(\cdot|s)}[\mathcal{A}_{\pi_2}(s, a)]$ as the difference between the expected advantages of following π_2 after choosing an action respectively by following policy π_1 and π_2 at any state s. Although the setting of Schulman et al. (2015) differs from ours by having the expected advantage $\mathop{\mathbb{E}}_{a\sim\pi_2(\cdot|s)}[\mathcal{A}_{\pi_2}(s, a)]$ equal to 0 due to the absence of entropy regularization, the following definition and lemmas from Schulman et al. (2015) remain valid in our setting.

Definition 7. Schulman et al. (2015), the protagonist policy π_1 and the antagonist policy π_2) are α-coupled if they defines a joint distribution over $(a, \tilde{a}) \in \mathbb{A} \times \mathbb{A}$, such that $Prob(a \neq \tilde{a}|s) \leq \alpha$ for all s.

Lemma 7. *Schulman et al. (2015) Given that the protagonist policy π_1 and the antagonist policy π_2 are α-coupled, then for all state s,*

$$|\Delta\mathcal{A}(s)| \leq 2\alpha \max_a |\mathcal{A}_{\pi_2}(s, a)| \tag{17}$$

Lemma 8. *Schulman et al. (2015) Given that the protagonist policy π_1 and the antagonist policy π_2 are α-coupled, then*

$$\left|\mathop{\mathbb{E}}_{s^{(t)}\sim\pi_1}\left[\Delta\mathcal{A}(s^{(t)})\right] - \mathop{\mathbb{E}}_{s^{(t)}\sim\pi_2}\left[\Delta\mathcal{A}(s^{(t)})\right]\right| \leq 4\alpha(1-(1-\alpha)^t)\max_{s,a}|\mathcal{A}_{\pi_2}(s, a)| \tag{18}$$

Lemma 9. *Given that the protagonist policy π_1 and the antagonist policy π_2 are α-coupled, then*

$$\mathop{\mathbb{E}}_{\substack{s^{(t)}\sim\pi_1\\a^{(t)}\sim\pi_2}}\left[\mathcal{A}_{\pi_2}(s^{(t)},a^{(t)})\right]-\mathop{\mathbb{E}}_{\substack{s^{(t)}\sim\pi_2\\a^{(t)}\sim\pi_2}}\left[\mathcal{A}_{\pi_2}(s^{(t)},a^{(t)})\right]\leq 2(1-(1-\alpha)^t)\max_{(s,a)}|\mathcal{A}_{\pi_2}(s,a)| \tag{19}$$

Proof. The proof is similar to that of Lemma 8 in Schulman et al. (2015). Let n_t be the number of times that $a^{(t')}\sim\pi_1$ does not equal $a^{(t')}\sim\pi_2$ for $t'<t$, i.e., the number of times that π_1 and π_2 disagree before timestep t. Then for $s^{(t)}\sim\pi_1$, we have the following.

$$\begin{aligned}&\mathop{\mathbb{E}}_{s^{(t)}\sim\pi_1}\left[\mathop{\mathbb{E}}_{a^{(t)}\sim\pi_2}\left[\mathcal{A}_{\pi_2}(s^{(t)},a^{(t)})\right]\right]\\ =\;& P(n_t=0)\mathop{\mathbb{E}}_{\substack{s^{(t)}\sim\pi_1\\n_t=0}}\left[\mathop{\mathbb{E}}_{a^{(t)}\sim\pi_2}\left[\mathcal{A}_{\pi_2}(s^{(t)},a^{(t)})\right]\right]+P(n_t>0)\mathop{\mathbb{E}}_{\substack{s^{(t)}\sim\pi_1\\n_t>0}}\left[\mathop{\mathbb{E}}_{a^{(t)}\sim\pi_2}\left[\mathcal{A}_{\pi_2}(s^{(t)},a^{(t)})\right]\right]\end{aligned}$$

The expectation decomposes similarly for $s^{(t)}\sim\pi_2$.

$$\begin{aligned}&\mathop{\mathbb{E}}_{\substack{s^{(t)}\sim\pi_2\\a^{(t)}\sim\pi_2}}\left[\mathcal{A}_{\pi_2}(s^{(t)},a^{(t)})\right]\\ =\;& P(n_t=0)\mathop{\mathbb{E}}_{\substack{s^{(t)}\sim\pi_2\\a^{(t)}\sim\pi_2\\n_t=0}}\left[\mathcal{A}_{\pi_2}(s^{(t)},a^{(t)})\right]+P(n_t>0)\mathop{\mathbb{E}}_{\substack{s^{(t)}\sim\pi_2\\a^{(t)}\sim\pi_2\\n_t>0}}\left[\mathcal{A}_{\pi_2}(s^{(t)},a^{(t)})\right]\end{aligned}$$

When computing $\mathop{\mathbb{E}}_{s^{(t)}\sim\pi_1}\left[\mathop{\mathbb{E}}_{a^{(t)}\sim\pi_2}\left[\mathcal{A}_{\pi_2}(s^{(t)},a^{(t)})\right]\right]-\mathop{\mathbb{E}}_{\substack{s^{(t)}\sim\pi_2\\a^{(t)}\sim\pi_2}}\left[\mathcal{A}_{\pi_2}(s^{(t)},a^{(t)})\right]$, the terms with $n_t=0$ cancel each other because $n_t=0$ indicates that π_1 and π_2 agreed on all timesteps less than t. That leads to the following.

$$\begin{aligned}&\mathop{\mathbb{E}}_{s^{(t)}\sim\pi_1}\left[\mathop{\mathbb{E}}_{a^{(t)}\sim\pi_2}\left[\mathcal{A}_{\pi_2}(s^{(t)},a^{(t)})\right]\right]-\mathop{\mathbb{E}}_{\substack{s^{(t)}\sim\pi_2\\a^{(t)}\sim\pi_2}}\left[\mathcal{A}_{\pi_2}(s^{(t)},a^{(t)})\right]\\ =\;& P(n_t>0)\mathop{\mathbb{E}}_{\substack{s^{(t)}\sim\pi_1\\n_t>0}}\left[\mathop{\mathbb{E}}_{a^{(t)}\sim\pi_2}\left[\mathcal{A}_{\pi_2}(s^{(t)},a^{(t)})\right]\right]-P(n_t>0)\mathop{\mathbb{E}}_{\substack{s^{(t)}\sim\pi_2\\a^{(t)}\sim\pi_2\\n_t>0}}\left[\mathcal{A}_{\pi_2}(s^{(t)},a^{(t)})\right]\end{aligned}$$

