Robust Inverse Reinforcement Learning under Transition Dynamics Mismatch

Luca Viano
LIONS, EPFL

Yu-Ting Huang
EPFL

Parameswaran Kamalaruban*
The Alan Turing Institute

Adrian Weller
University of Cambridge
& The Alan Turing Institute

Volkan Cevher
LIONS, EPFL

Abstract

We study the inverse reinforcement learning (IRL) problem under a transition dynamics mismatch between the expert and the learner. Specifically, we consider the Maximum Causal Entropy (MCE) IRL learner model and provide a tight upper bound on the learner's performance degradation based on the ℓ_1-distance between the transition dynamics of the expert and the learner. Leveraging insights from the Robust RL literature, we propose a robust MCE IRL algorithm, which is a principled approach to help with this mismatch. Finally, we empirically demonstrate the stable performance of our algorithm compared to the standard MCE IRL algorithm under transition dynamics mismatches in both finite and continuous MDP problems.

1 Introduction

Recent advances in Reinforcement Learning (RL) [1, 2, 3, 4] have demonstrated impressive performance in games [5, 6], continuous control [7], and robotics [8]. Despite these successes, a broader application of RL in real-world domains is hindered by the difficulty of designing a proper reward function. Inverse Reinforcement Learning (IRL) addresses this issue by inferring a reward function from a given set of demonstrations of the desired behavior [9, 10]. IRL has been extensively studied, and many algorithms have already been proposed [11, 12, 13, 14, 15, 16].

Almost all IRL algorithms assume that the expert demonstrations are collected from the same environment as the one in which the IRL agent is trained. However, this assumption rarely holds in real world because of many possible factors identified by [17]. For example, consider an autonomous car that should learn by observing expert demonstrations performed on another car with possibly different technical characteristics. There is often a mismatch between the learner and the expert's transition dynamics, resulting in poor performance that are critical in healthcare [18] or autonomous driving [19]. Indeed, the performance degradation of an IRL agent due to transition dynamics mismatch has been noted empirically [20, 21, 22, 23], but without theoretical guidance.

To this end, our work first provides a theoretical study on the effect of such mismatch in the context of the infinite horizon Maximum Causal Entropy (MCE) IRL framework [24, 25, 26]. Specifically, we bound the potential decrease in the IRL learner's performance as a function of the ℓ_1-distance between the expert and the learner's transition dynamics. We then propose a robust variant of the MCE IRL algorithm to effectively recover a reward function under transition dynamics mismatch, mitigating degradation. There is precedence to our robust IRL approach, such as [27] that employs

*Correspondence to: Parameswaran Kamalaruban <kparameswaran@turing.ac.uk>

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

an adversarial training method to learn a robust policy against adversarial changes in the learner's environment. The novel idea of our work is to incorporate this method within our IRL context, by viewing the expert's transition dynamics as a perturbed version of the learner's one.

Our robust MCE IRL algorithm leverages techniques from the robust RL literature [28, 29, 30, 27]. A few recent works [20, 21, 31] attempt to infer the expert's transition dynamics from the demonstration set or via additional information, and then apply the standard IRL method to recover the reward function based on the learned dynamics. Still, the transition dynamics can be estimated only up to a certain accuracy, i.e., a mismatch between the learner's belief and the dynamics of the expert's environment remains. Our robust IRL approach can be incorporated into this research vein to further improve the IRL agent's performance.

To our knowledge, this is the first work that rigorously reconciles model-mismatch in IRL with only one shot access to the expert environment. We highlight the following contributions:

1. We provide a tight upper bound for the suboptimality of an IRL learner that receives expert demonstrations from an MDP with different transition dynamics compared to a learner that receives demonstrations from an MDP with the same transition dynamics (Section 3.1).
2. We find suitable conditions under which a solution exists to the MCE IRL optimization problem with model mismatch (Section 3.2).
3. We propose a robust variant of the MCE IRL algorithm to learn a policy from expert demonstrations under transition dynamics mismatch (Section 4).
4. We demonstrate our method's robust performance compared to the standard MCE IRL in a broad set of experiments under both linear and non-linear reward settings (Section 5).
5. We extend our robust IRL method to the high dimensional continuous MDP setting with appropriate practical relaxations, and empirically demonstrate its effectiveness (Section 6).

2 Problem Setup

This section formalizes the IRL problem with an emphasis on the learner and expert environments. We use bold notation to represent vectors. A glossary of notation is given in Appendix C.

2.1 Environment and Reward

We formally represent the environment by a Markov decision process (MDP) $M_{\boldsymbol{\theta}} := \{\mathcal{S}, \mathcal{A}, T, \gamma, P_0, R_{\boldsymbol{\theta}}\}$, parameterized by $\boldsymbol{\theta} \in \mathbb{R}^d$. The state and action spaces are denoted as $\mathcal{S}$ and $\mathcal{A}$, respectively. We assume that $|\mathcal{S}|, |\mathcal{A}| < \infty$. $T : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$ represents the transition dynamics, i.e., $T(s'|s, a)$ is the probability of transitioning to state s' by taking action a from state s. The discount factor is given by $\gamma \in (0, 1)$, and P_0 is the initial state distribution. We consider a linear reward function $R_{\boldsymbol{\theta}} : \mathcal{S} \to \mathbb{R}$ of the form $R_{\boldsymbol{\theta}}(s) = \langle \boldsymbol{\theta}, \boldsymbol{\phi}(s) \rangle$, where $\boldsymbol{\theta} \in \mathbb{R}^d$ is the reward parameter, and $\boldsymbol{\phi} : \mathcal{S} \to \mathbb{R}^d$ is a feature map. We use a one-hot feature map $\boldsymbol{\phi} : \mathcal{S} \to \{0, 1\}^{|\mathcal{S}|}$, where the sth element of $\boldsymbol{\phi}(s)$ is 1 and 0 elsewhere. Our results can be extended to any general feature map (see empirical evidence in Fig. 6), but we use this particular choice as a running example for concreteness.

We focus on the state-only reward function since the state-action reward function is not that useful in the robustness context. Indeed, as [22] pointed out, the actions to achieve a specific goal under different transition dynamics will not necessarily be the same and, consequently, should not be imitated. Analogously, in the IRL context, the reward for taking a particular action should not be recovered since the quality of that action depends on the transition dynamics. We denote an MDP without a reward function by $M = M_{\boldsymbol{\theta}} \backslash R_{\boldsymbol{\theta}} = \{\mathcal{S}, \mathcal{A}, T, \gamma, P_0\}$.

2.2 Policy and Performance

A policy $\pi : \mathcal{S} \to \Delta_{\mathcal{A}}$ is a mapping from a state to a probability distribution over actions. The set of all valid stochastic policies is denoted by $\Pi := \{\pi : \sum_a \pi(a|s) = 1, \forall s \in \mathcal{S}; \pi(a|s) \geq 0, \forall (s, a) \in \mathcal{S} \times \mathcal{A}\}$. We are interested in two different performance measures of any policy π acting in the MDP $M_{\boldsymbol{\theta}}$: (i) the expected discounted return $V^{\pi}_{M_{\boldsymbol{\theta}}} := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_{\boldsymbol{\theta}}(s_t) \mid \pi, M\right]$, and (ii) its entropy regularized variant $V^{\pi,\text{soft}}_{M_{\boldsymbol{\theta}}} := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \left\{R_{\boldsymbol{\theta}}(s_t) - \log \pi(a_t|s_t)\right\} \mid \pi, M\right]$. The state occupancy measure of a policy π in the

MDP M is defined as $\rho_M^\pi(s) := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}\left[s_t = s \mid \pi, M\right]$, where $\mathbb{P}\left[s_t = s \mid \pi, M\right]$ denotes the probability of visiting the state s after t steps by following the policy π in M. Note that $\rho_M^\pi(s)$ does not depend on the reward function. Let $\boldsymbol{\rho}_M^\pi \in \mathbb{R}^{|\mathcal{S}|}$ be a vector whose s^{th} element is $\rho_M^\pi(s)$. For the one-hot feature map $\boldsymbol{\phi}$, we have that $V_{M_{\boldsymbol{\theta}}}^\pi = \frac{1}{1-\gamma}\sum_s \rho_M^\pi(s) R_{\boldsymbol{\theta}}(s) = \frac{1}{1-\gamma}\langle \boldsymbol{\theta}, \boldsymbol{\rho}_M^\pi \rangle$. A policy π is *optimal* for the MDP $M_{\boldsymbol{\theta}}$ if $\pi \in \arg\max_{\pi'} V_{M_{\boldsymbol{\theta}}}^{\pi'}$, in which case we denote it by $\pi_{M_{\boldsymbol{\theta}}}^*$. Similarly, the *soft-optimal* policy (always unique [32]) in $M_{\boldsymbol{\theta}}$ is defined as $\pi_{M_{\boldsymbol{\theta}}}^{\text{soft}} := \arg\max_{\pi'} V_{M_{\boldsymbol{\theta}}}^{\pi',\text{soft}}$ (see Appendix D for a parametric form of this policy).

2.3 Learner and Expert

Our setting has two entities: a learner implementing the MCE IRL algorithm, and an expert. We consider two MDPs, $M_{\boldsymbol{\theta}}^L = \left\{\mathcal{S}, \mathcal{A}, T^L, \gamma, P_0, R_{\boldsymbol{\theta}}\right\}$ and $M_{\boldsymbol{\theta}}^E = \left\{\mathcal{S}, \mathcal{A}, T^E, \gamma, P_0, R_{\boldsymbol{\theta}}\right\}$, that differ only in the transition dynamics. The true reward parameter $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ is known only to the expert. The expert provides demonstrations to the learner: (i) by following policy $\pi_{M_{\boldsymbol{\theta}^*}^E}^*$ in M^E when there is a *transition dynamics mismatch* between the learner and the expert, or (ii) by following policy $\pi_{M_{\boldsymbol{\theta}^*}^L}^*$ in M^L otherwise. The learner always operates in the MDP M^L and is not aware of the true reward parameter and of the expert dynamics T^E[2], i.e., it only has access to $M_{\boldsymbol{\theta}^*}^L \backslash R_{\boldsymbol{\theta}^*}$. It learns a reward parameter $\boldsymbol{\theta}$ and the corresponding soft-optimal policy $\pi_{M_{\boldsymbol{\theta}}^L}^{\text{soft}}$, based on the state occupancy measure $\boldsymbol{\rho}$ received from the expert. Here, $\boldsymbol{\rho}$ is either $\boldsymbol{\rho}_{M^E}^{\pi_{M^E_{\boldsymbol{\theta}^*}}^*}$ or $\boldsymbol{\rho}_{M^L}^{\pi_{M^L_{\boldsymbol{\theta}^*}}^*}$ depending on the case. Our results can be extended to the stochastic estimate of $\boldsymbol{\rho}$ using concentration inequalities [11].

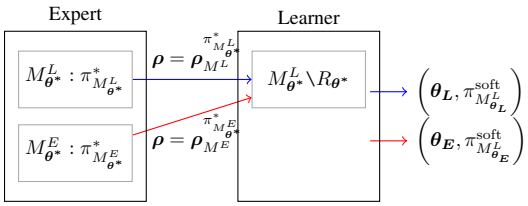


Figure 1: An illustration of the IRL problem under transition dynamics mismatch: See Section 2.

Our learner model builds on the MCE IRL [24, 25, 26] framework that matches the expert's state occupancy measure $\boldsymbol{\rho}$. In particular, the learner policy is obtained by maximizing its causal entropy while matching the expert's state occupancy:

$$\max_{\pi \in \Pi} \mathbb{E}\left[\sum_{t=0}^{\infty} -\gamma^t \log \pi(a_t|s_t) \;\middle|\; \pi, M^L\right] \quad \text{subject to} \quad \boldsymbol{\rho}_{M^L}^\pi = \boldsymbol{\rho}. \tag{1}$$

Note that this optimization problem only requires access to $M_{\boldsymbol{\theta}}^L \backslash R_{\boldsymbol{\theta}}$. The constraint in (1) follows from our choice of the one-hot feature map. We denote the optimal solution of the above problem by $\pi_{M_{\boldsymbol{\theta}}^L}^{\text{soft}}$ with a corresponding reward parameter: (i) $\boldsymbol{\theta} = \boldsymbol{\theta_E}$, when we use $\boldsymbol{\rho}_{M^E}^{\pi_{M^E_{\boldsymbol{\theta}^*}}^*}$ as $\boldsymbol{\rho}$, or (ii) $\boldsymbol{\theta} = \boldsymbol{\theta_L}$, when we use $\boldsymbol{\rho}_{M^L}^{\pi_{M^L_{\boldsymbol{\theta}^*}}^*}$ as $\boldsymbol{\rho}$. Here, the parameters $\boldsymbol{\theta_E}$ and $\boldsymbol{\theta_L}$ are obtained by solving the corresponding dual problems of (1). Finally, we are interested in the performance of the learner policy $\pi_{M_{\boldsymbol{\theta}}^L}^{\text{soft}}$ in the MDP $M_{\boldsymbol{\theta}^*}^L$. Our problem setup is illustrated in Figure 1.

3 MCE IRL under Transition Dynamics Mismatch

This section analyses the MCE IRL learner's suboptimality when there is a transition dynamics mismatch between the expert and the learner, as opposed to an ideal learner without this mismatch. The proofs of the theoretical statements of this section can be found in Appendix E.

3.1 Upper bound on the Performance Gap

First, we introduce an auxiliary lemma to be used later in our analysis. We define the distance between the two transition dynamics T and T', and the distance between the two poli-

[2]The setting with T^E known to the learner has been studied under the name of *imitation learning across embodiments* [33].

cies π and π' as follows, respectively: $d_{\text{dyn}}(T, T') := \max_{s,a} \|T(\cdot \mid s, a) - T'(\cdot \mid s, a)\|_1$, and $d_{\text{pol}}(\pi, \pi') := \max_s \|\pi(\cdot|s) - \pi'(\cdot|s)\|_1$. Consider the two MDPs $M_{\boldsymbol{\theta}} = \{\mathcal{S}, \mathcal{A}, T, \gamma, P_0, R_{\boldsymbol{\theta}}\}$ and $M'_{\boldsymbol{\theta}} = \{\mathcal{S}, \mathcal{A}, T', \gamma, P_0, R_{\boldsymbol{\theta}}\}$. We assume that the reward function is bounded, i.e., $R_{\boldsymbol{\theta}}(s) \in \left[R^{\min}_{\boldsymbol{\theta}}, R^{\max}_{\boldsymbol{\theta}}\right], \forall s \in \mathcal{S}$. Also, we define the following two constants: $\kappa_{\boldsymbol{\theta}} := \sqrt{\gamma \cdot \max\left\{R^{\max}_{\boldsymbol{\theta}} + \log|\mathcal{A}|, -\log|\mathcal{A}| - R^{\min}_{\boldsymbol{\theta}}\right\}}$ and $|R_{\boldsymbol{\theta}}|^{\max} := \max\left\{\left|R^{\min}_{\boldsymbol{\theta}}\right|, \left|R^{\max}_{\boldsymbol{\theta}}\right|\right\}$.

Lemma 1. *Let $\pi := \pi^{\text{soft}}_{M_{\boldsymbol{\theta}}}$ and $\pi' := \pi^{\text{soft}}_{M'_{\boldsymbol{\theta}}}$ be the soft optimal policies for the MDPs $M_{\boldsymbol{\theta}}$ and $M'_{\boldsymbol{\theta}}$ respectively. Then, the distance between π and π' is bounded as follows: $d_{\text{pol}}(\pi', \pi) \leq 2 \min\left\{\frac{\kappa_{\boldsymbol{\theta}}\sqrt{d_{\text{dyn}}(T',T)}}{(1-\gamma)}, \frac{\kappa^2_{\boldsymbol{\theta}} d_{\text{dyn}}(T',T)}{(1-\gamma)^2}\right\}$.*

The above result is obtained by bounding the KL divergence between the two soft optimal policies, and involves a non-standard derivation compared to the well-established performance difference theorems in the literature (see Appendix E.1). The lemma above bounds the maximum total variation distance between two soft optimal policies obtained by optimizing the same reward under different transition dynamics. It serves as a prerequisite result for our later theorems (Theorem 1 for soft optimal experts and Theorem 6). In addition, it may be a result of independent interest for entropy regularized MDP.

Now, we turn to our objective. Let $\pi_1 := \pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_L}}}$ be the policy returned by the MCE IRL algorithm when there is no transition dynamics mismatch. Similarly, let $\pi_2 := \pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}$ be the policy returned by the MCE IRL algorithm when there is a mismatch. Note that π_1 and π_2 are the corresponding solutions to the optimization problem (1), when $\boldsymbol{\rho} \leftarrow \boldsymbol{\rho}^{\pi^*_{M^L_{\boldsymbol{\theta}^*}}}_{M^L}$ and $\boldsymbol{\rho} \leftarrow \boldsymbol{\rho}^{\pi^*_{M^E_{\boldsymbol{\theta}^*}}}_{M^E}$, respectively. The following theorem bounds the performance degradation of the policy π_2 compared to the policy π_1 in the MDP $M^L_{\boldsymbol{\theta}^*}$, where the learner operates on:

Theorem 1. *The performance gap between the policies π_1 and π_2 on the MDP $M^L_{\boldsymbol{\theta}^*}$ is bounded as follows: $\left|V^{\pi_1}_{M^L_{\boldsymbol{\theta}^*}} - V^{\pi_2}_{M^L_{\boldsymbol{\theta}^*}}\right| \leq \frac{\gamma \cdot |R_{\boldsymbol{\theta}^*}|^{\max}}{(1-\gamma)^2} \cdot d_{\text{dyn}}\left(T^L, T^E\right)$.*

The above result is obtained from the optimality conditions of the problem (1), and using Theorem 7 from [34]. In Section 4.4, we show that the above bound is indeed tight. When the expert policy is soft-optimal, we can use Lemma 1 and Simulation Lemma [35, 36] to obtain an upper bound on the performance gap (see Appendix E.2). For an application of Theorem 1, consider an IRL learner that first learns a simulator of the expert environment, and then matches the expert behavior in the simulator. In this case, our upper bound provides an estimate (sufficient condition) of the accuracy required for the simulator.

3.2 Existence of Solution under Mismatch

The proof of the existence of a unique solution to the optimization problem (1), presented in [37], relies on the fact that both expert and learner environments are the same. This assumption implies that the expert policy is in the feasible set that is consequently non-empty. Theorem 2 presented in this section poses a condition under which we can ensure that the feasible set is non-empty when the expert and learner environments are not the same.

Given M^L and $\boldsymbol{\rho}$, we define the following quantities useful for stating our theorem. We define, for each state $s \in \mathcal{S}$, the probability flow matrix $\boldsymbol{F}(s) \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ as follows: $[\boldsymbol{F}(s)]_{i,j} := \rho(s) T^L_{s_i,s,a_j}$, where $T^L_{s_i,s,a_j} := T^L(s_i|s, a_j)$ for $i = 1, \ldots, |\mathcal{S}|$ and $j = 1, \ldots, |\mathcal{A}|$. Let $\boldsymbol{B}(s) \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ be a row matrix that contains only ones in row s and zero elsewhere. Then, we define the matrix $\boldsymbol{T} \in \mathbb{R}^{2|\mathcal{S}|\times|\mathcal{S}||\mathcal{A}|}$ by stacking the probability flow and the row matrices as follows: $\boldsymbol{T} := \begin{bmatrix} \boldsymbol{F}(s_1) & \boldsymbol{F}(s_2) & \ldots & \boldsymbol{F}(s_{|\mathcal{S}|}) \\ \boldsymbol{B}(s_1) & \boldsymbol{B}(s_2) & \ldots & \boldsymbol{B}(s_{|\mathcal{S}|}) \end{bmatrix}$. In addition, we define the vector $\boldsymbol{v} \in \mathbb{R}^{2|\mathcal{S}|}$ as follows: $\boldsymbol{v}_i = \rho(s_i) - (1-\gamma)P_0(s_i)$ if $i \leq |\mathcal{S}|$, and 1 otherwise.

Theorem 2. *The feasible set of the optimization problem* (1) *is non-empty iff the rank of the matrix $\boldsymbol{T}$ is equal to the rank of the augmented matrix $(\boldsymbol{T}|\boldsymbol{v})$.*

The proof of the above theorem leverages the fact that the Bellman flow constraints [15] must hold for any policy in an MDP. This requirement leads to the formulation of a linear system whose solutions

set corresponds to the feasible set of (1). The Rouché-Capelli theorem [38][Theorem 2.38] states that the solutions set is non-empty if and only if the condition in Theorem 2 holds. We note that the construction of the matrix $\boldsymbol{T}$ does not assume any restriction on the MDP structure since it leverages only on the Bellman flow constraints. Theorem 2 allows us to develop a robust MCE IRL scheme in Section 4 by ensuring the absence of duality gap. To this end, the following corollary provides a simple sufficient condition for the existence of a solution under transition dynamics mismatch.

Corollary 1. *Let $|\mathcal{A}| > 1$. Then, a sufficient condition for the non-emptiness of the feasible set of the optimization problem* (1) *is given by $\boldsymbol{T}$ being full rank.*

3.3 Reward Transfer under Mismatch

Consider a class $\mathcal{M}$ of MDPs such that it contains both the learner and the expert environments, i.e., $M^L, M^E \in \mathcal{M}$ (see Figure 2). We are given the expert's state occupancy measure $\boldsymbol{\rho} = \boldsymbol{\rho}_{M^E}^{\pi^*_{M^E_{\boldsymbol{\theta}^*}}}$; but the expert's policy $\pi^*_{M^E_{\boldsymbol{\theta}^*}}$ and the MDP M^E are unknown. Further, we assume that every MDP $M \in \mathcal{M}$ satisfies the condition in Theorem 2.