By definition of α, the probability of π_1 and π_2 agreeing at timestep t' is no less than $1-\alpha$. Hence, $P(n_t>0)\leq 1-(1-\alpha^t)^t$. Hence, we have the following bound.

$$\begin{aligned}&\left|\mathop{\mathbb{E}}_{s^{(t)}\sim\pi_1}\left[\mathop{\mathbb{E}}_{a^{(t)}\sim\pi_2}\left[\mathcal{A}_{\pi_2}(s^{(t)},a^{(t)})\right]\right]-\mathop{\mathbb{E}}_{\substack{s^{(t)}\sim\pi_2\\a^{(t)}\sim\pi_2}}\left[\mathcal{A}_{\pi_2}(s^{(t)},a^{(t)})\right]\right|\\ =\;&\left|P(n_t>0)\mathop{\mathbb{E}}_{\substack{s^{(t)}\sim\pi_1\\n_t>0}}\left[\mathop{\mathbb{E}}_{a^{(t)}\sim\pi_2}\left[\mathcal{A}_{\pi_2}(s^{(t)},a^{(t)})\right]\right]-P(n_t>0)\mathop{\mathbb{E}}_{\substack{s^{(t)}\sim\pi_2\\a^{(t)}\sim\pi_2\\n_t>0}}\left[\mathcal{A}_{\pi_2}(s^{(t)},a^{(t)})\right]\right|\\ \leq\;& P(n_t>0)\left(\left|\mathop{\mathbb{E}}_{\substack{s^{(t)}\sim\pi_1\\a^{(t)}\sim\pi_2\\n_t\geq 0}}\left[\mathcal{A}_{\pi_2}(s^{(t)},a^{(t)})\right]\right|+\left|\mathop{\mathbb{E}}_{\substack{s^{(t)}\sim\pi_2\\a^{(t)}\sim\pi_2\\n_t>0}}\left[\mathcal{A}_{\pi_2}(s^{(t)},a^{(t)})\right]\right|\right)\\ \leq\;& 2(1-(1-\alpha)^t)\max_{(s,a)}|\mathcal{A}_{\pi_2}(s,a)|\end{aligned} \tag{20}$$

□

The preceding lemmas lead to the proof for Theorem 4 in the main text.

Theorem 4. Suppose that π_2 is the optimal policy in terms of entropy regularized RL under r. Let $\alpha = \max_s D_{TV}(\pi_1(\cdot|s), \pi_2(\cdot|s))$, $\epsilon = \max_{s,a} |\mathcal{A}_{\pi_2}(s, a^{(t)})|$, and $\Delta\mathcal{A}(s) = \mathop{\mathbb{E}}_{a\sim\pi_1}[\mathcal{A}_{\pi_2}(s,a)] - \mathop{\mathbb{E}}_{a\sim\pi_2}[\mathcal{A}_{\pi_2}(s,a)]$. For any policy π_1, the following bounds hold.

$$\left|\eta_r(\pi_1) - \eta_r(\pi_2) - \sum_{t=0}^{\infty}\gamma^t \mathop{\mathbb{E}}_{s^{(t)}\sim\pi_1}\left[\Delta\mathcal{A}(s^{(t)})\right]\right| \leq \frac{2\alpha\gamma\epsilon}{(1-\gamma)^2} \tag{21}$$

$$\left|\eta_r(\pi_1) - \eta_r(\pi_2) - \sum_{t=0}^{\infty}\gamma^t \mathop{\mathbb{E}}_{s^{(t)}\sim\pi_2}\left[\Delta\mathcal{A}(s^{(t)})\right]\right| \leq \frac{2\alpha\gamma(2\alpha+1)\epsilon}{(1-\gamma)^2} \tag{22}$$

Proof. We first leverage Lemma 6 to derive Eq.23. Note that since π_2 is optimal under r, Remark 1 confirmed that $\mathcal{H}(\pi_2) = -\sum_{t=0}^{\infty}\gamma^t \mathop{\mathbb{E}}_{s^{(t)}\sim\pi_2}\left[\mathop{\mathbb{E}}_{a^{(t)}\sim\pi_2}\left[\mathcal{A}_{\pi_2}(s^{(t)}, a^{(t)})\right]\right]$.