We aim to find a policy π^L that performs well in the MDP $M^L_{\boldsymbol{\theta}^*}$, i.e., $V^{\pi^L}_{M^L_{\boldsymbol{\theta}^*}}$ is high. To this end, we can choose any MDP $M^{\text{train}} \in \mathcal{M}$, and solve the MCE IRL problem (1) with the constraint given by $\boldsymbol{\rho} = \boldsymbol{\rho}^{\pi}_{M^{\text{train}}}$. Then, we always obtain a reward parameter $\boldsymbol{\theta}^{\text{train}}$ s.t. $\boldsymbol{\rho} = \boldsymbol{\rho}_{M^{\text{train}}}^{\pi^{\text{soft}}_{M^{\text{train}}_{\boldsymbol{\theta}^{\text{train}}}}}$, since M^{train} satisfies the condition in Theorem 2. We can use this reward parameter $\boldsymbol{\theta}^{\text{train}}$ to learn a good policy π^L in the MDP $M^L_{\boldsymbol{\theta}^{\text{train}}}$, i.e., $\pi^L := \pi^*_{M^L_{\boldsymbol{\theta}^{\text{train}}}}$ or $\pi^L := \pi^{\text{soft}}_{M^L_{\boldsymbol{\theta}^{\text{train}}}}$. Using Lemma 1, we obtain a bound on the performance gap between π^L and $\pi_1 := \pi^{\text{soft}}_{M^L_{\boldsymbol{\theta}^L}}$ (see Theorem 6 in Appendix E.4).

However, there are two problems with this approach: (i) it requires access to multiple environments M^{train}, and (ii) unless M^{train} happened to be closer to the expert's MDP M^E, we cannot recover the true intention of the expert. Since the MDP M^E is unknown, one cannot compare the different reward parameters $\boldsymbol{\theta}^{\text{train}}$'s obtained with different MDPs M^{train}'s. Thus, with $\boldsymbol{\theta}^{\text{train}}$, it is impossible to ensure that the performance of π^L is high in the MDP $M^L_{\boldsymbol{\theta}^*}$. Instead, we try to learn a robust policy π^L over the class $\mathcal{M}$, while aligning with the expert's occupancy measure $\boldsymbol{\rho}$, and acting only in M^L. By doing this, we ensure that π^L performs reasonably well on any MDP $M_{\boldsymbol{\theta}^*} \in \mathcal{M}$ including $M^L_{\boldsymbol{\theta}^*}$. We further build upon this idea in the next section.

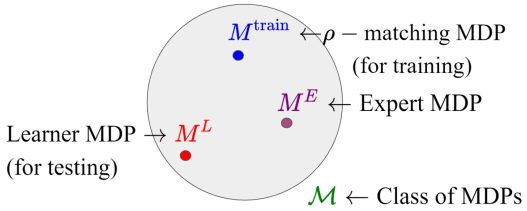


Figure 2: Illustrative example of learning a policy π^L to act in one MDP M^L, given the expert occupancy measure $\boldsymbol{\rho}$.

4 Robust MCE IRL via Two-Player Markov Game

4.1 Robust MCE IRL Formulation

This section focuses on recovering a learner policy via MCE IRL framework in a robust manner, under transition dynamics mismatch, i.e., $\boldsymbol{\rho} = \boldsymbol{\rho}_{M^E}^{\pi^{\text{soft}}_{M^E_{\boldsymbol{\theta}^*}}}$ in Eq. (1). In particular, our learner policy matches the expert state occupancy measure $\boldsymbol{\rho}$ under the most adversarial transition dynamics belonging to a set described as follows for a given $\alpha > 0$: $\mathcal{T}^{L,\alpha} := \left\{\alpha T^L + (1-\alpha)\bar{T}, \forall \bar{T} \in \Delta_{\mathcal{S}|\mathcal{S},\mathcal{A}}\right\}$, where $\Delta_{\mathcal{S}|\mathcal{S},\mathcal{A}}$ is the set of all the possible transition dynamics $T : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0,1]$. Note that the set $\mathcal{T}^{L,\alpha}$ is equivalent to the (s,a)-rectangular uncertainty set [28] centered around T^L, i.e., $\mathcal{T}^{L,\alpha} = \left\{T : d_{\text{dyn}}\left(T, T^L\right) \leq 2(1-\alpha)\right\}$. We need this set $\mathcal{T}^{L,\alpha}$ for establishing the equivalence between robust MDP and action-robust MDP formulations. The action-robust MDP formulation allows us to learn a robust policy while accessing only the MDP M^L.

We define a class of MDPs as follows: $\mathcal{M}^{L,\alpha} := \left\{\left\{\mathcal{S}, \mathcal{A}, T^{L,\alpha}, \gamma, P_0\right\}, \forall T^{L,\alpha} \in \mathcal{T}^{L,\alpha}\right\}$. Then, based on the discussions in Section 3.3, we propose the following robust MCE IRL problem:

$$\max_{\pi^{\mathrm{pl}} \in \Pi} \min_{M \in \mathcal{M}^{L,\alpha}} \mathbb{E}\left[\sum_{t=0}^{\infty} -\gamma^t \log \pi^{\mathrm{pl}}(a_t|s_t) \,\middle|\, \pi^{\mathrm{pl}}, M\right] \text{ subject to } \boldsymbol{\rho}_M^{\pi^{\mathrm{pl}}} = \boldsymbol{\rho} \tag{2}$$

The corresponding dual problem is given by:

$$\min_{\boldsymbol{\theta}} \max_{\pi^{\mathrm{pl}} \in \Pi} \min_{M \in \mathcal{M}^{L,\alpha}} \mathbb{E}\left[\sum_{t=0}^{\infty} -\gamma^t \log \pi^{\mathrm{pl}}(a_t|s_t) \,\middle|\, \pi^{\mathrm{pl}}, M\right] + \boldsymbol{\theta}^\top \left(\boldsymbol{\rho}_M^{\pi^{\mathrm{pl}}} - \boldsymbol{\rho}\right) \tag{3}$$

In the dual problem, for any $\boldsymbol{\theta}$, we attempt to learn a robust policy over the class $\mathcal{M}^{L,\alpha}$ with respect to the entropy regularized reward function. The parameter $\boldsymbol{\theta}$ plays the role of aligning the learner's policy with the expert's occupancy measure via constraint satisfaction.

4.2 Existence of Solution

We start by formulating the IRL problem for any MDP $M^{L,\alpha} \in \mathcal{M}^{L,\alpha}$, with transition dynamics $T^{L,\alpha} = \alpha T^L + (1-\alpha)\bar{T} \in \mathcal{T}^{L,\alpha}$, as follows:

$$\max_{\pi^{\mathrm{pl}} \in \Pi} \mathbb{E}\left[\sum_{t=0}^{\infty} -\gamma^t \log \pi^{\mathrm{pl}}(a_t|s_t) \,\middle|\, \pi^{\mathrm{pl}}, M^{L,\alpha}\right] \text{ subject to } \boldsymbol{\rho}_{M^{L,\alpha}}^{\pi^{\mathrm{pl}}} = \boldsymbol{\rho} \tag{4}$$

By introducing the Lagrangian vector $\boldsymbol{\theta} \in \mathbb{R}^{|\mathcal{S}|}$, we get:

$$\max_{\pi^{\mathrm{pl}} \in \Pi} \mathbb{E}\left[\sum_{t=0}^{\infty} -\gamma^t \log \pi^{\mathrm{pl}}(a_t|s_t) \,\middle|\, \pi^{\mathrm{pl}}, M^{L,\alpha}\right] + \boldsymbol{\theta}^\top \left(\boldsymbol{\rho}_{M^{L,\alpha}}^{\pi^{\mathrm{pl}}} - \boldsymbol{\rho}\right) \tag{5}$$

For any fixed $\boldsymbol{\theta}$, the problem (5) is feasible since Π is a closed and bounded set. We define $U(\boldsymbol{\theta})$ as the value of the program (5) for a given $\boldsymbol{\theta}$. By weak duality, $U(\boldsymbol{\theta})$ provides an upper bound on the optimization problem (4). Consequently, we introduce the dual problem aiming to find the value of $\boldsymbol{\theta}$ corresponding to the lowest upper bound, which can be written as

$$\min_{\boldsymbol{\theta}} U(\boldsymbol{\theta}) := \max_{\pi^{\mathrm{pl}} \in \Pi} \mathbb{E}\left[\sum_{t=0}^{\infty} -\gamma^t \log \pi^{\mathrm{pl}}(a_t|s_t) \,\middle|\, \pi^{\mathrm{pl}}, M^{L,\alpha}\right] + \boldsymbol{\theta}^\top \left(\boldsymbol{\rho}_{M^{L,\alpha}}^{\pi^{\mathrm{pl}}} - \boldsymbol{\rho}\right). \tag{6}$$

Given $\boldsymbol{\theta}$, we define $\pi^{\mathrm{pl},*} := \pi^{\mathrm{soft}}_{M^{L,\alpha}_{\boldsymbol{\theta}}}$. Due to [32][Theorem 1], for any fixed $M^{L,\alpha}_{\boldsymbol{\theta}}$, the policy $\pi^{\mathrm{pl},*}$ exists and it is unique. We can compute the gradient[3] $\nabla_{\boldsymbol{\theta}} U = \boldsymbol{\rho}_{M^{L,\alpha}}^{\pi^{\mathrm{pl},*}} - \boldsymbol{\rho}$, and update the parameter via gradient descent: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \nabla_{\boldsymbol{\theta}} U$. Note that, if the condition in Theorem 2 holds, the feasible set of (4) is non-empty. Then, according to [37][Lemma 2], there is no duality gap between the programs (4) and (6). Based on these observations, we argue that the program (2) is well-posed and admits a unique solution.

4.3 Solution via Markov Game

In the following, we outline a method (see Algorithm 1) to solve the robust MCE IRL dual problem (3). To this end, for any given $\boldsymbol{\theta}$, we need to solve the inner max-min problem of (3). First, we express the entropy term $\mathbb{E}\left[\sum_{t=0}^{\infty} -\gamma^t \log \pi^{\mathrm{pl}}(a_t|s_t) \middle| \pi^{\mathrm{pl}}, M\right]$ as follows:

$$\sum_{s \in \mathcal{S}} \rho_M^{\pi^{\mathrm{pl}}}(s) \sum_{a \in \mathcal{A}} \left\{-\pi^{\mathrm{pl}}(a|s) \log \pi^{\mathrm{pl}}(a|s)\right\} = \sum_{s \in \mathcal{S}} \rho_M^{\pi^{\mathrm{pl}}}(s) H^{\pi^{\mathrm{pl}}}(A \mid S = s) = \left(\boldsymbol{H}^{\pi^{\mathrm{pl}}}\right)^\top \boldsymbol{\rho}_M^{\pi^{\mathrm{pl}}},$$

where $\boldsymbol{H}^{\pi^{\mathrm{pl}}} \in \mathbb{R}^{|\mathcal{S}|}$ a vector whose s^{th} element is the entropy of the player policy given the state s. Since the quantity $\boldsymbol{H}^{\pi^{\mathrm{pl}}} + \boldsymbol{\theta}$ depends only on the states, to solve the dual problem, we can utilize the equivalence between the *robust MDP* [28, 29] formulation and the *action-robust MDP* [30, 27, 40] formulation shown in [27]. We can interpret the minimization over the environment class as the

[3] In Appendix F.2, we proved that this is indeed the gradient update under the transition dynamics mismatch.

Algorithm 1 Robust MCE IRL via Markov Game

Input: opponent strength $1-\alpha$
Initialize: player policy π^{pl}, opponent policy π^{op}, and parameter $\boldsymbol{\theta}$
while not converged **do**
  compute $\boldsymbol{\rho}_{M^L}^{\alpha\pi^{\mathrm{pl}}+(1-\alpha)\pi^{\mathrm{op}}}$ by dynamic programming [37][Section V.C].
  update $\boldsymbol{\theta}$ with Adam [39] using the gradient $\left(\boldsymbol{\rho}_{M^L}^{\alpha\pi^{\mathrm{pl}}+(1-\alpha)\pi^{\mathrm{op}}} - \boldsymbol{\rho}\right)$.
  use Algorithm 2 with $R = R_{\boldsymbol{\theta}}$ to update π^{pl} and π^{op} s.t. they solve the problem (9).
end while
Output: player policy π^{pl}

minimization over a set of opponent policies that with probability $1-\alpha$ take control of the agent and perform the worst possible move from the current agent state. Indeed, interpreting $\left(\boldsymbol{H}^{\pi^{\mathrm{pl}}}+\boldsymbol{\theta}\right)^\top \boldsymbol{\rho}_M^{\pi^{\mathrm{pl}}}$ as an entropy regularized value function, i.e., $\boldsymbol{\theta}$ as a reward parameter, we can write:

$$\max_{\pi^{\mathrm{pl}}\in\Pi}\min_{M\in\mathcal{M}^{L,\alpha}}\left(\boldsymbol{H}^{\pi^{\mathrm{pl}}}+\boldsymbol{\theta}\right)^\top \boldsymbol{\rho}_M^{\pi^{\mathrm{pl}}} = \max_{\pi^{\mathrm{pl}}\in\Pi}\min_{\bar{T}}\mathbb{E}\left[G \mid \pi^{\mathrm{pl}}, P_0, \alpha T^L + (1-\alpha)\bar{T}\right] \tag{7}$$

$$\leq \max_{\pi^{\mathrm{pl}}\in\Pi}\min_{\pi^{\mathrm{op}}\in\Pi}\mathbb{E}\left[G \mid \alpha\pi^{\mathrm{pl}} + (1-\alpha)\pi^{\mathrm{op}}, M^L\right], \tag{8}$$

where $G := \sum_{t=0}^{\infty}\gamma^t\left\{R_{\boldsymbol{\theta}}(s_t) + H^{\pi^{\mathrm{pl}}}\left(A \mid S=s_t\right)\right\}$. The above inequality holds due to the derivation in section 3.1 of [27]. Further details are in Appendix F.1.

Finally, we can formulate the problem (8) as a two-player zero-sum Markov game [41] with transition dynamics given by $T^{\mathrm{two},L,\alpha}(s'|s,a^{\mathrm{pl}},a^{\mathrm{op}}) = \alpha T^L(s'|s,a^{\mathrm{pl}}) + (1-\alpha)T^L(s'|s,a^{\mathrm{op}})$, where a^{pl} is an action chosen according to the player policy and a^{op} according to the opponent policy. Note that the opponent is restricted to take the worst possible action from the state of the player, i.e., there is no additional state variable for the opponent. As a result, we reach a two-player Markov game with a regularization term for the player as follows:

$$\arg\max_{\pi^{\mathrm{pl}}\in\Pi}\min_{\pi^{\mathrm{op}}\in\Pi}\mathbb{E}\left[G \mid \pi^{\mathrm{pl}}, \pi^{\mathrm{op}}, M^{\mathrm{two},L,\alpha}\right], \tag{9}$$

where $M^{\mathrm{two},L,\alpha} = \left\{\mathcal{S},\mathcal{A},\mathcal{A},T^{\mathrm{two},L,\alpha},\gamma,P_0,R_{\boldsymbol{\theta}}\right\}$ is the two-player MDP associated with the above game. The repetition of the action space $\mathcal{A}$ denotes the fact that player and adversary share the same action space. Inspired from [42], we propose a dynamic programming approach to find the player and opponent policies (see Algorithm 2 in Appendix F.3).

4.4 Performance Gap of Robust MCE IRL

Let π^{pl} be the policy returned by our Algorithm 1 when there is a transition dynamics mismatch. Recall that $\pi_1 := \pi^{\mathrm{soft}}_{M^L_{\boldsymbol{\theta}_L}}$ is the policy recovered without this mismatch. Then, we obtain the following upper-bound[4] for the performance gap of our algorithm via the triangle inequality:

Theorem 3. *The performance gap between the policies π_1 and π^{pl} on the MDP $M^L_{\boldsymbol{\theta}^*}$ is bounded as follows:* $\left|V^{\pi_1}_{M^L_{\boldsymbol{\theta}^*}} - V^{\pi^{\mathrm{pl}}}_{M^L_{\boldsymbol{\theta}^*}}\right| \leq \frac{|R_{\boldsymbol{\theta}^*}|^{\max}}{(1-\gamma)^2}\cdot\left\{\gamma\cdot d_{\mathrm{dyn}}\left(T^L,T^E\right) + 2\cdot(1-\alpha)\right\}$.

However, we now provide a constructive example, in which, by choosing the appropriate value for α, the performance gap of our Algorithm 1 vanishes. In contrast, the performance gap of the standard MCE IRL is proportional to the mismatch. Note that our Algorithm 1 with $\alpha = 1$ corresponds to the standard MCE-IRL algorithm.

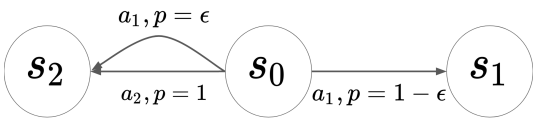


Figure 3: Constructive example to study the performance gap of Algorithm 1 and the MCE IRL.

[4]This bound is worst than the one given in Theorem 1. When the condition in Theorem 2 does not hold, the robust MCE IRL achieves a tighter bound than the MCE IRL for a proper choice of α (see Appendix F.5).

Consider a reference MDP $M^{(\epsilon)} = \left\{\mathcal{S}, \mathcal{A}, T^{(\epsilon)}, \gamma, P_0\right\}$ with variable ϵ (see Figure 3). The state space is $\mathcal{S} = \{s_0, s_1, s_1\}$, where s_1 and s_2 are absorbing states. The action space is $\mathcal{A} = \{a_1, a_2\}$ and the initial state distribution is $P_0\left(s_0\right) = 1$. The transition dynamics is defined as: $T^{(\epsilon)}(s_1|s_0, a_1) = 1-\epsilon$, $T^{(\epsilon)}(s_2|s_0, a_1) = \epsilon$, $T^{(\epsilon)}(s_1|s_0, a_2) = 0$, and $T^{(\epsilon)}(s_2|s_0, a_2) = 1$. The true reward function is given by: $R_{\boldsymbol{\theta}^*}\left(s_0\right) = 0$, $R_{\boldsymbol{\theta}^*}\left(s_1\right) = 1$, and $R_{\boldsymbol{\theta}^*}\left(s_2\right) = -1$. We define the learner and the expert environment as: $M^L := M^{(0)}$ and $M^L := M^{(\epsilon_E)}$. Note that the distance between the two transition dynamics is $d_{\text{dyn}}\left(T^L, T^E\right) = 2\epsilon_E$. Let π^{pl} and $\pi_2 := \pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}$ be the policies returned by Algorithm 1 and the MCE IRL algorithm, under the above mismatch. Recall that π_1 is the policy recovered by the MCE IRL algorithm without this mismatch. Then, the following holds:

Theorem 4. *For this example, the performance gap of Algorithm 1 vanishes by choosing* $\alpha = 1 - \frac{d_{\text{dyn}}(T^L, T^E)}{2}$*, i.e.,* $\left|V^{\pi_1}_{M^L_{\boldsymbol{\theta}^*}} - V^{\pi^{\text{pl}}}_{M^L_{\boldsymbol{\theta}^*}}\right| = 0$*. Whereas, the performance gap of the standard MCE IRL is given by:* $\left|V^{\pi_1}_{M^L_{\boldsymbol{\theta}^*}} - V^{\pi_2}_{M^L_{\boldsymbol{\theta}^*}}\right| = \frac{\gamma}{1-\gamma} \cdot d_{\text{dyn}}(T^L, T^E)$.

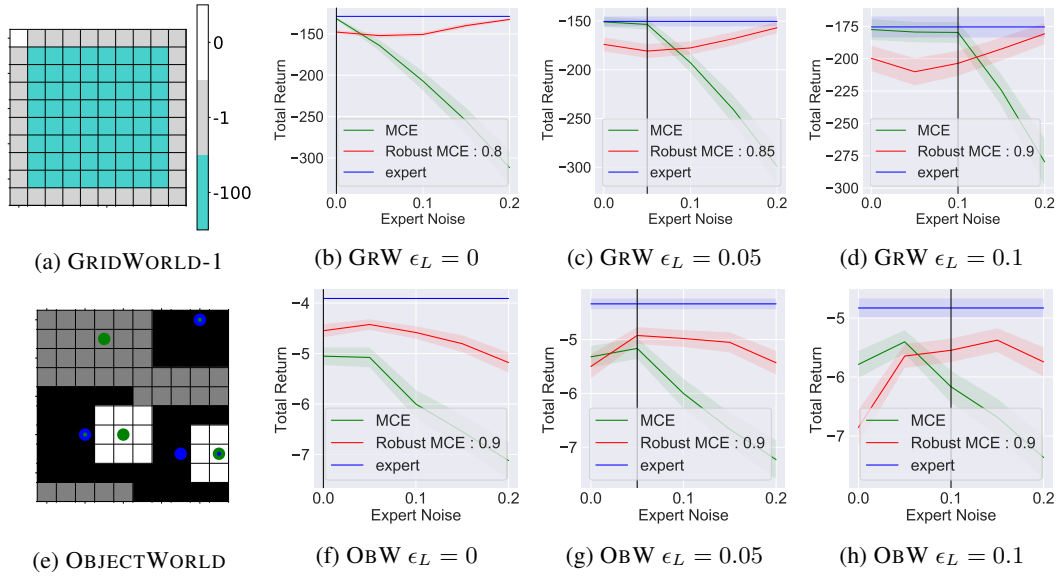
5 Experiments

This section demonstrates the superior performance of our Algorithm 1 compared to the standard MCE IRL algorithm, when there is a transition dynamics mismatch between the expert and the learner. All the missing figures and hyper-parameter details are reported in Appendix G.

Setup. Let $M^{\text{ref}}_{\boldsymbol{\theta}^*} = \left(\mathcal{S}, \mathcal{A}, T^{\text{ref}}, \gamma, P_0, R_{\boldsymbol{\theta}^*}\right)$ be a reference MDP. Given a *learner noise* $\epsilon_L \in [0, 1]$, we introduce a learner MDP without reward function as $M^{L,\epsilon_L} = \left(\mathcal{S}, \mathcal{A}, T^{L,\epsilon_L}, \gamma, P_0\right)$, where $T^{L,\epsilon_L} \in \Delta_{\mathcal{S}|\mathcal{S},\mathcal{A}}$ is defined as $T^{L,\epsilon_L} := (1-\epsilon_L)T^{\text{ref}} + \epsilon_L \bar{T}$ with $\bar{T} \in \Delta_{\mathcal{S}|\mathcal{S},\mathcal{A}}$. Similarly, given an *expert noise* $\epsilon_E \in [0, 1]$, we define an expert MDP $M^{E,\epsilon_E}_{\boldsymbol{\theta}^*} = \left(\mathcal{S}, \mathcal{A}, T^{E,\epsilon_E}, \gamma, P_0, R_{\boldsymbol{\theta}^*}\right)$, where $T^{E,\epsilon_E} \in \Delta_{\mathcal{S}|\mathcal{S},\mathcal{A}}$ is defined as $T^{E,\epsilon_E} := (1-\epsilon_E)T^{\text{ref}} + \epsilon_E \bar{T}$ with $\bar{T} \in \Delta_{\mathcal{S}|\mathcal{S},\mathcal{A}}$. Note that a pair (ϵ_E, ϵ_L) corresponds to an IRL problem under dynamics mismatch, where the expert acts in the MDP $M^{E,\epsilon_E}_{\boldsymbol{\theta}^*}$ and the learner in M^{L,ϵ_L}. In our experiments, we set T^{ref} to be deterministic, and $\bar{T}$ to be uniform. Then, one can easily show that $d_{\text{dyn}}\left(T^{L,\epsilon_L}, T^{E,\epsilon_E}\right) = 2\left(1 - \frac{1}{|\mathcal{S}|}\right)|\epsilon_L - \epsilon_E|$. The learned policies are evaluated in the MDP $M^{L,\epsilon_L}_{\boldsymbol{\theta}^*}$, i.e., M^{L,ϵ_L} endowed with the true reward function $R_{\boldsymbol{\theta}^*}$.

Baselines. We are not aware of any comparable prior IRL work that exactly matches our setting: (i) only one shot access to the expert environment, and (ii) do not explicitly model the expert environment. Note that Algorithm 2 in [33] requires online access to T^E (or the expert environment) to empirically estimate the gradient for every (time step) adversarial expert policy $\check{\pi}^*$, whereas we do not access the expert environment after obtaining a batch of demonstrations, i.e., $\boldsymbol{\rho}$. Thus, for each pair (ϵ_E, ϵ_L), we compare the performance of the following: (i) our robust MCE IRL algorithm with different values of $\alpha \in \{0.8, 0.85, 0.9, 0.95\}$, (ii) the standard MCE IRL algorithm, and (iii) the ideal baseline that utilizes the knowledge of the true reward function, i.e, $\pi^*_{M^{L,\epsilon_L}_{\boldsymbol{\theta}^*}}$.

Environments. We consider four GRIDWORLD environments and an OBJECTWORLD [43] environment. All of them are $N \times N$ grid, where a cell represents a state. There are four actions per state, corresponding to steps in one of the four cardinal directions; T^{ref} is defined accordingly. GRIDWORLD environments are endowed with a linear reward function $R_{\boldsymbol{\theta}^*}(s) = \langle \boldsymbol{\theta}^*, \boldsymbol{\phi}(s)\rangle$, where $\boldsymbol{\phi}$ is a one-hot feature map. The entries $\boldsymbol{\theta}^*_{\boldsymbol{s}}$ of the parameter $\boldsymbol{\theta}^*$ for each state $s \in \mathcal{S}$ are shown in Figures 4a, 10e, 10i, and 10m. OBJECTWORLD is endowed with a non-linear reward function, determined by the distance of the agent to the objects that are randomly placed in the environment. Each object has an outer and an inner color; however, only the former plays a role in determining the reward while the latter serves as a distractor. The reward is -2 in positions within three cells to an outer blue object (black areas of Figure 4e), 0 if they are also within two cells from an outer green object (white areas), and -1 otherwise (gray areas). We shift the rewards originally proposed by [43] to non-positive values, and we randomly placed the goal state in a white area. We also modify the reward features by augmenting them with binary features indicating whether the goal state has been reached. These changes simplify the application of the MCE IRL algorithm in the infinite horizon setting. For this non-linear reward setting, we used the deep MCE IRL algorithm from [44], where the reward function is parameterized by a neural network.



(a) GRIDWORLD-1 (b) GRW $\epsilon_L = 0$ (c) GRW $\epsilon_L = 0.05$ (d) GRW $\epsilon_L = 0.1$

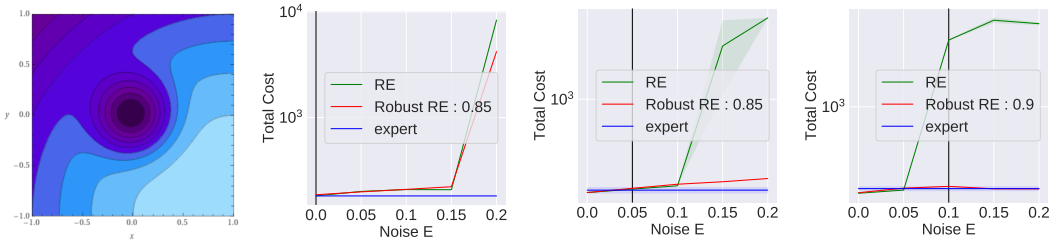
(e) OBJECTWORLD (f) OBW $\epsilon_L = 0$ (g) OBW $\epsilon_L = 0.05$ (h) OBW $\epsilon_L = 0.1$

Figure 4: Comparison of the performance our Algorithm 1 against the baselines, under different levels of mismatch: $(\epsilon_E, \epsilon_L) \in \{0.0, 0.05, 0.1, 0.15, 0.2\} \times \{0.0, 0.05, 0.1\}$. Each plot corresponds to a fixed leaner environment M^{L,ϵ_L} with $\epsilon_L \in \{0.0, 0.05, 0.1\}$. The values of α used for Algorithm 1 are reported in the legend. The vertical line indicates the position of the learner environment in the x-axis. We abbreviated the environment names as GRW, and OBW. Note that our Robust MCE IRL outperforms standard MCE IRL when the expert noise increases along the x-axis. At the same time, Robust MCE IRL might perform slightly worse in the low expert noise regime. This observation aligns with the overly conservative nature of robust training methods.

Results. In Figure 4, we have presented the results for two of the environments, and the complete results can be found in Figure 10. Also, in Figure 4, we have reported the results of our algorithm with the best performing value of α; and the performance of our algorithm with different values of α are presented in Figure 11. In all the plots, every point in the x-axis corresponds to a pair (ϵ_E, ϵ_L). For example, consider Figure 4b, for a fixed learner environment M^{L,ϵ_L} with $\epsilon_L = 0$, and different expert environments M^{E,ϵ_E} by varying ϵ_E along the x-axis. Note that, in this figure, the distance $d_{\text{dyn}}\left(T^{L,\epsilon_L}, T^{E,\epsilon_E}\right) \propto |\epsilon_L - \epsilon_E|$ increases along the x-axis. For each pair (ϵ_E, ϵ_L), in the y-axis, we present the performance of the learned polices in the MDP $M_{\boldsymbol{\theta}^*}^{L,\epsilon_L}$, i.e., $V^{\pi}_{M_{\boldsymbol{\theta}^*}^{L,\epsilon_L}}$. In alignment with our theory, the performance of the standard MCE IRL algorithm degrades along the x-axis. Whereas, our Algorithm 1 resulted in robust performance (even closer to the ideal baseline) across different levels of mismatch. These results confirm the efficacy of our method under mismatch. However, one has to carefully choose the value of $1-\alpha$ (s.t. $T^{E,\epsilon_E} \in \mathcal{T}^{L,\alpha}$): (i) underestimating it would lead to a linear decay in the performance, similar to the MCE IRL, (ii) overestimating it would also slightly hinder the performance, and (iii) given a rough estimate $\widehat{T}^E$ of the expert dynamics, choosing $1-\alpha \approx \frac{d_{\text{dyn}}(T^L, \widehat{T}^E)}{2}$ would lead to better performance in practice. The potential drop in the performance of our Robust MCE IRL method under the low expert noise regime (see Figures 4c, 4d, and 4h) can be related to the overly conservative nature of robust training. See Appendix G.3 for more discussion on the choice of $1-\alpha$. In addition, we have tested our method on a setting with low-dimensional feature mapping $\boldsymbol{\phi}$, where we observed significant improvement over the standard MCE IRL (see Appendix G.2).

6 Extension to Continuous MDP Setting

In this section, we extend our ideas to the continuous MDP setting, i.e., the environments with continuous state and action spaces. In particular, we implement a robust variant of the Relative Entropy IRL (RE IRL) [15] algorithm (see Algorithm 3 in Appendix H). We cannot use the dynamic programming approach to find the player and opponent policies in the continuous MDP setting. Therefore, we solve the two-player Markov game in a model-free manner using the policy gradient methods (see Algorithm 4 in Appendix H).



(a) GAUSSIANGRID (b) M^{L,ϵ_L} with $\epsilon_L = 0$ (c) M^{L,ϵ_L} with $\epsilon_L = 0.05$ (d) M^{L,ϵ_L} with $\epsilon_L = 0.1$

Figure 5: Comparison of the performance our Robust RE IRL (Algorithm 3) against the standard RE IRL, under different levels of mismatch: $(\epsilon_E, \epsilon_L) \in \{0.0, 0.05, 0.1, 0.15, 0.2\} \times \{0.0, 0.05, 0.1\}$. Each plot corresponds to a fixed leaner environment M^{L,ϵ_L} with $\epsilon_L \in \{0.0, 0.05, 0.1\}$. The values of α used for Algorithm 3 are reported in the legend. The vertical line indicates the position of the learner environment in the x-axis. The results are averaged across 5 seeds.

We evaluate the performance of our Robust RE IRL method on a continuous gridworld environment that we called GAUSSIANGRID. The details of the environment and the experimental setup are given in Appendix H. The results are reported in Figure 5, where we notice that our Robust RE IRL method outperforms standard RE IRL.

7 Related Work

In the context of forward RL, there are works that build on the robust MDP framework [28, 29, 45], for example, [46, 47, 48]. However, our work is closer to the line of work that leverages on the equivalence between action-robust and robust MDPs [49, 50, 30, 27, 40]. To our knowledge, this is the first work to adapt the robust RL methods in the IRL context. Other works study the IRL problem under a mismatch between the learner and the expert's worldviews [51, 52]. However, these works do not consider the dynamics mismatch.

Generative Adversarial Imitation Learning (GAIL) [53] and its variants are IRL methods that use a GAN-based reward to align the distribution of the state-action pairs between the expert and the learner. When there is a transition dynamics mismatch, the expert's actions are not quite useful for imitation. [54, 55] have considered state only distribution matching when the expert actions are not observable. Building on these works, [22, 23] have studied the imitation learning problem under transition dynamics mismatch. These works propose model-alignment based imitation learning algorithms in the high dimensional settings to address the dynamics mismatch. Finally, our work has the following important differences with AIRL [56]. In AIRL, the learner has access to the expert environment during the training phase, i.e., there is no transition dynamics mismatch during the training phase but only at test time. In contrast, we consider a different setting where the learner can not access the expert environment during the training phase. In addition, AIRL requires input demonstrations containing both states and actions, while our algorithm requires state-only demonstrations.

8 Conclusions

In this work, we theoretically analyze the MCE IRL algorithm under the transition dynamics mismatch: (i) we derive necessary and sufficient conditions for the existence of solution, and (ii) we provide a tight upper bound on the performance degradation. We propose a robust MCE IRL algorithm and empirically demonstrate its significant improvement over the standard MCE IRL under dynamics mismatch. Even though our Algorithm 1 is not essentially different from the standard robust RL methods, it poses additional theoretical challenges in the IRL context compared to the RL setup. In particular, we have proved: (i) the existence of solution for the robust MCE IRL formulation, and (ii) the performance gap improvement of our algorithm compared to the non-robust MCE IRL in a constructive example. We present empirical results for the settings not covered by our theory: MDPs with non-linear reward function and continuous state and action spaces.

Code Repository

`https://github.com/lviano/RobustMCE_IRL/tree/master/robustIRLcode`

Acknowledgments and Disclosure of Funding

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 725594 - time-data). Research was sponsored by the Army Research Office and was accomplished under Grant Number W911NF-19-1-0404, by the Department of the Navy, Office of Naval Research (ONR) under a grant number N62909-17-1-2111 and by Hasler Foundation Program: Cyber Human Systems (project number 16066). This work has been supported by 2021 gift from the Schindler Group for research excellence in reinforcement learning. This work has received financial support from the Enterprise for Society Center (E4S).

Parameswaran Kamalaruban acknowledges support from The Alan Turing Institute. He carried out part of this work while at LIONS, EPFL.

Adrian Weller acknowledges support from a Turing AI Fellowship under grant EP/V025379/1, The Alan Turing Institute, and the Leverhulme Trust via CFI.

References

[1] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2000.

[2] David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In *Proc. Intl Conf. on Machine Learning (ICML)*, 2014.

[3] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *Proc. Intl Conf. on Machine Learning (ICML)*, 2015.

[4] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[5] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 2015.

[6] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 2017.

[7] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *Proc. Intl Conf. on Learning Representations (ICLR)*, 2016.

[8] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 2016.

[9] Stuart Russell. Learning agents for uncertain environments. In *Proc. Conf. on Learning Theory (COLT)*, 1998.

[10] Andrew Y Ng and Stuart Russell. Algorithms for inverse reinforcement learning. In *Proc. Intl Conf. on Machine Learning (ICML)*, 2000.

[11] Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proc. Intl Conf. on Machine Learning (ICML)*, 2004.

[12] Nathan D Ratliff, J Andrew Bagnell, and Martin A Zinkevich. Maximum margin planning. In *Proc. Intl Conf. on Machine Learning (ICML)*, 2006.

[13] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. In *Proc. AAAI Conference on Artificial Intelligence*, 2008.

[14] Umar Syed and Robert E Schapire. A game-theoretic approach to apprenticeship learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2008.

[15] Abdeslam Boularias, Jens Kober, and Jan Peters. Relative entropy inverse reinforcement learning. In *Proc. Intl Conf. on Artificial Intelligence and Statistics (AISTATS)*, 2011.

[16] T Osa, J Pajarinen, G Neumann, JA Bagnell, P Abbeel, and J Peters. An algorithmic perspective on imitation learning. *Foundations and Trends in Robotics*, 2018.

[17] Gabriel Dulac-Arnold, Daniel Mankowitz, and Todd Hester. Challenges of real-world reinforcement learning. *arXiv preprint arXiv:1904.12901*, 2019.

[18] Chao Yu, Jiming Liu, and Shamim Nemati. Reinforcement learning in healthcare: A survey. *arXiv preprint arXiv:1908.08796*, 2019.

[19] B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A Al Sallab, Senthil Yogamani, and Patrick Pérez. Deep reinforcement learning for autonomous driving: A survey. *arXiv preprint arXiv:2002.00444*, 2020.

[20] Sid Reddy, Anca Dragan, and Sergey Levine. Where do you think you're going?: Inferring beliefs about dynamics from behavior. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[21] Ze Gong and Yu Zhang. What is it you really want of me? generalized reward learning with biased beliefs about domain dynamics. In *Proc. AAAI Conference on Artificial Intelligence*, 2020.

[22] Tanmay Gangwani and Jian Peng. State-only imitation with transition dynamics mismatch. In *Proc. Intl Conf. on Learning Representations (ICLR)*, 2020.

[23] Fangchen Liu, Zhan Ling, Tongzhou Mu, and Hao Su. State alignment-based imitation learning. In *Proc. Intl Conf. on Learning Representations (ICLR)*, 2020.

[24] Brian D Ziebart. *Modeling purposeful adaptive behavior with the principle of maximum causal entropy*. PhD thesis, Carnegie Mellon University, 2010.

[25] Brian D Ziebart, J Andrew Bagnell, and Anind K Dey. The principle of maximum causal entropy for estimating interacting processes. *IEEE Transactions on Information Theory*, 2013.

[26] Zhengyuan Zhou, Michael Bloem, and Nicholas Bambos. Infinite time horizon maximum causal entropy inverse reinforcement learning. *IEEE Transactions on Automatic Control*, 2017.

[27] Chen Tessler, Yonathan Efroni, and Shie Mannor. Action robust reinforcement learning and applications in continuous control. In *Proc. Intl Conf. on Machine Learning (ICML)*, 2019.

[28] Garud N Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 2005.

[29] Arnab Nilim and Laurent El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 2005.

[30] Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *Proc. Intl Conf. on Machine Learning (ICML)*, 2017.

[31] Michael Herman, Tobias Gindele, Jörg Wagner, Felix Schmitt, and Wolfram Burgard. Inverse reinforcement learning with simultaneous estimation of rewards and dynamics. In *Proc. Intl Conf. on Artificial Intelligence and Statistics (AISTATS)*, 2016.

[32] Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized markov decision processes. In *Proc. Intl Conf. on Machine Learning (ICML)*, 2019.

[33] Xiangli Chen, Mathew Monfort, Brian D Ziebart, and Peter Carr. Adversarial inverse optimal control for general imitation learning losses and embodiment transfer. In *Proc. Conf. on Uncertainty in Artificial Intelligence (UAI)*, 2016.

[34] Amy Zhang, Shagun Sodhani, Khimya Khetarpal, and Joelle Pineau. Multi-task reinforcement learning as a hidden-parameter block mdp. *arXiv preprint arXiv:2007.07206*, 2020.

[35] Michael J. Kearns and Satinder P. Singh. Near-optimal reinforcement learning in polynominal time. In *Proc. Intl Conf. on Machine Learning (ICML)*, 1998.

[36] Eyal Even-Dar and Yishay Mansour. Approximate equivalence of Markov decision processes. In *Learning Theory and Kernel Machines*. 2003.

[37] Michael Bloem and Nicholas Bambos. Infinite time horizon maximum causal entropy inverse reinforcement learning. In *IEEE Conference on Decision and Control*, 2014.

[38] I.R. Shafarevich, A.O. Remizov, D.P. Kramer, and L. Nekludova. *Linear Algebra and Geometry*. Springer Berlin Heidelberg, 2014.

[39] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. Intl Conf. on Learning Representations (ICLR)*, 2015.

[40] Parameswaran Kamalaruban, Yu-Ting Huang, Ya-Ping Hsieh, Paul Rolland, Cheng Shi, and Volkan Cevher. Robust reinforcement learning via adversarial training with langevin dynamics. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[41] Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings*. 1994.

[42] Jordi Grau-Moya, Felix Leibfried, and Haitham Bou-Ammar. Balancing two-player stochastic games with soft q-learning. In *Proc. Intl Joint Conf. on Artificial Intelligence (IJCAI)*, 2018.

[43] Sergey Levine, Zoran Popovic, and Vladlen Koltun. Nonlinear inverse reinforcement learning with gaussian processes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2011.

[44] Markus Wulfmeier, Peter Ondruska, and Ingmar Posner. Maximum entropy deep inverse reinforcement learning. *arXiv preprint arXiv:1507.04888*, 2015.

[45] Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust markov decision processes. *Mathematics of Operations Research*, 2013.

[46] Shirli Di-Castro Shashua and Shie Mannor. Deep robust kalman filter. *arXiv preprint arXiv:1703.02310*, 2017.

[47] Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In *IEEE international conference on robotics and automation (ICRA)*, 2018.

[48] Daniel J Mankowitz, Nir Levine, Rae Jeong, Abbas Abdolmaleki, Jost Tobias Springenberg, Yuanyuan Shi, Jackie Kay, Todd Hester, Timothy Mann, and Martin Riedmiller. Robust reinforcement learning for continuous control with model misspecification. In *Proc. Intl Conf. on Learning Representations (ICLR)*, 2020.

[49] Jun Morimoto and Kenji Doya. Robust reinforcement learning. *Neural computation*, 2005.

[50] John C Doyle, Bruce A Francis, and Allen R Tannenbaum. *Feedback control theory*. Courier Corporation, 2013.

[51] Luis Haug, Sebastian Tschiatschek, and Adish Singla. Teaching inverse reinforcement learners via features and demonstrations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[52] Sebastian Tschiatschek, Ahana Ghosh, Luis Haug, Rati Devidze, and Adish Singla. Learner-aware teaching: Inverse reinforcement learning with preferences and constraints. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[53] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[54] Faraz Torabi, Garrett Warnell, and Peter Stone. Generative adversarial imitation from observation. *arXiv preprint arXiv:1807.06158*, 2018.

[55] Wen Sun, Anirudh Vemula, Byron Boots, and J Andrew Bagnell. Provably efficient imitation learning from observation alone. In *Proc. Intl Conf. on Machine Learning (ICML)*, 2019.