$$\begin{aligned}
& \eta_r(\pi_1) - \eta_r(\pi_2) \\
= \; & (\eta_r(\pi_1) - \eta_r(\pi_2) - \mathcal{H}(\pi_2)) + \mathcal{H}(\pi_2) \\
= \; & \mathop{\mathbb{E}}_{\tau\sim\pi_1}\left[\sum_{t=0}^{\infty}\gamma^t\mathcal{A}_{\pi_2}(s^{(t)}, a^{(t)})\right] + \mathcal{H}(\pi_2) \\
= \; & \mathop{\mathbb{E}}_{\tau\sim\pi_1}\left[\sum_{t=0}^{\infty}\gamma^t\mathcal{A}_{\pi_2}(s^{(t)}, a^{(t)})\right] - \sum_{t=0}^{\infty}\gamma^t \mathop{\mathbb{E}}_{s^{(t)}\sim\pi_2}\left[\mathop{\mathbb{E}}_{a^{(t)}\sim\pi_2}\left[\mathcal{A}_{\pi_2}(s^{(t)}, a^{(t)})\right]\right] \\
= \; & \sum_{t=0}^{\infty}\gamma^t \mathop{\mathbb{E}}_{s^{(t)}\sim\pi_1}\left[\mathop{\mathbb{E}}_{a^{(t)}\sim\pi_1}\left[\mathcal{A}_{\pi_2}(s^{(t)}, a^{(t)})\right] - \mathop{\mathbb{E}}_{a^{(t)}\sim\pi_2}\left[\mathcal{A}_{\pi_2}(s^{(t)}, a^{(t)})\right]\right] + \\
& \sum_{t=0}^{\infty}\gamma^t\left(\mathop{\mathbb{E}}_{s^{(t)}\sim\pi_1}\left[\mathop{\mathbb{E}}_{a^{(t)}\sim\pi_2}\left[\mathcal{A}_{\pi_2}(s^{(t)}, a^{(t)})\right]\right] - \mathop{\mathbb{E}}_{s^{(t)}\sim\pi_2}\left[\mathop{\mathbb{E}}_{a^{(t)}\sim\pi_2}\left[\mathcal{A}_{\pi_2}(s^{(t)}, a^{(t)})\right]\right]\right) \\
= \; & \sum_{t=0}^{\infty}\gamma^t \mathop{\mathbb{E}}_{s^{(t)}\sim\pi_1}\left[\Delta\mathcal{A}(s^{(t)})\right] + \\
& \sum_{t=0}^{\infty}\gamma^t\left(\mathop{\mathbb{E}}_{s^{(t)}\sim\pi_1}\left[\mathop{\mathbb{E}}_{a^{(t)}\sim\pi_2}\left[\mathcal{A}_{\pi_2}(s^{(t)}, a^{(t)})\right]\right] - \mathop{\mathbb{E}}_{s^{(t)}\sim\pi_2}\left[\mathop{\mathbb{E}}_{a^{(t)}\sim\pi_2}\left[\mathcal{A}_{\pi_2}(s^{(t)}, a^{(t)})\right]\right]\right)
\end{aligned} \tag{23}$$

We switch terms between Eq.23 and $\eta_r(\pi_1) - \eta_r(\pi_2)$, then use Lemma 9 to derive Eq.24.

$$\begin{aligned}
& \left|\eta_r(\pi_1) - \eta_r(\pi_2) - \sum_{t=0}^{\infty}\gamma^t \mathop{\mathbb{E}}_{s^{(t)}\sim\pi_1}\left[\Delta\mathcal{A}(s^{(t)})\right]\right| \\
= \; & \left|\sum_{t=0}^{\infty}\gamma^t\left(\mathop{\mathbb{E}}_{s^{(t)}\sim\pi_1}\left[\mathop{\mathbb{E}}_{a^{(t)}\sim\pi_2}\left[\mathcal{A}_{\pi_2}(s^{(t)}, a^{(t)})\right]\right] - \mathop{\mathbb{E}}_{s^{(t)}\sim\pi_2}\left[\mathop{\mathbb{E}}_{a^{(t)}\sim\pi_2}\left[\mathcal{A}_{\pi_2}(s^{(t)}, a^{(t)})\right]\right]\right)\right| \\
\leq \; & \sum_{t=0}^{\infty}\gamma^t \cdot 2\max_{(s,a)}|\mathcal{A}_{\pi_2}(s,a)| \cdot (1-(1-\alpha)^t) \leq \frac{2\alpha\gamma\max_{(s,a)}|\mathcal{A}_{\pi_2}(s,a)|}{(1-\gamma)^2}
\end{aligned} \tag{24}$$

Alternatively, we can expand $\eta_r(\pi_2) - \eta_r(\pi_1)$ into Eq.25. During the process, $\mathcal{H}(\pi_2)$ is converted into $-\sum_{t=0}^{\infty}\gamma^t \mathop{\mathbb{E}}_{s^{(t)}\sim\pi_2}\left[\mathop{\mathbb{E}}_{a^{(t)}\sim\pi_2}\left[\mathcal{A}_{\pi_2}(s^{(t)},a^{(t)})\right]\right]$.

$$
\begin{aligned}
& \eta_r(\pi_1) - \eta_r(\pi_2) \\
= \;& (\eta_r(\pi_1) - \eta_r(\pi_2) - \mathcal{H}(\pi_2)) + \mathcal{H}(\pi_2) \\
= \;& \mathop{\mathbb{E}}_{\tau\sim\pi_1}\left[\sum_{t=0}^{\infty}\gamma^t \mathcal{A}_{\pi_2}(s^{(t)},a^{(t)})\right] + \mathcal{H}(\pi_2) \\
= \;& \sum_{t=0}^{\infty}\gamma^t \mathop{\mathbb{E}}_{s^{(t)}\sim\pi_1}\left[\mathop{\mathbb{E}}_{a^{(t)}\sim\pi_1}\left[\mathcal{A}_{\pi_2}(s^{(t)},a^{(t)})\right]\right] + \mathcal{H}(\pi_2) \\
= \;& \sum_{t=0}^{\infty}\gamma^t \mathop{\mathbb{E}}_{s^{(t)}\sim\pi_1}\left[\Delta A(s^{(t)}) + \mathop{\mathbb{E}}_{a^{(t)}\sim\pi_2}\left[\mathcal{A}_{\pi_2}(s^{(t)},a^{(t)})\right]\right] + \mathcal{H}(\pi_2) \\
= \;& \sum_{t=0}^{\infty}\gamma^t \mathop{\mathbb{E}}_{s^{(t)}\sim\pi_2}\left[\mathop{\mathbb{E}}_{a^{(t)}\sim\pi_1}\left[\mathcal{A}_{\pi_2}(s^{(t)},a^{(t)})\right] - \mathop{\mathbb{E}}_{a^{(t)}\sim\pi_2}\left[\mathcal{A}_{\pi_2}(s^{(t)},a^{(t)})\right] - \Delta\mathcal{A}(s^{(t)})\right] + \\
& \mathop{\mathbb{E}}_{s^{(t)}\sim\pi_1}\left[\Delta\mathcal{A}(s^{(t)}) + \mathop{\mathbb{E}}_{a^{(t)}\sim\pi_2}\left[\mathcal{A}_{\pi_2}(s^{(t)},a^{(t)})\right]\right] - \mathop{\mathbb{E}}_{\substack{s^{(t)}\sim\pi_2 \\ a^{(t)}\sim\pi_2}}\left[\mathcal{A}_{\pi_2}(s^{(t)},a^{(t)})\right] \\
= \;& \sum_{t=0}^{\infty}\gamma^t\left(\mathop{\mathbb{E}}_{s^{(t)}\sim\pi_1}\left[\mathop{\mathbb{E}}_{a^{(t)}\sim\pi_2}\left[\mathcal{A}_{\pi_2}(s^{(t)},a^{(t)})\right]\right] - 2\mathop{\mathbb{E}}_{s^{(t)}\sim\pi_2}\left[\mathop{\mathbb{E}}_{a^{(t)}\sim\pi_2}\left[\mathcal{A}_{\pi_2}(s^{(t)},a^{(t)})\right]\right]\right) + \\
& \sum_{t=0}^{\infty}\gamma^t\left(\mathop{\mathbb{E}}_{s^{(t)}\sim\pi_2}\left[\mathop{\mathbb{E}}_{a^{(t)}\sim\pi_1}\left[\mathcal{A}_{\pi_2}(s^{(t)},a^{(t)})\right]\right] - (\mathop{\mathbb{E}}_{s^{(t)}\sim\pi_2}\left[\Delta\mathcal{A}(s^{(t)})\right] - \mathop{\mathbb{E}}_{s^{(t)}\sim\pi_1}\left[\Delta\mathcal{A}(s^{(t)})\right])\right)
\end{aligned}
\tag{25}
$$