[56] Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adverserial inverse reinforcement learning. In *Proc. Intl Conf. on Learning Representations (ICLR)*, 2018.

[57] Alekh Agarwal, Nan Jiang, Sham M Kakade, and Wen Sun. Reinforcement Learning: Theory and Algorithms, 2019.

[58] Abdeslam Boularias and Brahim Chaib-Draa. Bootstrapping apprenticeship learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2010.

[59] Wen Sun, Geoffrey J Gordon, Byron Boots, and J Andrew Bagnell. Dual policy iteration. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[60] Parameswaran Kamalaruban, Rati Devidze, Volkan Cevher, and Adish Singla. Interactive teaching algorithms for inverse reinforcement learning. In *Proc. Intl Joint Conf. on Artificial Intelligence (IJCAI)*, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] The paper is organized exactly according to the contributions listed at the end of the introduction section.
   (b) Did you describe the limitations of your work? [Yes] Yes, we highlight the fact that the worst case guarantees of the robust MCE IRL can be worst than the standard MCE IRL, and describe how a good choice for α is crucial for better performance.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] This work presents a theoretical investigation of the imitation learning under transition dynamics mismatch (see Section 8). As such in the present form there are no direct negative societal impacts of our work. However, in future, when the proposed methods are applied on real-world systems, the practitioner has to be careful with the choice of α.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] We confirm that our paper conforms with the ethics review guidelines.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] For example, see Theorem 2.
   (b) Did you include complete proofs of all theoretical results? [Yes] Complete proofs of all the theoretical results can be found in the Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The code and instructions are included in the supplementary material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We have provided all the training and hyperparameters details in the Experiments section, and in the Appendix.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] This can be seen in all the Figures.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We used an internal cluster with CPU nodes for the experiments; but we do not have an estimate of the total amount of compute.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Appendix structure

Here, we provide an overview on the organization of the appendix:

- Appendix B summarizes the scope and contributions of the paper.
- Appendix C provides a glossary of notation.
- Appendix D provides further details of Section 2. In particular, we show that the expected feature count with one-hot feature map is proportional to the state occupancy measure.
- Appendix E provides further details of Section 3. In particular:
  1. In Appendix E.1, we provide the proof of Lemma 1 (performance difference between two soft optimal policies).
  2. In Appendix E.2, we provide the proof of Theorem 1 (performance gap of MCE IRL under model mismatch).
  3. In Appendix E.3.1, we explain why state-action reward function is not useful under model mismatch.
  4. In Appendix E.3.2, we provide the proof of Theorem 2 (existence of solution for MCE IRL under model mismatch).
  5. In Appendix E.4, we study the performance gap of the reward transfer strategy explained in Section 3.3.
- Appendix F provides further details of Section 4. In particular:
  1. In Appendix F.2, we derive the gradient update for MCE IRL under model mismatch.
  2. In Appendix F.3, we present Algorithm 2, with theoretical support, to solve the Markov Game in Section 4.3.
  3. In Appendix F.4, we provide the proof of Theorem 3 (performance gap of Algorithm 1 under model mismatch).
  4. In Appendix F.5, we study the performance gap of Algorithm 1 under model mismatch in the infeasible case (when exact occupancy measure matching is not possible).
  5. In Appendix F.6, we provide the proof of Theorem 4 (constructive example comparing MCE IRL and Algorithm 1).
- Appendix G provides further details of Section 5. In particular:
  1. In Appendix G.1, we report all the hyperparameter details, and present the figures mentioned in the main text.
  2. In Appendix G.2, we demonstrate superior performance of Algorithm 1 on a low-dimensional feature setting.
  3. In Appendix G.3, we study the impact of the opponent strength parameter $1 - \alpha$ on Robust MCE IRL.
- Appendix H provides further details of Section 6. In particular, we present a high-dimensional continuous control extension of our robust IRL method, and demonstrates its efficacy on a domain with continuous state and spaces under dynamics mismatch.

B Scope and Contributions

Our work is intended to:

1. provide a theoretical investigation of the transition dynamics mismatch issue in the standard MCE IRL formulation, including:
   (a) an upper bound on the performance gap due to dynamics mismatch (Theorem 1) + the tightness of the bound (Theorem 4)
   (b) existence of solution under dynamics mismatch (Theorem 2)
2. illustrate the issues with the reward transfer scheme under transition dynamics mismatch (Theorem 6 + Lemma 1; see Section 3.3, and Appendix E.4)
3. understand the role of robust RL methods in mitigating the mismatch issue
   (a) validity (existence of solution using Theorem 2) of the robust MCE IRL formulation (see Section 4.2)
   (b) an upper bound on the performance gap of robust MCE IRL (Theorem 3) + improvement over standard MCE IRL (Theorem 2)
   (c) an upper bound on the performance gap of robust MCE IRL when exact occupancy measure matching is not possible (Theorem 9)
   (d) different effect of over and underestimating the robustness parameter alpha (see Appendix G.3)
4. empirically validate our claims in a setting (finite MDP) without theory-practice gap (see Section 5, and Appendix G)
5. extend our robust IRL method to the high dimensional continuous MDP setting with appropriate practical relaxations, and empirically demonstrate its effectiveness (see Appendix H).

C Glossary of Notation

We have carefully developed the notation based on the best practices prescribed by the RL theory community [57], and do not want to compromise its rigorous nature. To help the reader, we provide a glossary of notation.

$\pi^*_{M^L_{\boldsymbol{\theta}^*}}$	optimal policy in the MDP $M^L_{\boldsymbol{\theta}^*} = \{\mathcal{S}, \mathcal{A}, T^L, \gamma, P_0, R_{\boldsymbol{\theta}^*}\}$
$\pi^*_{M^E_{\boldsymbol{\theta}^*}}$	optimal policy in the MDP $M^E_{\boldsymbol{\theta}^*} = \{\mathcal{S}, \mathcal{A}, T^E, \gamma, P_0, R_{\boldsymbol{\theta}^*}\}$
$\boldsymbol{\rho}_{M^L}^{\pi^*_{M^L_{\boldsymbol{\theta}^*}}}$	state occupancy measure of $\pi^*_{M^L_{\boldsymbol{\theta}^*}}$ in the MDP $M^L = \{\mathcal{S}, \mathcal{A}, T^L, \gamma, P_0\}$
$\boldsymbol{\rho}_{M^E}^{\pi^*_{M^E_{\boldsymbol{\theta}^*}}}$	state occupancy measure of $\pi^*_{M^E_{\boldsymbol{\theta}^*}}$ in the MDP $M^E = \{\mathcal{S}, \mathcal{A}, T^E, \gamma, P_0\}$
$\boldsymbol{\theta_L}$	reward parameter recovered when there is no transition dynamics mismatch
$\boldsymbol{\theta_E}$	reward parameter recovered under transition dynamics mismatch
$\pi_1 = \pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_L}}}$	soft optimal policy in the MDP $M^L_{\boldsymbol{\theta_L}} = \{\mathcal{S}, \mathcal{A}, T^L, \gamma, P_0, R_{\boldsymbol{\theta_L}}\}$
$\pi_2 = \pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}$	soft optimal policy in the MDP $M^L_{\boldsymbol{\theta_E}} = \{\mathcal{S}, \mathcal{A}, T^L, \gamma, P_0, R_{\boldsymbol{\theta_E}}\}$
$V^{\pi_1}_{M^L_{\boldsymbol{\theta}^*}}$	total expected return of π_1 in the MDP $M^L_{\boldsymbol{\theta}^*} = \{\mathcal{S}, \mathcal{A}, T^L, \gamma, P_0, R_{\boldsymbol{\theta}^*}\}$
$V^{\pi_2}_{M^L_{\boldsymbol{\theta}^*}}$	total expected return of π_2 in the MDP $M^L_{\boldsymbol{\theta}^*} = \{\mathcal{S}, \mathcal{A}, T^L, \gamma, P_0, R_{\boldsymbol{\theta}^*}\}$
$\boldsymbol{\rho}^{\pi^{\text{pl}}}_{M^{L,\alpha}}$	state occupancy measure of π^{pl} in the MDP $M^{L,\alpha} = \{\mathcal{S}, \mathcal{A}, T^{L,\alpha}, \gamma, P_0\}$
$\boldsymbol{\rho}_{M^L}^{\alpha\pi^{\text{pl}}+(1-\alpha)\pi^{\text{op}}}$	state occupancy measure of $\alpha\pi^{\text{pl}} + (1-\alpha)\pi^{\text{op}}$ in the MDP $M^L = \{\mathcal{S}, \mathcal{A}, T^L, \gamma, P_0\}$

Table 1: A glossary of notation.

D Further Details of Section 2

An optimal policy $\pi^*_{M_{\boldsymbol{\theta}}}$ in the MDP $M_{\boldsymbol{\theta}}$ satisfies the following *Bellman optimality equations* for all the state-action pairs $(s, a) \in \mathcal{S} \times \mathcal{A}$:

$$\begin{aligned}
\pi^*_{M_{\boldsymbol{\theta}}}(s) &= \arg\max_a Q^*_{M_{\boldsymbol{\theta}}}(s,a) \\
Q^*_{M_{\boldsymbol{\theta}}}(s,a) &= R_{\boldsymbol{\theta}}(s) + \gamma \sum_{s'} T(s'|s,a) V^*_{M_{\boldsymbol{\theta}}}(s') \\
V^*_{M_{\boldsymbol{\theta}}}(s) &= \max_a Q^*_{M_{\boldsymbol{\theta}}}(s,a)
\end{aligned}$$

The soft-optimal policy $\pi^{\text{soft}}_{M_{\boldsymbol{\theta}}}$ in the MDP $M_{\boldsymbol{\theta}}$ satisfies the following *soft Bellman optimality equations* for all the state-action pairs $(s, a) \in \mathcal{S} \times \mathcal{A}$:

$$\begin{aligned}
\pi^{\text{soft}}_{M_{\boldsymbol{\theta}}}(a|s) &= \exp\left(Q^{\text{soft}}_{M_{\boldsymbol{\theta}}}(s,a) - V^{\text{soft}}_{M_{\boldsymbol{\theta}}}(s)\right) \\
Q^{\text{soft}}_{M_{\boldsymbol{\theta}}}(s,a) &= R_{\boldsymbol{\theta}}(s) + \gamma \sum_{s'} T(s'|s,a) V^{\text{soft}}_{M_{\boldsymbol{\theta}}}(s') \\
V^{\text{soft}}_{M_{\boldsymbol{\theta}}}(s) &= \log \sum_a \exp Q^{\text{soft}}_{M_{\boldsymbol{\theta}}}(s,a)
\end{aligned}$$

The expected feature count of a policy π in the MDP M is defined as $\bar{\boldsymbol{\phi}}^{\pi}_M := \mathop{\mathbb{E}}_{\pi,M}\left[\sum_{t=0}^{\infty} \gamma^t \boldsymbol{\phi}(s_t)\right]$.

Fact 1. *If $\forall s \in \mathcal{S}$, $\boldsymbol{\phi}(s) \in \mathbb{R}^{|\mathcal{S}|}$ is a one-hot vector with only the element in position s being 1, then the expected feature count of a policy π in the MDP M is proportional to its state occupancy measure vector in the MDP M.*

Proof. For any M, π, we have:

$$\begin{aligned}
\bar{\boldsymbol{\phi}}^{\pi}_M &= \mathop{\mathbb{E}}_{\pi,M}\left[\sum_{t=0}^{\infty} \gamma^t \boldsymbol{\phi}(s_t)\right] \\
&= \mathop{\mathbb{E}}_{\pi,M}\left[\sum_{t=0}^{\infty} \gamma^t \sum_{s \in \mathcal{S}} \boldsymbol{\phi}(s) \mathbb{1}\left[s = s_t\right]\right] \\
&= \sum_{s \in \mathcal{S}} \boldsymbol{\phi}(s) \mathop{\mathbb{E}}_{\pi,M}\left[\sum_{t=0}^{\infty} \gamma^t \mathbb{1}\left[s = s_t\right]\right] \\
&= \sum_{s \in \mathcal{S}} \boldsymbol{\phi}(s) \sum_{t=0}^{\infty} \gamma^t \mathop{\mathbb{E}}_{\pi,M}\left[\mathbb{1}\left[s = s_t\right]\right] \\
&= \frac{1}{1-\gamma} \sum_{s \in \mathcal{S}} \rho^{\pi}_M(s) \boldsymbol{\phi}(s)
\end{aligned}$$

For the one-hot feature map, ignoring the normalizing factor, the above sum of vectors can be written as follows:

$$\left[\rho^{\pi}_M(s_1), \rho^{\pi}_M(s_2), \cdots\right]^{\top} = \boldsymbol{\rho}^{\pi}_M.$$

□

Leveraging on this fact, we formulate the MCE IRL problem (1) with the state occupancy measure $\boldsymbol{\rho}$ match rather than the usual expected feature count match. Note that if the occupancy measure match is attained, then the match of any expected feature count is also attained.

E Further Details of Section 3

E.1 Proof of Lemma 1

Proof. The soft-optimal policy of the MDP $M'_{\boldsymbol{\theta}}$ satisfies the following soft Bellman optimality equations:

$$\pi'(a|s) = \frac{Z'_{a|s}}{Z'_s} \tag{10}$$

$$\log Z'_s = \log \sum_a Z'_{a|s}$$

$$\log Z'_{a|s} = R_{\boldsymbol{\theta}}(s) + \gamma \sum_{s'} T'(s'|a,s) \log Z'_{s'} \tag{11}$$

Analogously, the soft-optimal policy of the MDP $M_{\boldsymbol{\theta}}$ satisfies the following soft Bellman optimality equations:

$$\pi(a|s) = \frac{Z_{a|s}}{Z_s} \tag{12}$$

$$\log Z_s = \log \sum_a Z_{a|s}$$

$$\log Z_{a|s} = R_{\boldsymbol{\theta}}(s) + \gamma \sum_{s'} T(s'|a,s) \log Z_{s'} \tag{13}$$

For any $s \in \mathcal{S}$, we have:

$$\begin{aligned} D_{\mathrm{KL}}\left(\pi'(\cdot|s), \pi(\cdot|s)\right) &= \sum_a \pi'(a|s) \log \frac{\pi'(a|s)}{\pi(a|s)} \\ &= \sum_a \frac{Z'_{a|s}}{Z'_s} \left(\log \frac{Z'_{a|s}}{Z_{a|s}} + \log \frac{Z_s}{Z'_s} \right) \\ &= \sum_a \frac{Z'_{a|s}}{Z'_s} \log \frac{Z'_{a|s}}{Z_{a|s}} + \log \frac{Z_s}{Z'_s} \end{aligned} \tag{14}$$

By using the log-sum inequality on the term depending on the states only:

$$\begin{aligned} \log \frac{Z_s}{Z'_s} &= \underbrace{\sum_a \frac{Z_{a|s}}{Z_s}}_{1} \log \frac{Z_s}{Z'_s} \\ &= \sum_a \frac{Z_{a|s}}{Z_s} \log \frac{\sum_a Z_{a|s}}{\sum_a Z'_{a|s}} \\ &\leq \frac{1}{Z_s} \sum_a Z_{a|s} \log \frac{Z_{a|s}}{Z'_{a|s}} \end{aligned} \tag{15}$$

Consequently, replacing (15) in (14), and using the definitions (12) and (10), we have:

$$\begin{aligned} D_{\mathrm{KL}}\left(\pi'(\cdot|s), \pi(\cdot|s)\right) &\leq \sum_a \left(\frac{Z'_{a|s}}{Z'_s} - \frac{Z_{a|s}}{Z_s} \right) \log \frac{Z'_{a|s}}{Z_{a|s}} \\ &= \sum_a \left(\pi'(a|s) - \pi(a|s)\right) \log \frac{Z'_{a|s}}{Z_{a|s}} \\ &\leq \sum_a |\pi'(a|s) - \pi(a|s)| \cdot \left| \log \frac{Z'_{a|s}}{Z_{a|s}} \right| \\ &\leq \sum_{a'} |\pi'(a'|s) - \pi(a'|s)| \cdot \max_a \left| \log \frac{Z'_{a|s}}{Z_{a|s}} \right| \end{aligned}$$

$$= \|\pi'(\cdot|s) - \pi(\cdot|s)\|_1 \cdot \max_a \left| \log \frac{Z'_{a|s}}{Z_{a|s}} \right|$$

Then, by taking $\max$ over s, we have:

$$\max_s D_{\mathrm{KL}}\left(\pi'(\cdot|s), \pi(\cdot|s)\right) \;\leq\; \max_s \|\pi'(\cdot|s) - \pi(\cdot|s)\|_1 \cdot \max_{s,a} \left| \log \frac{Z'_{a|s}}{Z_{a|s}} \right| \tag{16}$$

Further, we exploit the following fact:

$$\max_{s,a} \left| \log \frac{Z'_{a|s}}{Z_{a|s}} \right| \;=\; \max \left\{ \log \frac{Z'_{\bar{a}|\bar{s}}}{Z_{\bar{a}|\bar{s}}}, \log \frac{Z_{\underline{a}|\underline{s}}}{Z'_{\underline{a}|\underline{s}}} \right\}, \tag{17}$$

where we adopted the following notation:

$$(\bar{s}, \bar{a}) \;=\; \arg\max_{s,a} \log \frac{Z'_{a|s}}{Z_{a|s}} \tag{18}$$

$$(\underline{s}, \underline{a}) \;=\; \arg\min_{s,a} \log \frac{Z'_{a|s}}{Z_{a|s}} \tag{19}$$

At this point, we can bound separately the two arguments of the max in (17). Starting from (18):

$$\begin{aligned}
\log \frac{Z'_{\bar{a}|\bar{s}}}{Z_{\bar{a}|\bar{s}}} \;&=\; \log Z'_{\bar{a}|\bar{s}} - \log Z_{\bar{a}|\bar{s}} \\
&=\; \underbrace{R_{\boldsymbol{\theta}}(\bar{s}) - R_{\boldsymbol{\theta}}(\bar{s})}_{0} + \gamma \left\{ \sum_{s'} T'(s'|\bar{s}, \bar{a}) \log Z'_{s'} - T(s'|\bar{s}, \bar{a}) \log Z_{s'} \right\} \\
&=\; \gamma \left\{ \sum_{s'} T'(s'|\bar{s}, \bar{a}) \log \frac{Z'_{s'}}{Z_{s'}} + \left(T'(s'|\bar{s}, \bar{a}) - T(s'|\bar{s}, \bar{a})\right) \log Z_{s'} \right\} \\
&\leq\; \gamma \left\{ \sum_{s'} T'(s'|\bar{s}, \bar{a}) \left(\sum_a \pi'(a|s') \log \frac{Z'_{a|s'}}{Z_{a|s'}} \right) + \left(T'(s'|\bar{s}, \bar{a}) - T(s'|\bar{s}, \bar{a})\right) \log Z_{s'} \right\} \\
&\leq\; \gamma \log \frac{Z'_{\bar{a}|\bar{s}}}{Z_{\bar{a}|\bar{s}}} + \gamma \sum_{s'} \left(T'(s'|\bar{s}, \bar{a}) - T(s'|\bar{s}, \bar{a})\right) \log Z_{s'}
\end{aligned}$$

By rearranging the terms, we get:

$$\begin{aligned}
\log \frac{Z'_{\bar{a}|\bar{s}}}{Z_{\bar{a}|\bar{s}}} \;&\leq\; \frac{\gamma}{1-\gamma} \cdot \sum_{s'} \left(T'(s'|\bar{s}, \bar{a}) - T(s'|\bar{s}, \bar{a})\right) \log Z_{s'} \\
&\leq\; \frac{\gamma}{1-\gamma} \cdot \sum_{s'} \left|T'(s'|\bar{s}, \bar{a}) - T(s'|\bar{s}, \bar{a})\right| \cdot \left|\log Z_{s'}\right| \\
&\leq\; \frac{\gamma}{1-\gamma} \cdot \max_{s'} \left|\log Z_{s'}\right| \cdot \sum_{s'} \left|T'(s'|\bar{s}, \bar{a}) - T(s'|\bar{s}, \bar{a})\right|
\end{aligned} \tag{20}$$

Then, with analogous calculations for the second argument of the max operator in (17), we have

$$\begin{aligned}
\log \frac{Z_{\underline{a}|\underline{s}}}{Z'_{\underline{a}|\underline{s}}} \;&=\; \log Z_{\underline{a}|\underline{s}} - \log Z'_{\underline{a}|\underline{s}} \\
&=\; \underbrace{R_{\boldsymbol{\theta}}(\underline{s}) - R_{\boldsymbol{\theta}}(\underline{s})}_{0} + \gamma \left\{ \sum_{s'} T(s'|\underline{s}, \underline{a}) \log Z_{s'} - T'(s'|\underline{s}, \underline{a}) \log Z'_{s'} \right\} \\
&=\; \gamma \left\{ \sum_{s'} T(s'|\underline{s}, \underline{a}) \log \frac{Z_{s'}}{Z'_{s'}} + \left(T(s'|\underline{s}, \underline{a}) - T'(s'|\underline{s}, \underline{a})\right) \log Z'_{s'} \right\}
\end{aligned}$$

$$\leq \gamma \log \frac{Z_{\underline{a}|\underline{s}}}{Z'_{\underline{a}|\underline{s}}} + \gamma \sum_{s'} \left(T(s'|\underline{s},\underline{a}) - T'(s'|\underline{s},\underline{a})\right) \log Z'_{s'}$$