We switch terms between Eq.25 and $\eta_r(\pi_1) - \eta_r(\pi_2)$, then base on Lemma 8 and 9 to derive the inequality in Eq.26.

$$
\begin{aligned}
& \left|\eta_r(\pi_1) - \eta_r(\pi_2) - \sum_{t=0}^{\infty}\gamma^t \mathop{\mathbb{E}}_{s^{(t)}\sim\pi_2}\left[\Delta\mathcal{A}_{\pi}(s^{(t)},a^{(t)})\right]\right| \\
= \;& \Bigg|\eta_r(\pi_1) - \eta_r(\pi_2) - \\
& \quad \sum_{t=0}^{\infty}\gamma^t\left(\mathop{\mathbb{E}}_{s^{(t)}\sim\pi_2}\left[\mathop{\mathbb{E}}_{a^{(t)}\sim\pi_1}\left[\mathcal{A}_{\pi_2}(s^{(t)},a^{(t)})\right]\right] - \mathop{\mathbb{E}}_{s^{(t)}\sim\pi_2}\left[\mathop{\mathbb{E}}_{a^{(t)}\sim\pi_2}\left[\mathcal{A}_{\pi_2}(s^{(t)},a^{(t)})\right]\right]\right)\Bigg| \\
= \;& \Bigg|\sum_{t=0}^{\infty}\gamma^t\left(\mathop{\mathbb{E}}_{s^{(t)}\sim\pi_2}\left[\Delta\mathcal{A}(s^{(t)})\right] - \mathop{\mathbb{E}}_{s^{(t)}\sim\pi_1}\left[\Delta\mathcal{A}(s^{(t)})\right]\right) - \\
& \quad \sum_{t=0}^{\infty}\gamma^t\left(\mathop{\mathbb{E}}_{s^{(t)}\sim\pi_1}\left[\mathop{\mathbb{E}}_{a^{(t)}\sim\pi_2}\left[\mathcal{A}_{\pi_2}(s^{(t)},a^{(t)})\right]\right] - \mathop{\mathbb{E}}_{s^{(t)}\sim\pi_2}\left[\mathop{\mathbb{E}}_{a^{(t)}\sim\pi_2}\left[\mathcal{A}_{\pi_2}(s^{(t)},a^{(t)})\right]\right]\right)\Bigg| \\
\leq \;& \left|\sum_{t=0}^{\infty}\gamma^t\left(\mathop{\mathbb{E}}_{s^{(t)}\sim\pi_2}\left[\Delta\mathcal{A}(s^{(t)})\right] - \mathop{\mathbb{E}}_{s^{(t)}\sim\pi_1}\left[\Delta\mathcal{A}(s^{(t)})\right]\right)\right| + \\
& \quad \left|\sum_{t=0}^{\infty}\gamma^t\left(\mathop{\mathbb{E}}_{s^{(t)}\sim\pi_1}\left[\mathop{\mathbb{E}}_{a^{(t)}\sim\pi_2}\left[\mathcal{A}_{\pi_2}(s^{(t)},a^{(t)})\right]\right] - \mathop{\mathbb{E}}_{s^{(t)}\sim\pi_2}\left[\mathop{\mathbb{E}}_{a^{(t)}\sim\pi_2}\left[\mathcal{A}_{\pi_2}(s^{(t)},a^{(t)})\right]\right]\right)\right| \\
\leq \;& \sum_{t=0}^{\infty}\gamma^t\left((1-(1-\alpha)^t)(4\alpha\max_{s,a}|\mathcal{A}_{\pi_2}(s,a)| + 2\max_{(s,a)}|\mathcal{A}_{\pi_2}(s,a)|)\right) \\
\leq \;& \frac{2\alpha\gamma(2\alpha+1)\max_{s,a}|\mathcal{A}_{\pi_2}(s,a)|}{(1-\gamma)^2}
\end{aligned}
\tag{26}
$$

It is stated in Schulman et al. (2015) that $\max_s D_{TV}(\pi_2(\cdot|s), \pi_1(\cdot|s)) \leq \alpha$. Hence, by letting $\alpha := \max_s D_{TV}(\pi_2(\cdot|s), \pi_1(\cdot|s))$, Eq.23 and 26 still hold. Then, we have proved Theorem 4. □

B.2 Objective Functions of Reward Optimization

To derive $J_{R,1}$ and $J_{R,2}$, we let $\pi_1 = \pi_P$ and $\pi_2 = \pi_A$. Then based on Eq.21 and 22 we derive the following upper-bounds of $\eta_r(\pi_P) - \eta_r(\pi_A)$.

$$\eta_r(\pi_P) - \eta_r(\pi_A) \leq \sum_{t=0}^{\infty} \gamma^t \mathop{\mathbb{E}}_{s^{(t)}\sim\pi_P} \left[\Delta\mathcal{A}(s^{(t)})\right] + \frac{2\alpha\gamma(2\alpha+1)\epsilon}{(1-\gamma)^2} \tag{27}$$

$$\eta_r(\pi_P) - \eta_r(\pi_A) \geq \sum_{t=0}^{\infty} \gamma^t \mathop{\mathbb{E}}_{s^{(t)}\sim\pi_A} \left[\Delta\mathcal{A}(s^{(t)})\right] - \frac{2\alpha\gamma\epsilon}{(1-\gamma)^2} \tag{28}$$

By our assumption that π_A is optimal under r, we have $\mathcal{A}_{\pi_A} \equiv r$ Fu et al. (2018). This equivalence enables us to replace $\mathcal{A}_{\pi_A}$'s in $\Delta\mathcal{A}$ with r. As for the $\frac{2\alpha\gamma(2\alpha+1)\epsilon}{(1-\gamma)^2}$ and $\frac{2\alpha\gamma\epsilon}{(1-\gamma)^2}$ terms, since the objective is to maximize $\eta_r(\pi_A) - \eta_r(\pi_B)$, we heuristically estimate the ϵ in Eq.27 by using the samples from π_P and the ϵ in Eq.28 by using the samples from π_A. As a result we have the objective functions defined as Eq.29 and 30 where $\delta_1(s,a) = \frac{\pi_P(a^{(t)}|s^{(t)})}{\pi_A(a^{(t)}|s^{(t)})}$ and $\delta_2 = \frac{\pi_A(a^{(t)}|s^{(t)})}{\pi_P(a^{(t)}|s^{(t)})}$ are the importance sampling probability ratio derived from the definition of $\Delta\mathcal{A}$; $C_1 \propto -\frac{\gamma\hat{\alpha}}{(1-\gamma)}$ and $C_2 \propto \frac{\gamma\hat{\alpha}}{(1-\gamma)}$ where $\hat{\alpha}$ is either an estimated maximal KL-divergence between π_A and π_B since $D_{KL} \geq D_{TV}^2$ according to Schulman et al. (2015), or an estimated maximal D_{TV}^2 depending on whether the reward function is Gaussian or Categorical. We also note that for finite horizon tasks, we compute the average rewards instead of the discounted accumulated rewards in Eq.30 and 29.

$$J_{R,1}(r;\pi_P,\pi_A) := \mathop{\mathbb{E}}_{\tau\sim\pi_A}\left[\sum_{t=0}^{\infty}\gamma^t\left(\delta_1(s^{(t)},a^{(t)})-1\right)\cdot r(s^{(t)},a^{(t)})\right] + C_1 \max_{(s,a)\sim\pi_A}|r(s,a)| \tag{29}$$

$$J_{R,2}(r;\pi_P,\pi_A) := \mathop{\mathbb{E}}_{\tau\sim\pi_P}\left[\sum_{t=0}^{\infty}\gamma^t\left(1-\delta_2(s^{(t)},a^{(t)})\right)\cdot r(s^{(t)},a^{(t)})\right] + C_2 \max_{(s,a)\sim\pi_P}|r(s,a)| \tag{30}$$

Beside $J_{R,1}, J_{R,2}$, we additionally use two more objective functions based on the derived bounds. W $J_{R,r}(r;\pi_A,\pi_P)$. By denoting the optimal policy under r as π^*, $\alpha^* = \max_{s\in\mathbb{S}} D_{TV}(\pi^*(\cdot|s), \pi_A(\cdot|s))$, $\epsilon^* = \max_{(s,a^{(t)})}|\mathcal{A}_{\pi^*}(s,a^{(t)})|$, and $\Delta\mathcal{A}^*_A(s) = \mathop{\mathbb{E}}_{a\sim\pi_A}[\mathcal{A}_{\pi^*}(s,a)] - \mathop{\mathbb{E}}_{a\sim\pi^*}[\mathcal{A}_{\pi^*}(s,a)]$, we have the following.

$$\begin{aligned}
& \eta_r(\pi_P) - \eta_r(\pi^*) \\
= \;& \eta_r(\pi_P) - \eta_r(\pi_A) + \eta_r(\pi_A) - \eta_r(\pi^*) \\
\leq \;& \eta_r(\pi_P) - \eta_r(\pi_A) + \sum_{t=0}^{\infty}\gamma^t \mathop{\mathbb{E}}_{s^{(t)}\sim\pi_A}\left[\Delta\mathcal{A}^*_A(s^{(t)})\right] + \frac{2\alpha^*\gamma\epsilon^*}{(1-\gamma)^2} \\
= \;& \eta_r(\pi_P) - \sum_{t=0}^{\infty}\gamma^t \mathop{\mathbb{E}}_{s^{(t)}\sim\pi_A}\left[\mathop{\mathbb{E}}_{a^{(t)}\sim\pi_A}\left[r(s^{(t)},a^{(t)})\right]\right] + \\
& \sum_{t=0}^{\infty}\gamma^t \mathop{\mathbb{E}}_{s^{(t)}\sim\pi_A}\left[\mathop{\mathbb{E}}_{a^{(t)}\sim\pi_A}\left[\mathcal{A}_{\pi^*}(s^{(t)},a^{(t)})\right] - \mathop{\mathbb{E}}_{a^{(t)}\sim\pi^*}\left[\mathcal{A}_{\pi^*}(s^{(t)},a^{(t)})\right]\right] + \frac{2\alpha^*\gamma\epsilon^*}{(1-\gamma)^2} \\
= \;& \eta_r(\pi_P) - \sum_{t=0}^{\infty}\gamma^t \mathop{\mathbb{E}}_{s^{(t)}\sim\pi_A}\left[\mathop{\mathbb{E}}_{a^{(t)}\sim\pi^*}\left[\mathcal{A}_{\pi^*}(s^{(t)},a^{(t)})\right]\right] + \frac{2\alpha^*\gamma\epsilon^*}{(1-\gamma)^2} \\
= \;& \mathop{\mathbb{E}}_{\tau\sim\pi_P}\left[\sum_{t=0}^{\infty}\gamma^t r(s^{(t)},a^{(t)})\right] - \mathop{\mathbb{E}}_{\tau\sim\pi_A}\left[\sum_{t=0}^{\infty}\gamma^t\frac{\exp(r(s^{(t)},a^{(t)}))}{\pi_A(a^{(t)}|s^{(t)})}r(s^{(t)},a^{(t)})\right] + \frac{2\alpha^*\gamma\epsilon^*}{(1-\gamma)^2}
\end{aligned} \tag{31}$$