It follows that:

$$\begin{aligned}
\log \frac{Z_{\underline{a}|\underline{s}}}{Z'_{\underline{a}|\underline{s}}} &\leq \frac{\gamma}{1-\gamma} \cdot \sum_{s'} \left(T(s'|\underline{s},\underline{a}) - T'(s'|\underline{s},\underline{a})\right) \log Z'_{s'} \\
&\leq \frac{\gamma}{1-\gamma} \cdot \sum_{s'} |T(s'|\underline{s},\underline{a}) - T'(s'|\underline{s},\underline{a})| \cdot |\log Z'_{s'}| \\
&\leq \frac{\gamma}{1-\gamma} \cdot \max_{s'} |\log Z'_{s'}| \cdot \sum_{s'} |T(s'|\underline{s},\underline{a}) - T'(s'|\underline{s},\underline{a})| \qquad (21)
\end{aligned}$$

We can plug in the bounds obtained in (21) and (20) in (17):

$$\max_{s,a} \left| \log \frac{Z'_{a|s}}{Z_{a|s}} \right| \leq \frac{\gamma}{1-\gamma} \cdot \max \left\{ \max_{s'} |\log Z_{s'}|, \max_{s'} |\log Z'_{s'}| \right\} \cdot \max_{s,a} \sum_{s'} |T'(s'|s,a) - T(s'|s,a)| \qquad (22)$$

It still remains to bound the term $\max\{\max_{s'} |\log Z_{s'}|, \max_{s'} |\log Z'_{s'}|\}$. It can be done by a splitting procedure similar to the one in (17). Indeed:

$$\max_{s'} |\log Z_{s'}| = \max \left\{ \log Z_{\bar{s}}, \log \frac{1}{Z_{\underline{s}}} \right\} \qquad (23)$$

where, changing the previous definitions of $\bar{s}$ and $\underline{s}$, we set:

$$\bar{s} = \arg\max_{s} \log Z_s \qquad (24)$$

$$\underline{s} = \arg\min_{s} \log Z_s \qquad (25)$$

Starting from the first term in (23) and applying (13):

$$\begin{aligned}
\log Z_{\bar{s}} &= \log \sum_{a} Z_{a|\bar{s}} \\
&\leq \log \left(|\mathcal{A}| \max_{a} Z_{a|\bar{s}} \right) \\
&= \log |\mathcal{A}| + \log \max_{a} Z_{a|\bar{s}} \\
&= \log |\mathcal{A}| + \max_{a} \log Z_{a|\bar{s}} \qquad (26)
\end{aligned}$$

where the last equality follows from the fact that log is a monotonically increasing function. Furthermore, (24) implies that $\log Z_{s'} \leq \log Z_{\bar{s}}, \quad \forall s' \in \mathcal{S}$:

$$\begin{aligned}
\max_{a} \log Z_{a|\bar{s}} &\leq \max_{a} \left(R_{\boldsymbol{\theta}}(\bar{s}) + \gamma \log Z_{\bar{s}} \sum_{s'} T(s'|\bar{s},a) \right) \\
&\leq R_{\boldsymbol{\theta}}^{\max} + \gamma \log Z_{\bar{s}} \qquad (27)
\end{aligned}$$

In the last inequality we have used the quantity $R_{\boldsymbol{\theta}}^{\max}$ that satisfies $R_{\boldsymbol{\theta}}(s) \leq R_{\boldsymbol{\theta}}^{\max}, \quad \forall s \in \mathcal{S}$. In a similar fashion, we will use $R_{\boldsymbol{\theta}}^{\min}$ such that $R_{\boldsymbol{\theta}}(s) \geq R_{\boldsymbol{\theta}}^{\min}, \quad \forall s \in \mathcal{S}$. Finally, plugging (27) into (26), we get:

$$\log Z_{\bar{s}} \leq \frac{R_{\boldsymbol{\theta}}^{\max} + \log |\mathcal{A}|}{1-\gamma} \qquad (28)$$

We can proceed bounding the second argument of the max operator in (23). To this scope, we observe that $\sum_a \frac{1}{|\mathcal{A}|} = 1$, and, then, we apply the log-sum inequality as follows:

$$\begin{aligned}
\log \frac{1}{Z_{\underline{s}}} &= \sum_{a} \frac{1}{|\mathcal{A}|} \log \frac{\sum_a \frac{1}{|\mathcal{A}|}}{\sum_a Z_{a|\underline{s}}} \\
&\leq \sum_{a} \frac{1}{|\mathcal{A}|} \log \frac{\frac{1}{|\mathcal{A}|}}{Z_{a|\underline{s}}}
\end{aligned}$$

$$
\begin{aligned}
&= \log \frac{1}{|\mathcal{A}|} + \sum_a \frac{1}{|\mathcal{A}|} \log \frac{1}{Z_{a|\underline{s}}} \\
&\leq \log \frac{1}{|\mathcal{A}|} + \max_a \log \frac{1}{Z_{a|\underline{s}}} \qquad (29)
\end{aligned}
$$

Similarly to (27), we can apply one step of the soft Bellman equation to bound the term $\log \frac{1}{Z_{a|\underline{s}}}$:

$$
\begin{aligned}
\log \frac{1}{Z_{a|\underline{s}}} &= -\log Z_{a|\underline{s}} \\
&= -R_{\boldsymbol{\theta}}(\underline{s}) - \gamma \sum_{s'} T(s'|\underline{s}, a) \log Z_{s'} \\
&= -R_{\boldsymbol{\theta}}(\underline{s}) + \gamma \sum_{s'} T(s'|\underline{s}, a) \log \frac{1}{Z_{s'}} \\
&\leq -R_{\boldsymbol{\theta}}^{\min} + \gamma \log \frac{1}{Z_{\underline{s}}} \underbrace{\sum_{s'} T(s'|\underline{s}, a)}_{1} \qquad (30)
\end{aligned}
$$

where in the last inequality we used (25), $R_{\boldsymbol{\theta}}(s) \geq R_{\boldsymbol{\theta}}^{\min}, \quad \forall s \in \mathcal{S}$. Since the upper bound in (30) does not depend on a, we have:

$$
\max_a \log \frac{1}{Z_{a|\underline{s}}} \leq -R_{\boldsymbol{\theta}}^{\min} + \gamma \log \frac{1}{Z_{\underline{s}}} \qquad (31)
$$

Replacing (31) into (29), we have:

$$
\log \frac{1}{Z_{\underline{s}}} \leq \log \frac{1}{|\mathcal{A}|} - R_{\boldsymbol{\theta}}^{\min} + \gamma \log \frac{1}{Z_{\underline{s}}}
$$

and, consequently:

$$
\log \frac{1}{Z_{\underline{s}}} \leq \frac{-\log |\mathcal{A}| - R_{\boldsymbol{\theta}}^{\min}}{1-\gamma} \qquad (32)
$$

Finally, using (28) and (32) in (23):

$$
\max_{s'} |\log Z_{s'}| \leq \frac{1}{1-\gamma} \cdot \max \left\{ R_{\boldsymbol{\theta}}^{\max} + \log |\mathcal{A}|, -\log |\mathcal{A}| - R_{\boldsymbol{\theta}}^{\min} \right\} \qquad (33)
$$

In addition, one can notice that the bound (33) holds also for $\max_{s'} |\log Z'_{s'}|$:

$$
\max_{s'} |\log Z'_{s'}| \leq \frac{1}{1-\gamma} \cdot \max \left\{ R_{\boldsymbol{\theta}}^{\max} + \log |\mathcal{A}|, -\log |\mathcal{A}| - R_{\boldsymbol{\theta}}^{\min} \right\}
$$

Thus, we can finally replace (33) in (22) that gives:

$$
\max_{s,a} \left| \log \frac{Z'_{a|s}}{Z_{a|s}} \right| \leq \frac{\gamma}{(1-\gamma)^2} \cdot \max \left\{ R_{\boldsymbol{\theta}}^{\max} + \log |\mathcal{A}|, -\log |\mathcal{A}| - R_{\boldsymbol{\theta}}^{\min} \right\} \cdot \max_{s,a} \sum_{s'} |T'(s'|s,a) - T(s'|s,a)| \qquad (34)
$$

We can now go back through the inequality chain to eventually state the bound in the Theorem. First, plugging in (34) into (16) gives:

$$
\max_s D_{\mathrm{KL}} \left(\pi'(\cdot|s), \pi(\cdot|s) \right) \leq \frac{\max_s \|\pi'(\cdot|s) - \pi(\cdot|s)\|_1 \cdot \kappa_{\boldsymbol{\theta}}^2}{(1-\gamma)^2} \cdot d_{\mathrm{dyn}} \left(T', T \right) \qquad (35)
$$

First, by using Pinsker's inequality and the fact that $\max_s \|\pi'(\cdot|s) - \pi(\cdot|s)\|_1 \leq 2$, we get:

$$
\max_s \|\pi'(\cdot|s) - \pi(\cdot|s)\|_1 \leq \sqrt{2 \max_s D_{\mathrm{KL}} \left(\pi'(\cdot|s), \pi(\cdot|s) \right)} \leq \frac{2 \cdot \kappa_{\boldsymbol{\theta}}}{(1-\gamma)} \cdot \sqrt{d_{\mathrm{dyn}} \left(T', T \right)}
$$

Similarly, by using Pinsker's inequality, we get:

$$
\max_s \|\pi'(\cdot|s) - \pi(\cdot|s)\|_1 \leq \sqrt{2 \max_s D_{\mathrm{KL}} \left(\pi'(\cdot|s), \pi(\cdot|s) \right)} \leq \frac{\kappa_{\boldsymbol{\theta}}}{(1-\gamma)} \cdot \sqrt{2 \max_s \|\pi'(\cdot|s) - \pi(\cdot|s)\|_1 \, d_{\mathrm{dyn}} \left(T', T \right)}
$$

Thus, we have:

$$\max_s \|\pi'(\cdot|s) - \pi(\cdot|s)\|_1 \;\leq\; \frac{2 \cdot \kappa_{\boldsymbol{\theta}}^2}{(1-\gamma)^2} \cdot d_{\text{dyn}}\left(T', T\right)$$

Finally, we get:

$$d_{\text{pol}}\left(\pi', \pi\right) \;\leq\; 2 \min \left\{ \frac{\kappa_{\boldsymbol{\theta}} \cdot \sqrt{d_{\text{dyn}}\left(T', T\right)}}{(1-\gamma)}, \frac{\kappa_{\boldsymbol{\theta}}^2 \cdot d_{\text{dyn}}\left(T', T\right)}{(1-\gamma)^2} \right\}$$

□

E.2 Proof of Theorem 1

Proof. Consider the following:

$$\begin{aligned}
\left| V^{\pi_1}_{M^L_{\boldsymbol{\theta}^*}} - V^{\pi_2}_{M^L_{\boldsymbol{\theta}^*}} \right| &\leq \left| V^{\pi_1}_{M^L_{\boldsymbol{\theta}^*}} - V^{\pi^*_{M^L_{\boldsymbol{\theta}^*}}}_{M^L_{\boldsymbol{\theta}^*}} \right| + \left| V^{\pi^*_{M^L_{\boldsymbol{\theta}^*}}}_{M^L_{\boldsymbol{\theta}^*}} - V^{\pi^*_{M^E_{\boldsymbol{\theta}^*}}}_{M^E_{\boldsymbol{\theta}^*}} \right| + \left| V^{\pi^*_{M^E_{\boldsymbol{\theta}^*}}}_{M^E_{\boldsymbol{\theta}^*}} - V^{\pi_2}_{M^L_{\boldsymbol{\theta}^*}} \right| \\
&= \frac{1}{1-\gamma} \left| \left\langle \boldsymbol{\theta}^*, \boldsymbol{\rho}^{\pi_1}_{M^L} - \boldsymbol{\rho}^{\pi^*_{M^L_{\boldsymbol{\theta}^*}}}_{M^L} \right\rangle \right| + \left| V^{\pi^*_{M^L_{\boldsymbol{\theta}^*}}}_{M^L_{\boldsymbol{\theta}^*}} - V^{\pi^*_{M^E_{\boldsymbol{\theta}^*}}}_{M^E_{\boldsymbol{\theta}^*}} \right| + \frac{1}{1-\gamma} \left| \left\langle \boldsymbol{\theta}^*, \boldsymbol{\rho}^{\pi^*_{M^E_{\boldsymbol{\theta}^*}}}_{M^E} - \boldsymbol{\rho}^{\pi_2}_{M^L} \right\rangle \right| \\
&= \left| V^{\pi^*_{M^L_{\boldsymbol{\theta}^*}}}_{M^L_{\boldsymbol{\theta}^*}} - V^{\pi^*_{M^E_{\boldsymbol{\theta}^*}}}_{M^E_{\boldsymbol{\theta}^*}} \right| \\
&\leq \frac{\gamma \cdot |R_{\boldsymbol{\theta}^*}|^{\max}}{(1-\gamma)^2} \cdot d_{\text{dyn}}\left(T^L, T^E\right)
\end{aligned}$$

The first and third terms vanish, since:

1. π_1 is the optimal (thus feasible) solution to the optimization problem (1) with $\boldsymbol{\rho} \leftarrow \boldsymbol{\rho}^{\pi^*_{M^L_{\boldsymbol{\theta}^*}}}_{M^L}$, and
2. π_2 is the optimal (thus feasible) solution to the optimization problem (1) with $\boldsymbol{\rho} \leftarrow \boldsymbol{\rho}^{\pi^*_{M^E_{\boldsymbol{\theta}^*}}}_{M^E}$.

The last inequality is obtained from the Bellman optimality condition (see Theorem 7 in [34]). □

For completeness, we restate Theorem 7 in [34] adapting the notation to our framework and considering bounded rewards instead of normalized rewards as in [34].

Theorem 5 (Theorem 7 in [34])**.** *Consider two MDPs $M_1 = \{\mathcal{S}, \mathcal{A}, T_1, \gamma, P_0, R\}$ and $M_2 = \{\mathcal{S}, \mathcal{A}, T_2, \gamma, P_0, R\}$ with bounded reward function $|R| \leq |R|^{\max}$ and policies π_1^* optimal in M_1 and π_2^* optimal in M_2. Then, we have that:*

$$|V^{\pi_1^*}_{M_1} - V^{\pi_2^*}_{M_2}| \leq \frac{\gamma \cdot |R|^{\max}}{(1-\gamma)^2} \cdot d_{\text{dyn}}\left(T_1, T_2\right). \tag{36}$$

When the expert policy is soft-optimal, we use Lemma 1 and Simulation Lemma [35, 36] to obtain the following bound on the performance gap:

$$\begin{aligned}
\left| V^{\pi_1}_{M^L_{\boldsymbol{\theta}^*}} - V^{\pi_2}_{M^L_{\boldsymbol{\theta}^*}} \right| &\leq \left| V^{\pi_1}_{M^L_{\boldsymbol{\theta}^*}} - V^{\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta}^*}}}_{M^L_{\boldsymbol{\theta}^*}} \right| + \left| V^{\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta}^*}}}_{M^L_{\boldsymbol{\theta}^*}} - V^{\pi^{\text{soft}}_{M^E_{\boldsymbol{\theta}^*}}}_{M^E_{\boldsymbol{\theta}^*}} \right| + \left| V^{\pi^{\text{soft}}_{M^E_{\boldsymbol{\theta}^*}}}_{M^E_{\boldsymbol{\theta}^*}} - V^{\pi_2}_{M^L_{\boldsymbol{\theta}^*}} \right| \\
&= \left| \left\langle \boldsymbol{\theta}^*, \boldsymbol{\rho}^{\pi_1}_{M^L} - \boldsymbol{\rho}^{\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta}^*}}}_{M^L} \right\rangle \right| + \left| V^{\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta}^*}}}_{M^L_{\boldsymbol{\theta}^*}} - V^{\pi^{\text{soft}}_{M^E_{\boldsymbol{\theta}^*}}}_{M^E_{\boldsymbol{\theta}^*}} \right| + \left| \left\langle \boldsymbol{\theta}^*, \boldsymbol{\rho}^{\pi^{\text{soft}}_{M^E_{\boldsymbol{\theta}^*}}}_{M^E} - \boldsymbol{\rho}^{\pi_2}_{M^L} \right\rangle \right| \\
&= \left| V^{\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta}^*}}}_{M^L_{\boldsymbol{\theta}^*}} - V^{\pi^{\text{soft}}_{M^E_{\boldsymbol{\theta}^*}}}_{M^E_{\boldsymbol{\theta}^*}} \right|
\end{aligned}$$

$$
\begin{aligned}
&\leq \left| V^{\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta}^*}}}_{M^L_{\boldsymbol{\theta}^*}} - V^{\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta}^*}}}_{M^E_{\boldsymbol{\theta}^*}} \right| + \left| V^{\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta}^*}}}_{M^E_{\boldsymbol{\theta}^*}} - V^{\pi^{\text{soft}}_{M^E_{\boldsymbol{\theta}^*}}}_{M^E_{\boldsymbol{\theta}^*}} \right| \\
&\leq \frac{\gamma \cdot |R_{\boldsymbol{\theta}^*}|^{\max}}{(1-\gamma)^2} \cdot d_{\text{dyn}}\left(T^L, T^E\right) + \frac{2 \cdot \kappa_{\boldsymbol{\theta}^*} \cdot |R_{\boldsymbol{\theta}^*}|^{\max}}{(1-\gamma)^3} \cdot \sqrt{d_{\text{dyn}}\left(T^L, T^E\right)}
\end{aligned}
$$

E.3 Proof of Theorem 2

E.3.1 Impossibility to match the State-action Occupancy Measure

We overload the notation ρ^{π}_{M} to denote the state-action occupancy measure as well, which is defined as follows:

$$
\rho^{\pi}_{M}(s,a) \;:=\; \pi(a|s)\rho^{\pi}_{M}(s).
$$

Before proving the theorem, we show that finding the policy π^L whose state-action occupancy measure matches the state-action visitation frequency $\boldsymbol{\rho}$ of the expert policy[5] π^E is impossible in case of model mismatch. Consider:

$$
\begin{aligned}
\rho(s,a) \;&=\; \rho^{\pi^L}_{M^L}(s,a) \\
\rho(s)\pi^E(a|s) \;&=\; \rho^{\pi^L}_{M^L}(s)\pi^L(a|s) \\
\pi^L(a|s) \;&=\; \pi^E(a|s)\frac{\rho(s)}{\rho^{\pi^L}_{M^L}(s)}
\end{aligned}
$$

Notice that the policy π^L is normalized only if we require that $\frac{\rho(s)}{\rho^{\pi^L}_{M^L}(s)} = 1$. This implies that $\pi^L(s|a) = \pi^E(s|a)$. However, the same policy can not induce the same state occupancy measure under different transition dynamics, it follows that $\frac{\rho(s)}{\rho^{\pi^L}_{M^L}(s)} \neq 1$. We reached a contradiction that allows us to conclude that π^L can match the state-action occupancy measure only in absence of model mismatch. Therefore, when there is a model mismatch, the feasible set of (1) would be empty if state-action occupancy measures were used in posing the constraint. In addition, even if the two environments were the same, only the expert policy would have been in the feasible set because there exists an injective mapping from state-action visitation frequencies to policies as already noted in [37, 58].

E.3.2 Theorem Proof

Proof. If there exists a policy π^L that matches the expert state occupancy measure $\boldsymbol{\rho}$ in the environment M^L, the Bellman flow constraints [58] lead to the following equation for each state $s \in \mathcal{S}$:

$$
\rho(s) - (1-\gamma)P_0(s) \;=\; \gamma \sum_{s',a'} \rho(s')\pi^L(a'|s')T^L(s|s',a') \tag{37}
$$

This can be seen by writing the Bellman flow constraints for the expert policy π^E with transition dynamics T^E, and for the policy π^L with transition dynamics T^L:

$$
\rho(s) - (1-\gamma)P_0(s) \;=\; \gamma \sum_{s',a'} \rho(s')\pi^E(a'|s')T^E(s|s',a') \tag{38}
$$

$$
\rho^{\pi^L}_{M^L}(s) - (1-\gamma)P_0(s) \;=\; \gamma \sum_{s',a'} \rho^{\pi^L}_{M^L}(s')\pi^L(a'|s')T^L(s|s',a') \tag{39}
$$

By definition of π^L, the two occupancy measures are equal, so we can equate the LHS of (38) to the RHS of (39), obtaining:

$$
\rho(s) - (1-\gamma)P_0(s) \;=\; \gamma \sum_{s',a'} \rho^{\pi^L}_{M^L}(s')\pi^L(a'|s')T^L(s|s',a')
$$

[5] In this proof, the expert policy is denoted by π^E. In the specific case of our paper, it stands for either $\pi^*_{M^L_{\boldsymbol{\theta}}}$ or $\pi^*_{M^E_{\boldsymbol{\theta}}}$. However, the result holds for every valid expert policy.

Finally, replacing ρ in the RHS, one obtains the equation in (37). In addition, for each state we have the condition on the normalization of the policy:

$$1 \;=\; \sum_a \pi^L(a|s), \quad \forall s \in \mathcal{S}$$

All these conditions can be seen as an underdetermined system with $2\,|\mathcal{S}|$ equations ($|\mathcal{S}|$ for normalization, and $|\mathcal{S}|$ for the Bellman flow constraints). The unknown is the policy π^* represented by the $|S|\,|A|$ entries of the vector $\boldsymbol{\pi}^*$, formally defined in (43).