Let $\delta_3 = \frac{\exp(r(s^{(t)},a^{(t)}))}{\pi_A(a^{(t)}|s^{(t)})}$ be the importance sampling probability ratio. It is suggested in Schulman et al. (2017) that instead of directly optimizing the objective function Eq.31, optimizing a surrogate

objective function as in Eq.32, which is an upper-bound of Eq.31, with some small $\delta \in (0, 1)$ can be much less expensive and still effective.

$$J_{R,3}(r; \pi_P, \pi_A) := \mathop{\mathbb{E}}_{\tau \sim \pi_P} \left[\sum_{t=0}^{\infty} \gamma^t r(s^{(t)}, a^{(t)})\right] - \mathop{\mathbb{E}}_{\tau \sim \pi_A} \left[\sum_{t=0}^{\infty} \gamma^t \min\left(\delta_3 \cdot r(s^{(t)}, a^{(t)}), clip(\delta_3, 1-\delta, 1+\delta) \cdot r(s^{(t)}, a^{(t)})\right)\right] \tag{32}$$

Alternatively, we let $\Delta\mathcal{A}^*_P(s) = \mathop{\mathbb{E}}_{a \sim \pi_P} [\mathcal{A}_{\pi^*}(s, a)] - \mathop{\mathbb{E}}_{a \sim \pi^*} [\mathcal{A}_{\pi^*}(s, a)]$. The according to Eq.27, we have the following.

$$\begin{aligned} & \eta_r(\pi_P) - \eta_r(\pi^*) \\ \leq & \sum_{t=0}^{\infty} \gamma^t \mathop{\mathbb{E}}_{s^{(t)} \sim \pi_P} \left[\Delta\mathcal{A}^*_P(s^{(t)})\right] + \frac{2\alpha^*\gamma(2\alpha^*+1)\epsilon^*}{(1-\gamma)^2} \\ = & \sum_{t=0}^{\infty} \gamma^t \mathop{\mathbb{E}}_{s^{(t)} \sim \pi_P} \left[\mathop{\mathbb{E}}_{a^{(t)} \sim \pi_P} \left[\mathcal{A}_{\pi^*}(s^{(t)}, a^{(t)})\right] - \mathop{\mathbb{E}}_{a^{(t)} \sim \pi^*} \left[\mathcal{A}_{\pi^*}(s^{(t)}, a)^{(t)}\right]\right] + \frac{2\alpha^*\gamma(2\alpha^*+1)\epsilon^*}{(1-\gamma)^2} \end{aligned} \tag{33}$$

Then a new objective function $J_{R,4}$ is formulated in Eq.34 where $\delta_4 = \frac{\exp(r(s^{(t)}, a^{(t)}))}{\pi_P(a^{(t)}|s^{(t)})}$.

$$J_{R,4}(r; \pi_P, \pi_A) := \mathop{\mathbb{E}}_{\tau \sim \pi_P} \left[\sum_{t=0}^{\infty} \gamma^t r(s^{(t)}, a^{(t)})\right] - \mathop{\mathbb{E}}_{\tau \sim \pi_P} \left[\sum_{t=0}^{\infty} \gamma^t \min\left(\delta_4 \cdot r(s^{(t)}, a^{(t)}), clip(\delta_4, 1-\delta, 1+\delta) \cdot r(s^{(t)}, a^{(t)})\right)\right] \tag{34}$$

B.3 Incorporating IRL Algorithms

In our implementation, we combine PAGAR with GAIL and VAIL, respectively. When PAGAR is combined with GAIL, the meta-algorithm Algorithm 1 becomes Algorithm 2. When PAGAR is combined with VAIL, it becomes Algorithm 3. Both of the two algorithms are GAN-based IRL, indicating that both algorithms use Eq.1 as the IRL objective function. In our implementation, we use a neural network to approximate D, the discriminator in Eq.1. To get the reward function r, we follow Fu et al. (2018) and denote $r(s, a) = \log\left(\frac{\pi_A(a|s)}{D(s,a)} - \pi_A(a|s)\right)$ as mentioned in Section 1. Hence, the only difference between Algorithm 2 and Algorithm 1 is in the representation of the reward function. Regarding VAIL, since it additionally learns a representation for the state-action pairs, a bottleneck constraint $J_{IC}(D) \leq i_c$ is added where the bottleneck J_{IC} is estimated from policy roll-outs. VAIL introduces a Lagrangian parameter β to integrate $J_{IC}(D) - i_c$ in the objective function. As a result its objective function becomes $J_{IRL}(r) + \beta \cdot (J_{IC}(D) - i_c)$. VAIL not only learns the policy and the discriminator but also optimizes β. In our case, we utilize the samples from both protagonist and antagonist policies to optimize β as in line 10, where we follow Peng et al. (2019) by using projected gradient descent with a step size δ

In our implementation, depending on the difficulty of the benchmarks, we choose to maintain λ as a constant or update λ with the IRL loss $J_{IRL}(r)$. In *HalfCheetah-v2* and all the maze navigation tasks, we update λ by introducing a hyperparameter μ. As described in the maintext, we regard δ as the target IRL loss of $J_{IRL}(r)$, i.e., $J_{IRL}(r) \leq \delta$. Then we update λ by $\lambda := \lambda \cdot \exp(\mu \cdot (J_{IRL}(r) - \delta))$ after every iteration. In other benchmarks, we keep λ as a constant by letting $\mu := 0$. Besides, we use PPO Schulman et al. (2017) to train all policies in Algorithm 2 and 3.