We introduce the matrix $\boldsymbol{T}$. In the first $|\mathcal{S}|$ rows, the entry in the sth row and $(s'\,|\mathcal{A}| + a')^{\text{th}}$ column is the element $\rho(s')T^L(s|s', a')$. In the last $|\mathcal{S}|$ rows, the entries are instead given by 1 from position $s'\,|\mathcal{A}|$ to position $s'\,|\mathcal{A}| + |\mathcal{A}|$. These rows of the matrix serves to impose the normalization condition for each possible state. A clearer block structure representation is given in Section 3.2.

We can thus write the underdetermined system as:

$$\begin{bmatrix} \boldsymbol{\rho} - (1-\gamma)\boldsymbol{P}_0 \\ \mathbf{1}_{|\mathcal{S}|} \end{bmatrix} \;=\; \boldsymbol{T}\boldsymbol{\pi}^L, \tag{40}$$

where the left hand side is a vector whose first $|\mathcal{S}|$ positions are the element-wise difference between the state occupancy measure and the initial probability distribution for each state, and the second half are all ones. Recognising that this matches the vector $\boldsymbol{v}$ described in Section 3.2, we can rewrite the system as:

$$\boldsymbol{v} \;=\; \boldsymbol{T}\boldsymbol{\pi}^L \tag{41}$$

The right hand side is instead written using the matrix $\boldsymbol{T}$, and the unknown matching policy vector $\boldsymbol{\pi}^L$. A direct application of the Rouché-Capelli theorem gives that a linear system admits solutions if and only if the rank of the coeffient matrix is equal to the rank of the coefficient matrix augmented with the known vector. In our case it is:

$$\text{rank}\,(\boldsymbol{T}) = \text{rank}\,(\boldsymbol{T}|\boldsymbol{v}) \tag{42}$$

This fact limits the class of perturbation in the dynamics that can be considered still achieving perfect matching. Corollary 1 follows because in the case of determined or underdetermined system, i.e. when $|\mathcal{A}| > 1$, the matrix $\boldsymbol{T}$ has rank no larger than $\min(2\,|\mathcal{S}|\,, |\mathcal{S}|\,|\mathcal{A}|) = 2\,|\mathcal{S}|$ that is the number of rows of the matrix. It follows that under this assumption, $\boldsymbol{T}$ is full rank when its rank is equal to $2\,|\mathcal{S}|$. The augmented matrix $(\boldsymbol{T}|\boldsymbol{v})$ will also have a rank upper bounded by $\min(2\,|\mathcal{S}|\,, |\mathcal{S}|\,|\mathcal{A}| + 1) = 2\,|\mathcal{S}|$ since it has constructed adding one column. This implies that, when $\boldsymbol{T}$ is full rank, equation (42) holds.

Block Representation of the Matching Policy Vector $\boldsymbol{\pi}^L$. For each state $s \in \mathcal{S}$, we can define a local matching policy vector $\boldsymbol{\pi}^L(s) \in \mathbb{R}^{|\mathcal{A}|}$ as:

$$\boldsymbol{\pi}^L(s) \;=\; \begin{bmatrix} \pi(a_1|s) \\ \pi(a_2|s) \\ \vdots \\ \pi(a_{|\mathcal{A}|}|s) \end{bmatrix}$$

Then, the matching policy vector $\boldsymbol{\pi}^L \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ is given by the vertical stacking of the local matching vectors:

$$\boldsymbol{\pi}^L \;=\; \begin{bmatrix} \boldsymbol{\pi^L}(s_1) \\ \boldsymbol{\pi^L}(s_2) \\ \vdots \\ \boldsymbol{\pi^L}(s_{|\mathcal{S}|}) \end{bmatrix} \tag{43}$$

□

E.4 Upper bound for the Reward Transfer Strategy

Let π^L be the policy obtained from the reward transfer strategy explained in Section 3.3, and $\pi_1 := \pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_L}}}$.

Theorem 6. *The performance gap between the policies π_1 and π^L on the MDP $M^L_{\boldsymbol{\theta}^*}$ is bounded as follows:*

$$\begin{aligned}
&\left|V^{\pi_1}_{M^L_{\boldsymbol{\theta}^*}} - V^{\pi^L}_{M^L_{\boldsymbol{\theta}^*}}\right| \\
\leq\ & \frac{|R_{\boldsymbol{\theta}^*}|^{\max}}{(1-\gamma)^2} \cdot \left\{\gamma \cdot d_{\text{dyn}}\left(T^L, T^E\right) + \frac{2 \cdot \kappa_{\boldsymbol{\theta}^{\text{train}}}}{1-\gamma} \cdot \sqrt{d_{\text{dyn}}\left(T^{\text{train}}, T^L\right)} + \gamma \cdot d_{\text{dyn}}\left(T^{\text{train}}, T^L\right) + d_{\text{pol}}\left(\pi_4, \pi^L\right)\right\}
\end{aligned}$$

Proof. We define $\pi_3 := \pi^{\text{soft}}_{M^{\text{train}}_{\boldsymbol{\theta}^{\text{train}}}}$ and $\pi_4 := \pi^{\text{soft}}_{M^L_{\boldsymbol{\theta}^{\text{train}}}}$. First, consider the following:

$$\begin{aligned}
\left|V^{\pi_3}_{M^{\text{train}}_{\boldsymbol{\theta}^*}} - V^{\pi_4}_{M^{\text{train}}_{\boldsymbol{\theta}^*}}\right| &= \frac{1}{1-\gamma} \cdot \left|\sum_s \left\{\rho^{\pi_3}_{M^{\text{train}}}(s) - \rho^{\pi_4}_{M^{\text{train}}}(s)\right\} R_{\boldsymbol{\theta}^*}(s)\right| \\
&\leq \frac{1}{1-\gamma} \cdot \sum_s \left|\rho^{\pi_3}_{M^{\text{train}}}(s) - \rho^{\pi_4}_{M^{\text{train}}}(s)\right| \cdot |R_{\boldsymbol{\theta}^*}(s)| \\
&\leq \frac{|R_{\boldsymbol{\theta}^*}|^{\max}}{1-\gamma} \cdot \sum_s \left|\rho^{\pi_3}_{M^{\text{train}}}(s) - \rho^{\pi_4}_{M^{\text{train}}}(s)\right| \\
&= \frac{|R_{\boldsymbol{\theta}^*}|^{\max}}{1-\gamma} \cdot \left\|\boldsymbol{\rho}^{\pi_3}_{M^{\text{train}}} - \boldsymbol{\rho}^{\pi_4}_{M^{\text{train}}}\right\|_1 \\
&\overset{\text{a}}{\leq} \frac{|R_{\boldsymbol{\theta}^*}|^{\max}}{(1-\gamma)^2} \cdot d_{\text{pol}}\left(\pi_3, \pi_4\right) \\
&\overset{\text{b}}{\leq} \frac{2 \cdot \kappa_{\boldsymbol{\theta}^{\text{train}}} \cdot |R_{\boldsymbol{\theta}^*}|^{\max}}{(1-\gamma)^3} \cdot \sqrt{d_{\text{dyn}}\left(T^{\text{train}}, T^L\right)},
\end{aligned}$$

where a is due to Lemma A.1 in [59], and b is due to Lemma 1. Then, consider the following:

$$\begin{aligned}
&\left|V^{\pi_1}_{M^L_{\boldsymbol{\theta}^*}} - V^{\pi^L}_{M^L_{\boldsymbol{\theta}^*}}\right| \\
\leq\ & \left|V^{\pi_1}_{M^L_{\boldsymbol{\theta}^*}} - V^{\pi^*_{M^L_{\boldsymbol{\theta}^*}}}_{M^L_{\boldsymbol{\theta}^*}}\right| + \left|V^{\pi^*_{M^L_{\boldsymbol{\theta}^*}}}_{M^L_{\boldsymbol{\theta}^*}} - V^{\pi^*_{M^E_{\boldsymbol{\theta}^*}}}_{M^E_{\boldsymbol{\theta}^*}}\right| + \left|V^{\pi^*_{M^E_{\boldsymbol{\theta}^*}}}_{M^E_{\boldsymbol{\theta}^*}} - V^{\pi_3}_{M^{\text{train}}_{\boldsymbol{\theta}^*}}\right| + \\
& \left|V^{\pi_3}_{M^{\text{train}}_{\boldsymbol{\theta}^*}} - V^{\pi_4}_{M^{\text{train}}_{\boldsymbol{\theta}^*}}\right| + \left|V^{\pi_4}_{M^{\text{train}}_{\boldsymbol{\theta}^*}} - V^{\pi_4}_{M^L_{\boldsymbol{\theta}^*}}\right| + \left|V^{\pi_4}_{M^L_{\boldsymbol{\theta}^*}} - V^{\pi^L}_{M^L_{\boldsymbol{\theta}^*}}\right| \\
\overset{\text{a}}{=}\ & \left|V^{\pi^*_{M^L_{\boldsymbol{\theta}^*}}}_{M^L_{\boldsymbol{\theta}^*}} - V^{\pi^*_{M^E_{\boldsymbol{\theta}^*}}}_{M^E_{\boldsymbol{\theta}^*}}\right| + \left|V^{\pi_3}_{M^{\text{train}}_{\boldsymbol{\theta}^*}} - V^{\pi_4}_{M^{\text{train}}_{\boldsymbol{\theta}^*}}\right| + \left|V^{\pi_4}_{M^{\text{train}}_{\boldsymbol{\theta}^*}} - V^{\pi_4}_{M^L_{\boldsymbol{\theta}^*}}\right| + \left|V^{\pi_4}_{M^L_{\boldsymbol{\theta}^*}} - V^{\pi^L}_{M^L_{\boldsymbol{\theta}^*}}\right| \\
\leq\ & \left|V^{\pi^*_{M^L_{\boldsymbol{\theta}^*}}}_{M^L_{\boldsymbol{\theta}^*}} - V^{\pi^*_{M^E_{\boldsymbol{\theta}^*}}}_{M^E_{\boldsymbol{\theta}^*}}\right| + \left|V^{\pi_3}_{M^{\text{train}}_{\boldsymbol{\theta}^*}} - V^{\pi_4}_{M^{\text{train}}_{\boldsymbol{\theta}^*}}\right| + \left|V^{\pi_4}_{M^{\text{train}}_{\boldsymbol{\theta}^*}} - V^{\pi_4}_{M^L_{\boldsymbol{\theta}^*}}\right| + \frac{|R_{\boldsymbol{\theta}^*}|^{\max}}{(1-\gamma)^2} \cdot d_{\text{pol}}\left(\pi_4, \pi^L\right) \\
\overset{\text{b}}{\leq}\ & \frac{\gamma \cdot |R_{\boldsymbol{\theta}^*}|^{\max}}{(1-\gamma)^2} \cdot d_{\text{dyn}}\left(T^L, T^E\right) + \left|V^{\pi_3}_{M^{\text{train}}_{\boldsymbol{\theta}^*}} - V^{\pi_4}_{M^{\text{train}}_{\boldsymbol{\theta}^*}}\right| + \left|V^{\pi_4}_{M^{\text{train}}_{\boldsymbol{\theta}^*}} - V^{\pi_4}_{M^L_{\boldsymbol{\theta}^*}}\right| + \frac{|R_{\boldsymbol{\theta}^*}|^{\max}}{(1-\gamma)^2} \cdot d_{\text{pol}}\left(\pi_4, \pi^L\right) \\
\overset{\text{c}}{\leq}\ & \frac{\gamma \cdot |R_{\boldsymbol{\theta}^*}|^{\max}}{(1-\gamma)^2} \cdot d_{\text{dyn}}\left(T^L, T^E\right) + \left|V^{\pi_3}_{M^{\text{train}}_{\boldsymbol{\theta}^*}} - V^{\pi_4}_{M^{\text{train}}_{\boldsymbol{\theta}^*}}\right| + \frac{\gamma \cdot |R_{\boldsymbol{\theta}^*}|^{\max}}{(1-\gamma)^2} \cdot d_{\text{dyn}}\left(T^{\text{train}}, T^L\right) + \frac{|R_{\boldsymbol{\theta}^*}|^{\max}}{(1-\gamma)^2} \cdot d_{\text{pol}}\left(\pi_4, \pi^L\right) \\
\leq\ & \frac{|R_{\boldsymbol{\theta}^*}|^{\max}}{(1-\gamma)^2} \cdot \left\{\gamma \cdot d_{\text{dyn}}\left(T^L, T^E\right) + \frac{2 \cdot \kappa_{\boldsymbol{\theta}^{\text{train}}}}{1-\gamma} \cdot \sqrt{d_{\text{dyn}}\left(T^{\text{train}}, T^L\right)} + \gamma \cdot d_{\text{dyn}}\left(T^{\text{train}}, T^L\right) + d_{\text{pol}}\left(\pi_4, \pi^L\right)\right\},
\end{aligned}$$

where a is due to the fact that $\boldsymbol{\rho}^{\pi_1}_{M^L} = \boldsymbol{\rho}^{\pi^*_{M^L_{\boldsymbol{\theta}^*}}}_{M^L}$ and $\boldsymbol{\rho}^{\pi^*_{M^E_{\boldsymbol{\theta}^*}}}_{M^E} = \boldsymbol{\rho}^{\pi_3}_{M^{\text{train}}}$; b is due to Theorem 7 in [34]; and c is due to Simulation Lemma [35, 36]. □

When $M^{\text{train}} = M^E$ and $\pi^L = \pi_4$, the above bound simplifies to:

$$\left| V^{\pi_1}_{M^L_{\boldsymbol{\theta}^*}} - V^{\pi^L}_{M^L_{\boldsymbol{\theta}^*}} \right| \;\leq\; \frac{2 \cdot |R_{\boldsymbol{\theta}^*}|^{\max}}{(1-\gamma)^2} \cdot \left\{ \gamma \cdot d_{\text{dyn}}\left(T^L, T^E\right) + \frac{\kappa_{\boldsymbol{\theta}^{\text{E}}}}{1-\gamma} \cdot \sqrt{d_{\text{dyn}}\left(T^L, T^E\right)} \right\}.$$

F Further Details of Section 4

F.1 Relation between Robust MDP and Markov Games

This section gives a proof for the inequality in equation (8):

Proof. We first introduce the set:

$$\underline{\mathcal{T}}^{L,\alpha} = \left\{ T \quad | \quad T(s'|s,a) = \alpha T^L(s'|s,a) + (1-\alpha)\bar{T}(s'|s), \quad \bar{T}(s'|s) = \sum_a \pi(a|s) T(s'|s,a), \forall \pi \in \Delta_{A|S} \right\}$$

Clearly, it holds that: $\underline{\mathcal{T}}^{L,\alpha} \subset \mathcal{T}^{L,\alpha}$ that implies:

$$\max_{\pi^{\mathrm{pl}} \in \Pi} \min_{T \in \mathcal{T}^{L,\alpha}} \mathbb{E}\left[G \mid \pi^{\mathrm{pl}}, P_0, T\right] \leq \max_{\pi^{\mathrm{pl}} \in \Pi} \min_{T \in \underline{\mathcal{T}}^{L,\alpha}} \mathbb{E}\left[G \mid \pi^{\mathrm{pl}}, P_0, T\right]$$

Finally, from [27, Section 3.1] we have:

$$\max_{\pi^{\mathrm{pl}} \in \Pi} \min_{T \in \underline{\mathcal{T}}^{L,\alpha}} \mathbb{E}\left[G \mid \pi^{\mathrm{pl}}, P_0, T\right] = \max_{\pi^{\mathrm{pl}} \in \Pi} \min_{\pi^{\mathrm{op}} \in \Pi} \mathbb{E}\left[G \mid \alpha\pi^{\mathrm{pl}} + (1-\alpha)\pi^{\mathrm{op}}, M^L\right]$$

We conclude that:

$$\max_{\pi^{\mathrm{pl}} \in \Pi} \min_{T \in \mathcal{T}^{L,\alpha}} \mathbb{E}\left[G \mid \pi^{\mathrm{pl}}, P_0, T\right] \leq \max_{\pi^{\mathrm{pl}} \in \Pi} \min_{\pi^{\mathrm{op}} \in \Pi} \mathbb{E}\left[G \mid \alpha\pi^{\mathrm{pl}} + (1-\alpha)\pi^{\mathrm{op}}, M^L\right]$$

Therefore the inequality in (8) holds. □

A natural question is whether the tightness of the bound can be controlled. An affirmative answer come from the following theorem relying on Lemma 1.

Theorem 7. *Let T^* be a saddle point when the* min *acts over the set $\mathcal{T}^{L,\alpha}$ and $\underline{T}^*$ be a saddle point when the* min *acts over the set $\underline{\mathcal{T}}^{L,\alpha}$. Then, the following holds:*

$$\max_{\pi^{\mathrm{pl}} \in \Pi} \min_{\pi^{\mathrm{op}} \in \Pi} \mathbb{E}\left[G \mid \alpha\pi^{\mathrm{pl}} + (1-\alpha)\pi^{\mathrm{op}}, M^L\right] - \max_{\pi^{\mathrm{pl}} \in \Pi} \min_{T \in \mathcal{T}^{L,\alpha}} \mathbb{E}\left[G \mid \pi^{\mathrm{pl}}, P_0, T\right]$$
$$\leq \frac{2|R_\theta^{\max}|}{(1-\gamma)^2} \min\left\{ \frac{\kappa_\theta \sqrt{d(T^*, \underline{T}^*)}}{(1-\gamma)}, \frac{\kappa_\theta^2 d(T^*, \underline{T}^*)}{(1-\gamma)^2} \right\}$$

Proof.

$$\begin{aligned}
&\max_{\pi^{\mathrm{pl}} \in \Pi} \min_{\pi^{\mathrm{op}} \in \Pi} \mathbb{E}\left[G \mid \alpha\pi^{\mathrm{pl}} + (1-\alpha)\pi^{\mathrm{op}}, M^L\right] - \max_{\pi^{\mathrm{pl}} \in \Pi} \min_{T \in \mathcal{T}^{L,\alpha}} \mathbb{E}\left[G \mid \pi^{\mathrm{pl}}, P_0, T\right] \\
&= \max_{\pi^{\mathrm{pl}} \in \Pi} \min_{T \in \underline{\mathcal{T}}^{L,\alpha}} \mathbb{E}\left[G \mid \pi^{\mathrm{pl}}, P_0, T\right] - \max_{\pi^{\mathrm{pl}} \in \Pi} \min_{T \in \mathcal{T}^{L,\alpha}} \mathbb{E}\left[G \mid \pi^{\mathrm{pl}}, P_0, T\right] \\
&= \max_{\pi^{\mathrm{pl}} \in \Pi} \mathbb{E}\left[G \mid \pi^{\mathrm{pl}}, P_0, \underline{T}^*\right] - \max_{\pi^{\mathrm{pl}} \in \Pi} \mathbb{E}\left[G \mid \pi^{\mathrm{pl}}, P_0, T^*\right] \\
&\leq \frac{|R_\theta|^{\max}}{(1-\gamma)^2} d_{\mathrm{pol}}\left(\pi^{\mathrm{soft}}_{\underline{T}^*}, \pi^{\mathrm{soft}}_{T^*}\right) \leq \frac{2|R_\theta^{\max}|}{(1-\gamma)^2} \min\left\{ \frac{\kappa_\theta \sqrt{d(T^*, \underline{T}^*)}}{(1-\gamma)}, \frac{\kappa_\theta^2 d(T^*, \underline{T}^*)}{(1-\gamma)^2} \right\}
\end{aligned}$$

Where the second last inequality holds with similar steps of the proof of Theorem 6 and the last inequality applies thanks to Lemma 1. □

F.2 Deriving Gradient-based Method from Worst-case Predictive Log-loss

We consider again in this section the optimization problem given in (1) with model mismatch, i.e., using $\boldsymbol{\rho}_{M^E_{\boldsymbol{\theta}^*}}^{\pi^*_{M^E}}$ as $\boldsymbol{\rho}$. The aim of this section is to give an alternative point of view on this program based on a proper adaptation of the worst-case predictive log-loss [24][Corollary 6.3] to the model mismatch case.

[24] proved that the maximum causal entropy policy satisfying the optimization constraints is also the distribution that minimizes the worst-case predictive log-loss. However, the proof leverages on the fact that learner and expert MDPs coincide, an assumption that fails in the scenario of our work.