C Experiment Details

This section presents some details of the experiments and additional results.

C.1 Experimental Details

Network Architectures. Our algorithm involves a protagonist policy π_P, and an antagonist policy π_A. In our implementation, the two policies have the same structures. Each structure contains two

Algorithm 2 GAIL w/ PAGAR

Input: Expert demonstration E, discriminator loss bound δ, initial protagonist policy π_P, antagonist policy π_A, discriminator D (representing $r(s,a) = \log\left(\frac{\pi_A(a|s)}{D(s,a)} - \pi_A(a|s)\right)$), Lagrangian parameter λ, iteration number $i = 0$, maximum iteration number N

Output: π_P

while iteration number $i < N$ **do**
  Sample trajectory sets $\mathbb{D}_A \sim \pi_A$ and $\mathbb{D}_P \sim \pi_P$
  Estimate $J_{RL}(\pi_A; r)$ with $\mathbb{D}_A$
  Optimize π_A to maximize $J_{RL}(\pi_A; r)$.
  Estimate $J_{RL}(\pi_P; r)$ with $\mathbb{D}_P$; $J_{PPO}(\pi_P; \pi_A, r)$ with $\mathbb{D}_P$ and $\mathbb{D}_A$;
  Optimize π_P to maximize $J_{RL}(\pi_P; r) + J_{PPO}(\pi_P; \pi_A, r)$.
  Estimate $J_{PAGAR}(r; \pi_P, \pi_A)$ with $\mathbb{D}_P$ and $\mathbb{D}_A$
  Estimate $J_{IRL}(\pi_A; r)$ with $\mathbb{D}_A$ and E by following the IRL algorithm
  Optimize D to minimize $J_{PAGAR}(r; \pi_P, \pi_A) + \lambda \cdot max(J_{IRL}(r) + \delta, 0)$
end while
return π_P

Algorithm 3 VAIL w/ PAGAR

Input: Expert demonstration E, discriminator loss bound δ, initial protagonist policy π_P, antagonist policy π_A, discriminator D (representing $r(s,a) = \log\left(\frac{\pi_A(a|s)}{D(s,a)} - \pi_A(a|s)\right)$), Lagrangian parameter λ for PAGAR, iteration number $i = 0$, maximum iteration number N, Lagrangian parameter β for bottleneck constraint, bounds on the bottleneck penalty i_c, learning rate δ.

Output: π_P

while iteration number $i < N$ **do**
  Sample trajectory sets $\mathbb{D}_A \sim \pi_A$ and $\mathbb{D}_P \sim \pi_P$
  Estimate $J_{RL}(\pi_A; r)$ with $\mathbb{D}_A$
  Optimize π_A to maximize $J_{RL}(\pi_A; r)$.
  Estimate $J_{RL}(\pi_P; r)$ with $\mathbb{D}_P$; $J_{PPO}(\pi_P; \pi_A, r)$ with $\mathbb{D}_P$ and $\mathbb{D}_A$;
  Optimize π_P to maximize $J_{RL}(\pi_P; r) + J_{PPO}(\pi_P; \pi_A, r)$.
  Estimate $J_{PAGAR}(r; \pi_P, \pi_A)$ with $\mathbb{D}_P$ and $\mathbb{D}_A$
  Estimate $J_{IRL}(\pi_A; r)$ with $\mathbb{D}_A$ and E by following the IRL algorithm
  Estimate $J_{IC}(D)$ with $\mathbb{D}_A, \mathbb{D}_P$ and E
  Optimize D to minimize $J_{PAGAR}(r; \pi_P, \pi_A) + \lambda \cdot max(J_{IRL}(r) + \delta, 0) + \beta \cdot J_{IC}(D)$
  Update $\beta := \max\left(0, \beta - \delta \cdot (\frac{J_{IC}(D)}{3} - i_c)\right)$
end while
return π_P

neural networks, an actor network, and a critic network. When associated with GAN-based IRL, we use a discriminator D to represent the reward function as mentioned in Appendix B.3.

- **Protagonist and Antagonist policies**. We prepare two versions of actor-critic networks, an MLP version, and a CNN version, respectively, for the Mujoco and Mini-Grid benchmarks. The MLP version, the actor and critic networks have 3 layers. Each hidden layer has 100 neurons and a $tanh$ activation function. The output layer output the mean and standard deviation of the actions. In the CNN version, the actor and critic networks share 3 convolutional layers, each having 5, 2, 2 filters, 2×2 kernel size, and $ReLU$ activation function. Then 2 fully connected networks are used to simulate the actor and critic networks. The fully connected networks have one hidden layer, of which the sizes are 64.
- **Discriminator D for GAIL w/ PAGAR in Algorithm 2**. We prepare two versions of discriminator networks, an MLP version and a CNN version, respectively, for the Mujoco and Mini-Grid benchmarks. The MLP version has 3 linear layers. Each hidden layer has 100 neurons and a $tanh$ activation function. The output layer uses the $Sigmoid$ function to output the confidence. In the CNN version, the actor and critic networks share 3 convolutional layers, each having 5, 2, 2 filters, 2×2 kernel size, and $ReLU$ activation function. The last

convolutional layer is concatenated with a fully connected network with one hidden layer with 64 neurons and $tanh$ activation function. The output layer uses the $Sigmoid$ function as the activation function.

- **Discriminator D for VAIL w/ PAGAR in Algorithm 3.** We prepare two versions of discriminator networks, an MLP version and a CNN version, respectively, for the Mujoco and Mini-Grid benchmarks. The MLP version uses 3 linear layers to generate the mean and standard deviation of the embedding of the input. Then a two-layer fully connected network takes a sampled embedding vector as input and outputs the confidence. The hidden layer in this fully connected network has 100 neurons and a $tanh$ activation function. The output layer uses the $Sigmoid$ function to output the confidence. In the CNN version, the actor and critic networks share 3 convolutional layers, each having 5, 2, 2 filters, 2×2 kernel size, and $ReLU$ activation function. The last convolutional layer is concatenated with a two-layer fully connected network. The hidden layer has 64 neurons and uses $tanh$ as the activation function. The output layer uses the $Sigmoid$ function as the activation function.