This section extends the result to the general case, where expert and learner MDP do not coincide, thanks to the two following contributions: (i) we show that the MCE constrained maximization given in (4) in the main text can be recast as a worst-case predictive log-loss constrained minimization and (ii) that this alternative problem leads to the same reward weights update found in the main text for the dual of the program (4). We start reporting again the optimization problem of interest:

$$\arg\max_{\pi\in\Pi} \quad \mathbb{E}\left[\sum_{t=0}^{\infty} -\gamma^t \log \pi(a_t|s_t) \;\middle|\; \pi, M^L\right] \tag{44}$$

$$\text{subject to} \quad \boldsymbol{\rho}_{M^E}^{\pi^*_{M^E_{\boldsymbol{\theta}^*}}} \;=\; \boldsymbol{\rho}_{M^L}^{\pi} \tag{45}$$

An alternative interpretation of the entropy is given by the following property:

$$\mathbb{E}\left[\sum_{t=0}^{\infty} -\gamma^t \log \pi(a_t|s_t) \;\middle|\; \pi, M^L\right] \;=\; \inf_{\bar{\pi}} \mathbb{E}\left[\sum_{t=0}^{\infty} -\gamma^t \log \bar{\pi}(a_t|s_t) \;\middle|\; \pi, M^L\right], \quad \forall \pi$$

Thus, it holds also for $\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}$ solution of the primal optimization problem (44)- (45), that exists if Theorem 2 is satisfied. In addition, to maintain the equivalence with the program (44)-(45), we restrict the $\inf$ search space to the feasible set of (44)-(45) that we denote $\widetilde{\Pi}$.

$$\mathbb{E}\left[\sum_{t=0}^{\infty} -\gamma^t \log \pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}(a_t|s_t) \;\middle|\; \pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}, M^L\right] \;=\; \inf_{\bar{\pi}\in\widetilde{\Pi}} \mathbb{E}\left[\sum_{t=0}^{\infty} -\gamma^t \log \bar{\pi}(a_t|s_t) \;\middle|\; \pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}, M^L\right]$$

Notice that since $\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}$ is solution of the maximization problem, we can indicate the the previous equality as:

$$\begin{aligned}\sup_{\tilde{\pi}\in\widetilde{\Pi}} \mathbb{E}\left[\sum_{t=0}^{\infty} -\gamma^t \log \tilde{\pi}(a_t|s_t) \;\middle|\; \tilde{\pi}, M^L\right] \;&=\; \sup_{\tilde{\pi}\in\widetilde{\Pi}} \inf_{\bar{\pi}\in\widetilde{\Pi}} \mathbb{E}\left[\sum_{t=0}^{\infty} -\gamma^t \log \bar{\pi}(a_t|s_t) \;\middle|\; \tilde{\pi}, M^L\right] \\ &=\; \inf_{\bar{\pi}\in\widetilde{\Pi}} \sup_{\tilde{\pi}\in\widetilde{\Pi}} \mathbb{E}\left[\sum_{t=0}^{\infty} -\gamma^t \log \bar{\pi}(a_t|s_t) \;\middle|\; \tilde{\pi}, M^L\right]\end{aligned} \tag{46}$$

The last equality follows by min-max equality that holds since the objective is convex in $\bar{\pi}$ and concave in $\tilde{\pi}$. It is thus natural to interpret the quantity:

$$c(\pi) = \mathbb{E}\left[\sum_{t=0}^{\infty} -\gamma^t \log \pi(a_t|s_t) \;\middle|\; \pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}, M^L\right] \tag{47}$$

as the cost function associated to the policy π because, according to (46), this quantity is equivalent to the worst-case predictive log-loss among the policies of the feasible set $\widetilde{\Pi}$. It can be seen that the loss inherits the feasible set of the original MCE maximization problem as search space for the $\inf$ and $\sup$ operations. It follows that in case of model mismatch, the loss studied in [24][Corollary 6.3] is modified because a different set must be used as search space for the $\inf$ and $\sup$.

In the following, we develop a gradient based method to minimize this cost and, thus, the worst case predictive log-loss. [6]

Furthermore, we can already consider that π belongs to the family of soft Bellman policies parametrized by the parameter $\boldsymbol{\theta}$ in the environment $M^L_{\boldsymbol{\theta}}$ because they are the family of distributions attaining maximum discounted causal entropy (see [37][Lemma 3]). The cost is, in this case, expressed for the parameter $\boldsymbol{\theta}$:

$$c(\boldsymbol{\theta}) = \mathbb{E}\left[\sum_{t=0}^{\infty} -\gamma^t \log \pi^{\text{soft}}_{M^L_{\boldsymbol{\theta}}}(a_t|s_t) \;\middle|\; \pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}, M^L\right] \tag{48}$$

[6] If we used $\boldsymbol{\rho}_{M^L}^{\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta}^*}}}$ as $\boldsymbol{\rho}$, we would have obtained the cost $c(\pi) = \mathbb{E}\left[\sum_{t=0}^{\infty} -\gamma^t \log \pi(a_t|s_t) \;\middle|\; \pi^{\text{soft}}_{M^L_{\boldsymbol{\theta}^*}}, M^L\right]$. In this case, the gradient is known see [60].

Theorem 8. *If $\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}$ exists, the gradient of the cost function given in* (48) *is equal to:*

$$\nabla_{\boldsymbol{\theta}} c(\boldsymbol{\theta}) = \sum_s \left(\rho^{\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta}}}}_{M_L}(s) - \rho^{\pi^*_{M^E_{\boldsymbol{\theta^*}}}}_{M_E}(s) \right) \nabla_{\boldsymbol{\theta}} R_{\boldsymbol{\theta}}(s)$$

In addition, this result generalizes when the expectation in the cost function is taken with respect to any of the policies in the feasible set of the primal problem (44)-(45)*.*

Note that choosing one-hot features, we have $\nabla_{\boldsymbol{\theta}} c(\boldsymbol{\theta}) = \boldsymbol{\rho}^{\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta}}}}_{M_L} - \boldsymbol{\rho}^{\pi^*_{M^E_{\boldsymbol{\theta^*}}}}_{M_E}$ as used in Section 4.

Uniqueness of the Solution. The cost in equation (48) is strictly convex in the soft max policy $\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta}}}$ because $-\log(\cdot)$ is a strictly convex function and the cost consists in a linear composition of these strictly convex functions. Thus the gradient descent converges to a unique soft optimal policy. In addition, the fact that for each possible $\boldsymbol{\theta}$, the quantity $\log \pi^{\text{soft}}_{M^L_{\boldsymbol{\theta}}} = Q^{\text{soft}}_{M_{\boldsymbol{\theta}}}(s,a) - V^{\text{soft}}_{M_{\boldsymbol{\theta}}}(s)$ is convex in $\boldsymbol{\theta}$ since the soft value functions ($Q^{\text{soft}}_{M_{\boldsymbol{\theta}}}(s,a)$ and $V^{\text{soft}}_{M_{\boldsymbol{\theta}}}(s)$) are given by a sum of rewards that are linear in $\boldsymbol{\theta}$ and LogSumExp funtions that are convex. It follows that $\log \pi^{\text{soft}}_{M^L_{\boldsymbol{\theta}}}$ is a composition of linear and convex functions for each state actions pairs. Consequently the cost given in (48) is convex in $\boldsymbol{\theta}$. It follows that alternating an update of the parameter $\boldsymbol{\theta}$ using a gradient descent scheme based on the gradient given by Theorem 8 with a derivation of the corresponding soft-optimal policy by Soft-Value-Iteration, one can converge to $\boldsymbol{\theta_E}$ whose corresponding soft optimal policy is $\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}$. However, considering that the function LogSumExp is convex but not strictly convex there is no unique $\boldsymbol{\theta_E}$ corresponding to the soft optimal policy $\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}$.

F.2.1 Proof of Theorem 8

Proof. We will make use of the following quantities:

- $P_t^{\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta}}}}(s)$ defined as the probability of visiting state s at time t by the policy $\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta}}}$ acting in $M^L_{\boldsymbol{\theta}}$
- $P_t^{\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta}}}}(s,a)$ defined as the probability of visiting state s and taking action a from state s at time t by the policy $\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta}}}$ acting in $M^L_{\boldsymbol{\theta}}$
- $P_t^{\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}}(s)$ defined as the probability of visiting state s at time t by the policy $\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}$ acting in $M^L_{\boldsymbol{\theta}}$
- $P_t^{\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}}(s,a)$ defined as the probability of visiting state s and taking action a from state s at time t by the policy $\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}$ acting in $M^L_{\boldsymbol{\theta}}$

The cost can be rewritten as:

$$\begin{aligned}
c(\boldsymbol{\theta}) &= -\sum_{t=0}^{\infty} \gamma^t \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} P_t^{\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}}(s,a) \log \pi^{\text{soft}}_{M^L_{\boldsymbol{\theta}}}(a|s) \\
&= -\sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} P_0^{\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}}(s,a) \left(Q^{\text{soft}}_{M^L_{\boldsymbol{\theta}}}(s,a) - V^{\text{soft}}_{M^L_{\boldsymbol{\theta}}}(s) \right) \\
&\quad -\sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} P_1^{\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}}(s,a)\gamma \left(Q^{\text{soft}}_{M^L_{\boldsymbol{\theta}}}(s,a) - V^{\text{soft}}_{M^L_{\boldsymbol{\theta}}}(s) \right) \\
&\quad -\sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} P_2^{\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}}(s,a)\gamma^2 \left(Q^{\text{soft}}_{M^L_{\boldsymbol{\theta}}}(s,a) - V^{\text{soft}}_{M^L_{\boldsymbol{\theta}}}(s) \right)
\end{aligned}$$

$$
\begin{aligned}
& - \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} P_3^{\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}}(s,a)\gamma^3\left(Q^{\text{soft}}_{M^L_{\boldsymbol{\theta}}}(s,a) - V^{\text{soft}}_{M^L_{\boldsymbol{\theta}}}(s)\right) \\
& \dots \\
= & \sum_{s,a} P_0(s)\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}(a|s)V^{\text{soft}}_{M^L_{\boldsymbol{\theta}}}(s) && (49) \\
& - \sum_{s,a} P_0(s)\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}(a|s)Q^{\text{soft}}_{M^L_{\boldsymbol{\theta}}}(s,a) + \gamma\sum_{s,a} P_1^{\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}}(s)\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}(a|s)V^{\text{soft}}_{M^L_{\boldsymbol{\theta}}}(s) && (50) \\
& - \gamma\sum_{s,a} P_1^{\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}}(s)\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}(a|s)Q^{\text{soft}}_{M^L_{\boldsymbol{\theta}}}(s,a) + \gamma^2\sum_{s,a} P_2^{\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}}(s)\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}(a|s)V^{\text{soft}}_{M^L_{\boldsymbol{\theta}}}(s) && (51) \\
& - \gamma^2\sum_{s,a} P_2^{\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}}(s)\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}(a|s)Q^{\text{soft}}_{M^L_{\boldsymbol{\theta}}}(s,a) + \gamma^3\sum_{s,a} P_3^{\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}}(s)\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}(a|s)V^{\text{soft}}_{M^L_{\boldsymbol{\theta}}}(s) \\
& \dots
\end{aligned}
$$

The gradient of the term in (49) has already been derived in [60] and it is given by:

$$
\nabla_{\boldsymbol{\theta}}\sum_{s,a} P_0(s)\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}(a|s)V^{\text{soft}}_{M^L_{\boldsymbol{\theta}}}(s) \;=\; \nabla_{\boldsymbol{\theta}}\sum_{s} P_0(s)V^{\text{soft}}_{M^L_{\boldsymbol{\theta}}}(s) \;=\; \sum_{s,a}\rho^{\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta}}}}_{M_L}(s,a)\nabla_{\boldsymbol{\theta}}R_{\boldsymbol{\theta}}(s,a)
$$

Now, we compute the gradient of the following terms starting from (50). We notice that this term can be simplified as follows:

$$
\begin{aligned}
& - \sum_{s,a} P_0(s)\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}(a|s)Q^{\text{soft}}_{M^L_{\boldsymbol{\theta}}}(s,a) + \gamma\sum_{s,a} P_1^{\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}}(s)\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}(a|s)V^{\text{soft}}_{M^L_{\boldsymbol{\theta}}}(s) \\
= & - \sum_{s,a} P_0(s)\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}(a|s)\left(R_{\boldsymbol{\theta}}(s,a) + \gamma\sum_{s'} T^L(s'|s,a)V^{\text{soft}}_{M^L_{\boldsymbol{\theta}}}(s')\right) + \gamma\sum_{s,a} P_1^{\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}}(s)\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}(a|s)V^{\text{soft}}_{M^L_{\boldsymbol{\theta}}}(s) \\
= & - \sum_{s,a} P_0(s)\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}(a|s)R_{\boldsymbol{\theta}}(s,a) - \gamma\sum_{s'}\sum_{s,a} T^L(s'|s,a)P_0(s)\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}(a|s)V^{\text{soft}}_{M^L_{\boldsymbol{\theta}}}(s') + \gamma\sum_{s} P_1^{\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}}(s)V^{\text{soft}}_{M^L_{\boldsymbol{\theta}}}(s) \\
= & - \sum_{s,a} P_0(s)\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}(a|s)R_{\boldsymbol{\theta}}(s,a) - \gamma\sum_{s'} P_1^{\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}}(s')V^{\text{soft}}_{M^L_{\boldsymbol{\theta}}}(s') + \gamma\sum_{s} P_1^{\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}}(s)V^{\text{soft}}_{M^L_{\boldsymbol{\theta}}}(s) \\
= & - \sum_{s,a} P_0(s)\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}(a|s)R_{\boldsymbol{\theta}}(s,a)
\end{aligned}
$$

With similar steps, all the terms except the first one are given by

$$
-\sum_{t=0}^{\infty}\sum_{s,a} P_t^{\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}}(s,a)\gamma^t R_{\boldsymbol{\theta}}(s,a) \;=\; -\sum_{s,a}\rho^{\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}}_{M^L}(s,a)R_{\boldsymbol{\theta}}(s,a)
$$

If the reward is state only, then and we can marginalize the sum over the action and then exploiting the fact that $\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}$ is in the feasible set of the primal problem (44)-(45):

$$
-\sum_{t=0}^{\infty}\sum_{s,a} P_t^{\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}}(s,a)\gamma^t R_{\boldsymbol{\theta}}(s) \;=\; -\sum_{s}\rho^{\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta_E}}}}_{M^L}(s)R_{\boldsymbol{\theta}}(s) \;=\; -\sum_{s}\rho^{\pi^{*}_{M^E_{\boldsymbol{\theta^*}}}}_{M^E}(s)R_{\boldsymbol{\theta}}(s)
$$

It follows that the gradient of all the terms but the first term (49) is given by:

$$
-\sum_{s}\rho^{\pi^{*}_{M^E_{\boldsymbol{\theta^*}}}}_{M^E}(s)R_{\boldsymbol{\theta}}(s)
$$

Finally, the proof is concluded by summing the latest result to the gradient of (49) that gives:

$$\nabla_{\boldsymbol{\theta}} c(\boldsymbol{\theta}) \;=\; \sum_{s} \left(\rho_{M^L}^{\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta}}}}(s) - \rho_{M^E}^{\pi^{*}_{M^E_{\boldsymbol{\theta}^*}}}(s) \right) \nabla_{\boldsymbol{\theta}} R_{\boldsymbol{\theta}}(s)$$

It can be noticed that the computation of this gradient exploits only the fact that $\pi^{\text{soft}}_{M^L_{\boldsymbol{\theta}_{\boldsymbol{E}}}}$ is in the primal feasible set and not the fact that it maximizes the discounted causal entropy. It follows that all the policies in the primal feasible set share this gradient. This means that this gradients aim to move the learner policy towards the primal feasible set while the causal entropy is then maximized by Soft-Value-Iteration. □

F.3 Solving the Two-Player Markov Game

Algorithm 2 Value Iteration for Two-Player Markov Game

Initialize: $Q(s, a^{\text{pl}}, a^{\text{op}}) \leftarrow 0$, $V(s) \leftarrow 0$
while not converged **do**
  for $s \in \mathcal{S}$ **do**
    for $(a^{\text{pl}}, a^{\text{op}}) \in \mathcal{A} \times \mathcal{A}$ **do**
      update joint Q-function as follows:

$$Q(s, a^{\text{pl}}, a^{\text{op}}) \;=\; R(s) + \gamma \sum_{s'} T^{\text{two},L,\alpha}(s'|s, a^{\text{pl}}, a^{\text{op}}) V(s') \tag{52}$$

    end for
    update joint V-function as follows:

$$V(s) \;=\; \log \sum_{a^{\text{pl}}} \exp \left(\min_{a^{\text{op}}} Q(s, a^{\text{pl}}, a^{\text{op}}) \right) \tag{53}$$

  end for
end while
compute the marginal Q values for player and opponent, for all $(s, a^{\text{pl}}, a^{\text{op}}) \in \mathcal{S} \times \mathcal{A} \times \mathcal{A}$:

$$Q^{\text{pl}}(s, a^{\text{pl}}) \;=\; \min_{a^{\text{op}}} Q(s, a^{\text{pl}}, a^{\text{op}}) \quad \text{and}$$

$$Q^{\text{op}}(s, a^{\text{op}}) \;=\; \log \sum_{a^{\text{pl}}} \exp Q(s, a^{\text{pl}}, a^{\text{op}})$$

compute the player (soft-max) and opponent (greedy) policies, for all $(s, a^{\text{pl}}, a^{\text{op}}) \in \mathcal{S} \times \mathcal{A} \times \mathcal{A}$:

$$\pi^{\text{pl}}(a^{\text{pl}}|s) \;=\; \frac{\exp Q^{\text{pl}}(s, a^{\text{pl}})}{\sum_{a'} Q^{\text{pl}}(s, a')} \quad \text{and}$$

$$\pi^{\text{op}}(a^{\text{op}}|s) \;=\; \mathbb{1}\left[a^{\text{op}} \in \arg\min_{a'} Q^{\text{op}}(s, a') \right]$$

Output: player policy π^{pl}, opponent policy π^{op}

Here, we prove that the optimization problem in (9) can be solved by the Algorithm 2. First of all, one can rewrite (9) as:

$$\mathop{\mathbb{E}}_{s \sim P_0} \left[\mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t \left\{ R_{\boldsymbol{\theta}}(s_t) + H^{\pi^{\text{pl}}}(A \mid S = s_t) \right\} \;\middle|\; \pi^{\text{pl}}, \pi^{\text{op}}, M^{\text{two},L,\alpha}, s_0 = s \right] \right]$$

The quantity inside the expectation over P_0 is usually known as free energy, and for each state $s \in \mathcal{S}$, it is equal to:

$$F(\pi^{\text{pl}}, \pi^{\text{op}}, s) \;=\; \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t \left\{ R_{\boldsymbol{\theta}}(s_t) + H^{\pi^{\text{pl}}}(A \mid S = s_t) \right\} \;\middle|\; \pi^{\text{pl}}, \pi^{\text{op}}, M^{\text{two},L,\alpha}, s_0 = s \right]$$

Separating the first term of the sum over temporal steps, one can observe a recursive relation that is useful for the development of the algorithm:

$F(\pi^{\text{pl}}, \pi^{\text{op}}, s)$

$$
\begin{aligned}
&= R_{\boldsymbol{\theta}}(s) + H^{\pi^{\mathrm{pl}}}(A|S=s) \\
&+ \mathop{\mathbb{E}}_{a^{\mathrm{pl}}\sim\pi^{\mathrm{pl}},a^{\mathrm{op}}\sim\pi^{\mathrm{op}}}\left[\mathop{\mathbb{E}}_{s'\sim T^{\mathrm{two},L,\alpha}(\cdot|s,a^{\mathrm{pl}},a^{\mathrm{op}})}\left[\mathbb{E}\left[\sum_{t=1}^{\infty}\gamma^t\left\{R_{\boldsymbol{\theta}}(s_t)+H^{\pi^{\mathrm{pl}}}\left(A\mid S=s_t\right)\right\}\,\middle|\,\pi^{\mathrm{pl}},\pi^{\mathrm{op}},M^{\mathrm{two},L,\alpha},s_1=s'\right]\right]\right] \\
&= R_{\boldsymbol{\theta}}(s) + H^{\pi^{\mathrm{pl}}}(A|S=s) \\
&+ \gamma\mathop{\mathbb{E}}_{a^{\mathrm{pl}}\sim\pi^{\mathrm{pl}},a^{\mathrm{op}}\sim\pi^{\mathrm{op}}}\left[\mathop{\mathbb{E}}_{s'\sim T^{\mathrm{two},L,\alpha}(\cdot|s,a^{\mathrm{pl}},a^{\mathrm{op}})}\left[\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t\left\{R_{\boldsymbol{\theta}}(s_t)+H^{\pi^{\mathrm{pl}}}\left(A\mid S=s_t\right)\right\}\,\middle|\,\pi^{\mathrm{pl}},\pi^{\mathrm{op}},M^{\mathrm{two},L,\alpha},s_0=s'\right]\right]\right] \\
&= R_{\boldsymbol{\theta}}(s) + H^{\pi^{\mathrm{pl}}}(A|S=s) + \gamma\mathop{\mathbb{E}}_{a^{\mathrm{pl}}\sim\pi^{\mathrm{pl}},a^{\mathrm{op}}\sim\pi^{\mathrm{op}}}\left[\mathop{\mathbb{E}}_{s'\sim T^{\mathrm{two},L,\alpha}(\cdot|s,a^{\mathrm{pl}},a^{\mathrm{op}})}\left[F(\pi^{\mathrm{pl}},\pi^{\mathrm{op}},s')\right]\right] \\
&= \mathop{\mathbb{E}}_{a^{\mathrm{pl}}\sim\pi^{\mathrm{pl}},a^{\mathrm{op}}\sim\pi^{\mathrm{op}}}\left[R_{\boldsymbol{\theta}}(s) - \log\pi^{\mathrm{pl}}(a^{\mathrm{pl}}|s) + \gamma\mathop{\mathbb{E}}_{s'\sim T^{\mathrm{two},L,\alpha}(\cdot|s,a^{\mathrm{pl}},a^{\mathrm{op}})}\left[F(\pi^{\mathrm{pl}},\pi^{\mathrm{op}},s')\right]\right]
\end{aligned}
$$

Then, our aim is to find the saddle point:

$$
V(s) \;=\; \max_{\pi^{\mathrm{pl}}}\min_{\pi^{\mathrm{op}}} F(\pi^{\mathrm{pl}},\pi^{\mathrm{op}},s)
$$

and the policies attaining it. Define the joint quality function for a triplet $(s,a^{\mathrm{pl}},a^{\mathrm{op}})$ as:

$$
Q(s,a^{\mathrm{pl}},a^{\mathrm{op}}) \;=\; R_{\boldsymbol{\theta}}(s) + \gamma\mathop{\mathbb{E}}_{s'\sim T(\cdot|s,a^{\mathrm{pl}},a^{\mathrm{op}})}\left[V(s')\right]
$$

In a dynamic programming context, the previous equation gives the quality function based on the observed reward and the current estimate of the saddle point V. This is done by step (52) in the Algorithm 2. It remains now to motivate the update of the saddle point estimate V in (53). Consider:

$$
\begin{aligned}
&\max_{\pi^{\mathrm{pl}}}\min_{\pi^{\mathrm{op}}} F(\pi^{\mathrm{pl}},\pi^{\mathrm{op}},s) \\
&= \max_{\pi^{\mathrm{pl}}}\min_{\pi^{\mathrm{op}}}\mathop{\mathbb{E}}_{a^{\mathrm{pl}}\sim\pi^{\mathrm{pl}}(\cdot|s),a^{\mathrm{op}}\sim\pi^{\mathrm{op}}(\cdot|s)}\left[Q(s,a^{\mathrm{pl}},a^{\mathrm{op}}) - \log\pi^{\mathrm{pl}}(a^{\mathrm{pl}}|s)\right] \\
&= \max_{\pi^{\mathrm{pl}}}\min_{\pi^{\mathrm{op}}}\mathop{\mathbb{E}}_{a^{\mathrm{pl}}\sim\pi^{\mathrm{pl}}(\cdot|s)}\left[\mathop{\mathbb{E}}_{a^{\mathrm{op}}\sim\pi^{\mathrm{op}}(\cdot|s)}\left[Q(s,a^{\mathrm{pl}},a^{\mathrm{op}}) - \log\pi^{\mathrm{pl}}(a^{\mathrm{pl}}|s)|a^{\mathrm{pl}}\right]\right] \\
&= \max_{\pi^{\mathrm{pl}}}\mathop{\mathbb{E}}_{a^{\mathrm{pl}}\sim\pi^{\mathrm{pl}}(\cdot|s)}\left[\min_{\pi^{\mathrm{op}}}\mathop{\mathbb{E}}_{a^{\mathrm{op}}\sim\pi^{\mathrm{op}}(\cdot|s)}\left[Q(s,a^{\mathrm{pl}},a^{\mathrm{op}}) - \log\pi^{\mathrm{pl}}(a^{\mathrm{pl}}|s)|a^{\mathrm{pl}}\right]\right] \\
&= \max_{\pi^{\mathrm{pl}}}\mathop{\mathbb{E}}_{a^{\mathrm{pl}}\sim\pi^{\mathrm{pl}}(\cdot|s)}\left[\underbrace{\min_{a^{\mathrm{op}}} Q(s,a^{\mathrm{pl}},a^{\mathrm{op}})}_{Q^{\mathrm{pl}}(s,a^{\mathrm{pl}})} - \log\pi^{\mathrm{pl}}(a^{\mathrm{pl}}|s)\right] \\
&= \log\sum_{a^{\mathrm{pl}}}\exp Q^{\mathrm{pl}}(s,a^{\mathrm{pl}}),
\end{aligned}
$$

where the second last equality follows choosing a greedy policy π^{op} that selects the opponent action that minimizes the joint quality function $Q(s,a^{\mathrm{pl}},a^{\mathrm{op}})$.