Hyperparameters The hyperparameters that appear in Algorithm 3 and 3 are summarized in Table 2 where we use `N/A` to indicate using the maximal δ as mentioned in Section 4.2, in which case we let $\mu = 0$. The values vary depending on the task and IRL algorithm for the hyperparameters δ, μ explained in Appendix B.3.

Parameter	Continuous Control Domain	Partially Observable Domain
Policy training batch size	64	256
Discount factor	0.99	0.99
GAE parameter	0.95	0.95
PPO clipping parameter	0.2	0.2
λ	1e3	1e3
σ	0.2	0.2
i_c	0.5	0.5
β	0.0	0.0
μ	VAIL(HalfCheetah): 0.5; others: 0.0	VAIL: 1.0; GAIL: 1.0
δ	VAIL(HalfCheetah): 1.0; others: N/A	VAIL: 0.8; GAIL: 1.2

Table 2: Hyperparameters used in the training processes

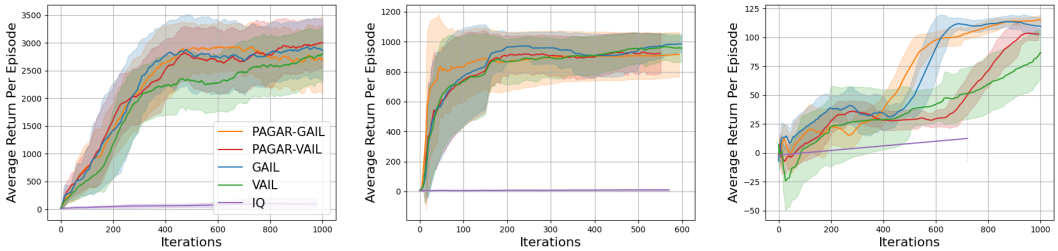
Expert Demonstrations. Our expert demonstrations all achieve high rewards in the task. The number of trajectories and the average trajectory total rewards are listed in Table 3.

C.2 Additional Results

We append the results in three Mujoco benchmarks: *Hopper-v2*, *InvertedPendulum-v2* and *Swimmer-v2* in Figure 6. Algorithm 1 performs similarly to VAIL and GAIL in those two benchmarks. IQ-learn does not perform well in Walker2d-v2 but performs better than ours and other baselines by a large margin.

Task	Number of Trajectories	Average Tot.Rewards
Walker2d-v2	10	4133
HalfCheetah-v2	100	1798
Hopper-v2	100	3586
InvertedPendulum-v2	10	1000
Swimmer-v2	10	122
DoorKey-6x6-v0	10	0.92
SimpleCrossingS9N1-v0	10	0.93

Table 3: The number of demonstrated trajectories and the average trajectory rewards

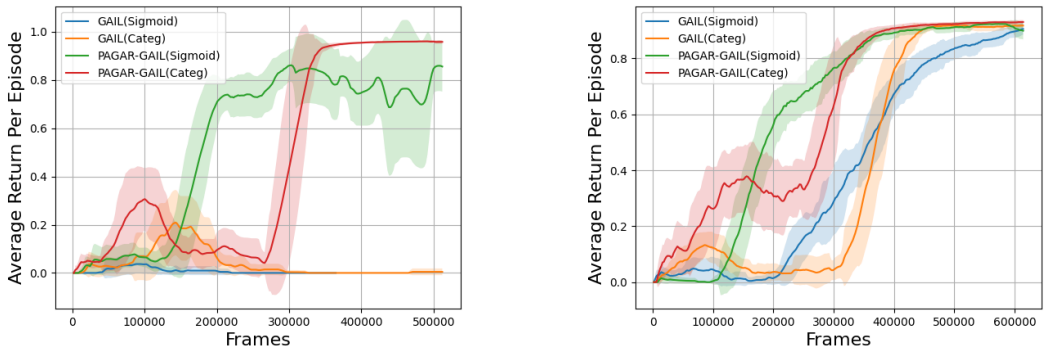


(a) Hopper-v2 (b) InvertedPendulum-v2 (c) Swimmer-v2

Figure 6: Comparing Algorithm 1 with baselines. The suffix after each '`PAGAR-`' indicates which IRL algorithm is utilized in Algorithm 1. The y axis is the average return per step. The x axis is the number of iterations in GAIL, VAIL, and ours. The policy is executed between each iteration for 2048 timesteps for sample collection. One exception is that IQ-learn updates the policy at every timestep, making its actual number of iterations 2048 times larger than indicated in the figures.

C.3 Influence of Reward Hypothesis Space

We study whether choosing a different reward function set R can influence the performance of Algorithm 1. Specifically, we use the Mini-Grid benchmarks to show the difference between using the $Sigmoid$ function and a Categorical distribution in the output layer of the discriminator network. When using the $Sigmoid$ function, the confidence of D classifying a state-action pair as being sampled from the antagonist policy's roll-outs is not normalized over the action space, i.e., $\sum_{a\in\mathbb{A}} D(s,a) \neq 1$. When using a Categorical distribution, the confidence sum to one for all the actions, i.e., $\sum_{a\in\mathbb{A}} D(s,a) = 1$. We fixed the IRL algorithm as the one used in GAIL. The results are shown in Figure 7. When changing the discriminator's output layer, GAIL, and ours' training efficiency change. However, our algorithm outperforms GAIL in both cases by using fewer samples to train the protagonist policy to attain high performance.



(a) MiniGrid-DoorKey-8x8-v0 (b) MiniGrid-SimpleCrossingS9N2-v0

Figure 7: Comparing Algorithm 1 with baselines. The prefix '`protagonist_GAIL`' indicates that the IRL algorithm utilized in Algorithm 1 is the same as in GAIL. The '`_Sigmoid`' and '`_Categ`' suffixes indicate whether the output layer of the discriminator is using the $Sigmoid$ function or Categorical distribution. The x axis is the number of sampled frames. The y axis is the average return per episode.