The last equality is more involved and it is explained in the following lines:

$$
\mathop{\mathbb{E}}_{a^{\mathrm{pl}}\sim\pi^{\mathrm{pl}}(\cdot|s)}\left[Q^{\mathrm{pl}}(s,a^{\mathrm{pl}}) - \log\pi^{\mathrm{pl}}(a^{\mathrm{pl}}|s)\right] \;=\; \sum_{a^{\mathrm{pl}}}\pi^{\mathrm{pl}}(a^{\mathrm{pl}}|s)\left(Q^{\mathrm{pl}}(s,a^{\mathrm{pl}}) - \log\pi^{\mathrm{pl}}(a^{\mathrm{pl}}|s)\right)
$$

The latter expression is a strictly concave with respect to each decision variable $\pi(a|s)$. So if the derivative with respect to each decision variable $\pi^{\mathrm{pl}}(a^{\mathrm{pl}}|s)$ is zero, we have found the desired global maximum. The normalization is imposed once the maximum has been found. Taking the derivative for a particular decision variable, and equating to zero, we have:

$$
\left(Q^{\mathrm{pl}}(s,a^{\mathrm{pl}}) - \log\pi^{\mathrm{pl}}(a^{\mathrm{pl}}|s)\right) - 1 \;=\; 0
$$

It follows that:

$$
\pi^{\mathrm{pl}}(a|s) \propto \exp Q^{\mathrm{pl}}(s,a^{\mathrm{pl}})
$$

and imposing the proper normalization, we obtain the maximizing policy $\pi^{\mathrm{pl},*}$ with the form:

$$\pi^{\mathrm{pl},*}(a^{\mathrm{pl}}|s) \;=\; \frac{\exp Q^{\mathrm{pl}}(s,a^{\mathrm{pl}})}{\sum_{a^{\mathrm{pl}}}\exp Q^{\mathrm{pl}}(s,a^{\mathrm{pl}})}$$

Finally, computing the expectation with respect to the maximizing policy:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{a^{\mathrm{pl}}\sim\pi^{\mathrm{pl},*}(\cdot|s)}\left[Q^{\mathrm{pl}}(s,a^{\mathrm{pl}}) - \log\pi^{\mathrm{pl}}(a^{\mathrm{pl}}|s)\right] \\
&= \sum_{a^{\mathrm{pl}}}\pi^{\mathrm{pl},*}(a^{\mathrm{pl}}|s)\left(Q^{\mathrm{pl}}(s,a^{\mathrm{pl}}) - \log\pi^{\mathrm{pl},*}(a^{\mathrm{pl}}|s)\right) \\
&= \sum_{a^{\mathrm{pl}}}\frac{\exp Q^{\mathrm{pl}}(s,a^{\mathrm{pl}})}{\sum_{a^{\mathrm{pl}}}\exp Q^{\mathrm{pl}}(s,a^{\mathrm{pl}})}\left(Q^{\mathrm{pl}}(s,a^{\mathrm{pl}}) - \log\frac{\exp Q^{\mathrm{pl}}(s,a^{\mathrm{pl}})}{\sum_{a^{\mathrm{pl}}}\exp Q^{\mathrm{pl}}(s,a^{\mathrm{pl}})}\right) \\
&= \sum_{a^{\mathrm{pl}}}\frac{\exp Q^{\mathrm{pl}}(s,a^{\mathrm{pl}})}{\sum_{a^{\mathrm{pl}}}\exp Q^{\mathrm{pl}}(s,a^{\mathrm{pl}})}\left(Q^{\mathrm{pl}}(s,a^{\mathrm{pl}}) - Q^{\mathrm{pl}}(s,a^{\mathrm{pl}}) + \log\sum_{a^{\mathrm{pl}}}\exp Q^{\mathrm{pl}}(s,a^{\mathrm{pl}})\right) \\
&= \sum_{a^{\mathrm{pl}}}\frac{\exp Q^{\mathrm{pl}}(s,a^{\mathrm{pl}})}{\sum_{a^{\mathrm{pl}}}\exp Q^{\mathrm{pl}}(s,a^{\mathrm{pl}})}\left(\log\sum_{a^{\mathrm{pl}}}\exp Q^{\mathrm{pl}}(s,a^{\mathrm{pl}})\right) \\
&= \log\sum_{a^{\mathrm{pl}}}\exp Q^{\mathrm{pl}}(s,a^{\mathrm{pl}}) \qquad (54)
\end{aligned}$$

Basically, we have shown that the optimization problem is solved when the player follows a soft-max policy with respect to the quality function $Q^{\mathrm{pl}}(a^{\mathrm{pl}}|s) = \min_{a^{\mathrm{op}}} Q(s,a^{\mathrm{pl}},a^{\mathrm{op}})$. This explains the steps for the player policy in Algorithm 2. In addition, replacing the definition $Q^{\mathrm{pl}}(a^{\mathrm{pl}}|s) = \min_{a^{\mathrm{op}}} Q(s,a^{\mathrm{pl}},a^{\mathrm{op}})$ in (54), one gets the saddle point update (53) in Algorithm 2.

We still need to proceed similarly to motivate the opponent policy derivation from the quality function (52). To this end, we maximize with respect to the player before minimizing for the opponent, we have:

$$\begin{aligned}
&\min_{\pi^{\mathrm{op}}}\max_{\pi^{\mathrm{pl}}} F(\pi^{\mathrm{pl}},\pi^{\mathrm{op}},s) \\
&= \min_{\pi^{\mathrm{op}}}\max_{\pi^{\mathrm{pl}}}\mathop{\mathbb{E}}_{a^{\mathrm{pl}}\sim\pi^{\mathrm{pl}}(\cdot|s),a^{\mathrm{op}}\sim\pi^{\mathrm{op}}(\cdot|s)}\left[Q(s,a^{\mathrm{pl}},a^{\mathrm{op}}) - \log\pi^{\mathrm{pl}}(a^{\mathrm{pl}}|s)\right] \\
&= \min_{\pi^{\mathrm{op}}}\max_{\pi^{\mathrm{pl}}}\mathop{\mathbb{E}}_{a^{\mathrm{op}}\sim\pi^{\mathrm{op}}(\cdot|s)}\left[\mathop{\mathbb{E}}_{a^{\mathrm{pl}}\sim\pi^{\mathrm{pl}}(\cdot|s)}\left[Q(s,a^{\mathrm{pl}},a^{\mathrm{op}}) - \log\pi^{\mathrm{pl}}(a^{\mathrm{pl}}|s)|a^{\mathrm{op}}\right]\right] \\
&= \min_{\pi^{\mathrm{op}}}\mathop{\mathbb{E}}_{a^{\mathrm{op}}\sim\pi^{\mathrm{op}}(\cdot|s)}\left[\max_{\pi^{\mathrm{pl}}}\mathop{\mathbb{E}}_{a^{\mathrm{pl}}\sim\pi^{\mathrm{pl}}(\cdot|s)}\left[Q(s,a^{\mathrm{pl}},a^{\mathrm{op}}) - \log\pi^{\mathrm{pl}}(a^{\mathrm{pl}}|s)|a^{\mathrm{op}}\right]\right]
\end{aligned}$$

The innermost maximization is solved again by observing that it is a concave function in the decision variables, normalizing one obtains the maximizer policy, and plugging that in the expectation gives the soft-max function with respect to the player action a^{pl}. We define this function as the quality function of the opponent, because it is the amount of information that can be used by the opponent to decide its move.

$$Q^{\mathrm{op}}(s,a^{\mathrm{op}}) \;=\; \log\sum_{a^{\mathrm{pl}}}\exp Q(s,a^{\mathrm{pl}},a^{\mathrm{op}})$$

It remains to face the external minimization with respect to the opponent policy. This is trivial, the opponent can simply act greedly since it is not regularized :

$$\min_{\pi^{\mathrm{op}}}\mathop{\mathbb{E}}_{a^{\mathrm{op}}\sim\pi^{\mathrm{op}}(\cdot|s)}\left[Q^{\mathrm{op}}(s,a^{\mathrm{op}})\right] \;=\; \min_{a^{\mathrm{op}}} Q^{\mathrm{op}}(s,a^{\mathrm{op}})$$

This second part clarifies the updates relative to the opponent in Algorithm 2.

Notice that the algorithm iterates in order to obtain a more and more precise estimate of the joint quality function $Q(s,a^{\mathrm{pl}},a^{\mathrm{op}})$. When it converges, the quality functions for the player and the agent respectively are obtained, thanks to the transformations illustrated here and in the body of Algorithm 2.

F.4 Proof of Theorem 3

Proof. Consider the following:

$$
\begin{aligned}
\left|V^{\pi_1}_{M^L_{\boldsymbol{\theta}^*}} - V^{\pi^{\mathrm{pl}}}_{M^L_{\boldsymbol{\theta}^*}}\right| &\leq \left|V^{\pi_1}_{M^L_{\boldsymbol{\theta}^*}} - V^{\pi^*_{M^L_{\boldsymbol{\theta}^*}}}_{M^L_{\boldsymbol{\theta}^*}}\right| + \left|V^{\pi^*_{M^L_{\boldsymbol{\theta}^*}}}_{M^L_{\boldsymbol{\theta}^*}} - V^{\pi^*_{M^E_{\boldsymbol{\theta}^*}}}_{M^E_{\boldsymbol{\theta}^*}}\right| + \left|V^{\pi^*_{M^E_{\boldsymbol{\theta}^*}}}_{M^E_{\boldsymbol{\theta}^*}} - V^{\alpha\pi^{\mathrm{pl}}+(1-\alpha)\pi^{\mathrm{op}}}_{M^L_{\boldsymbol{\theta}^*}}\right| + \left|V^{\alpha\pi^{\mathrm{pl}}+(1-\alpha)\pi^{\mathrm{op}}}_{M^L_{\boldsymbol{\theta}^*}} - V^{\pi^{\mathrm{pl}}}_{M^L_{\boldsymbol{\theta}^*}}\right| \\
&\overset{\mathrm{a}}{=} \left|V^{\pi^*_{M^L_{\boldsymbol{\theta}^*}}}_{M^L_{\boldsymbol{\theta}^*}} - V^{\pi^*_{M^E_{\boldsymbol{\theta}^*}}}_{M^E_{\boldsymbol{\theta}^*}}\right| + \left|V^{\alpha\pi^{\mathrm{pl}}+(1-\alpha)\pi^{\mathrm{op}}}_{M^L_{\boldsymbol{\theta}^*}} - V^{\pi^{\mathrm{pl}}}_{M^L_{\boldsymbol{\theta}^*}}\right| \\
&\overset{\mathrm{b}}{\leq} \frac{\gamma\cdot|R_{\boldsymbol{\theta}^*}|^{\max}}{(1-\gamma)^2}\cdot d_{\mathrm{dyn}}\left(T^L,T^E\right) + \left|V^{\alpha\pi^{\mathrm{pl}}+(1-\alpha)\pi^{\mathrm{op}}}_{M^L_{\boldsymbol{\theta}^*}} - V^{\pi^{\mathrm{pl}}}_{M^L_{\boldsymbol{\theta}^*}}\right| \\
&\leq \frac{\gamma\cdot|R_{\boldsymbol{\theta}^*}|^{\max}}{(1-\gamma)^2}\cdot d_{\mathrm{dyn}}\left(T^L,T^E\right) + \frac{|R_{\boldsymbol{\theta}^*}|^{\max}}{1-\gamma}\cdot\left\|\boldsymbol{\rho}^{\alpha\pi^{\mathrm{pl}}+(1-\alpha)\pi^{\mathrm{op}}}_{M^L} - \boldsymbol{\rho}^{\pi^{\mathrm{pl}}}_{M^L}\right\|_1 \\
&\overset{\mathrm{c}}{\leq} \frac{\gamma\cdot|R_{\boldsymbol{\theta}^*}|^{\max}}{(1-\gamma)^2}\cdot d_{\mathrm{dyn}}\left(T^L,T^E\right) + \frac{|R_{\boldsymbol{\theta}^*}|^{\max}}{(1-\gamma)^2}\cdot\max_s\left\|\alpha\pi^{\mathrm{pl}} + (1-\alpha)\pi^{\mathrm{op}}(\cdot|s) - \pi^{\mathrm{pl}}(\cdot|s)\right\|_1 \\
&= \frac{\gamma\cdot|R_{\boldsymbol{\theta}^*}|^{\max}}{(1-\gamma)^2}\cdot d_{\mathrm{dyn}}\left(T^L,T^E\right) + \frac{|R_{\boldsymbol{\theta}^*}|^{\max}}{(1-\gamma)^2}\cdot(1-\alpha)\cdot\max_s\left\|\pi^{\mathrm{op}}(\cdot|s) - \pi^{\mathrm{pl}}(\cdot|s)\right\|_1 \\
&\leq \frac{\gamma\cdot|R_{\boldsymbol{\theta}^*}|^{\max}}{(1-\gamma)^2}\cdot d_{\mathrm{dyn}}\left(T^L,T^E\right) + \frac{|R_{\boldsymbol{\theta}^*}|^{\max}}{(1-\gamma)^2}\cdot(1-\alpha)\cdot 2
\end{aligned}
$$

where a is due to the fact that $\boldsymbol{\rho}^{\pi_1}_{M^L} = \boldsymbol{\rho}^{\pi^*_{M^L_{\boldsymbol{\theta}^*}}}_{M^L}$ and $\boldsymbol{\rho}^{\pi^*_{M^E_{\boldsymbol{\theta}^*}}}_{M^E} = \boldsymbol{\rho}^{\alpha\pi^{\mathrm{pl}}+(1-\alpha)\pi^{\mathrm{op}}}_{M^L}$; b is due to Theorem 7 in [34]; and c is due to Lemma A.1 in [59].

□

F.5 Suboptimality gap for the Robust MCE-IRL in the infeasible case

In the main text, we always assume that the condition of Theorem 2 holds. In that case, the problem (1) is feasible, and the performance gap guarantee of Robust MCE IRL provided by Theorem 3 is weaker than that of the standard MCE IRL. Here, instead we consider the case where the condition of Theorem 2 does not hold[7].

Theorem 9. *When the condition in Theorem 2 does not hold, the performance gap between the policies π_1 and π^{pl} in the MDP $M^L_{\boldsymbol{\theta}^*}$ is bounded as follows:*

$$
\begin{aligned}
\left|V^{\pi_1}_{M^L_{\boldsymbol{\theta}^*}} - V^{\pi^{\mathrm{pl}}}_{M^L_{\boldsymbol{\theta}^*}}\right| \leq\ & \frac{\gamma\cdot|R_{\boldsymbol{\theta}^*}|^{\max}}{(1-\gamma)^2}\cdot d_{\mathrm{dyn}}\left(T^L,T^E\right) + \\
& \frac{\gamma\cdot|R_{\boldsymbol{\theta}^*}|^{\max}}{(1-\gamma)^2}2(1-\alpha)^2 + \frac{|R_{\boldsymbol{\theta}^*}|^{\max}}{(1-\gamma)^2}d_{\mathrm{pol}}\left(\pi^*_{M^E_{\boldsymbol{\theta}^*}},\pi^{\mathrm{soft}}_{M^E_{\boldsymbol{\theta}^*}}\right) \\
& \frac{2\cdot\kappa^2_{\boldsymbol{\theta}^*}\cdot|R_{\boldsymbol{\theta}^*}|^{\max}}{(1-\gamma)^4}\left[\alpha\cdot d_{\mathrm{dyn}}\left(T^E,T^L\right) + (1-\alpha)\cdot d_{\mathrm{dyn}}\left(T^E,T^*\right)\right]
\end{aligned}
$$

where T^* minimizes (7).

Proof.

$$
\begin{aligned}
\left|V^{\pi_1}_{M^L_{\boldsymbol{\theta}^*}} - V^{\pi^{\mathrm{pl}}}_{M^L_{\boldsymbol{\theta}^*}}\right| &\leq \left|V^{\pi_1}_{M^L_{\boldsymbol{\theta}^*}} - V^{\pi^*_{M^L_{\boldsymbol{\theta}^*}}}_{M^L_{\boldsymbol{\theta}^*}}\right| + \left|V^{\pi^*_{M^L_{\boldsymbol{\theta}^*}}}_{M^L_{\boldsymbol{\theta}^*}} - V^{\pi^*_{M^E_{\boldsymbol{\theta}^*}}}_{M^E_{\boldsymbol{\theta}^*}}\right| + \left|V^{\pi^*_{M^E_{\boldsymbol{\theta}^*}}}_{M^E_{\boldsymbol{\theta}^*}} - V^{\pi^{\mathrm{pl}}}_{\boldsymbol{\theta}^*,\alpha T^L+(1-\alpha)T^*}\right| + \left|V^{\pi^{\mathrm{pl}}}_{\boldsymbol{\theta}^*,\alpha T^L+(1-\alpha)T^*} - V^{\pi^{\mathrm{pl}}}_{M^L_{\boldsymbol{\theta}^*}}\right| \\
&\overset{\mathrm{a}}{=} \underbrace{\left|V^{\pi^*_{M^L_{\boldsymbol{\theta}^*}}}_{M^L_{\boldsymbol{\theta}^*}} - V^{\pi^*_{M^E_{\boldsymbol{\theta}^*}}}_{M^E_{\boldsymbol{\theta}^*}}\right|}_{\text{Demonstration difference}} + \underbrace{\left|V^{\pi^{\mathrm{pl}}}_{\boldsymbol{\theta}^*,\alpha T^L+(1-\alpha)T^*} - V^{\pi^{\mathrm{pl}}}_{M^L_{\boldsymbol{\theta}^*}}\right|}_{\text{Transfer difference}} + \underbrace{\left|V^{\pi^*_{M^E_{\boldsymbol{\theta}^*}}}_{M^E_{\boldsymbol{\theta}^*}} - V^{\pi^{\mathrm{pl}}}_{\boldsymbol{\theta}^*,\alpha T^L+(1-\alpha)T^*}\right|}_{\text{infeasibility error}}
\end{aligned}
$$

[7] It follows that the policy output by Algorithm 1 is not in the feasible set of the problem 1

The Demonstration difference is bounded using Theorem 7 in [34], i.e.

$$\left| V^{\pi^*_{M^L_{\boldsymbol{\theta}^*}}}_{M^L_{\boldsymbol{\theta}^*}} - V^{\pi^*_{M^E_{\boldsymbol{\theta}^*}}}_{M^E_{\boldsymbol{\theta}^*}} \right| \leq \frac{\gamma \cdot |R_{\boldsymbol{\theta}^*}|^{\max}}{(1-\gamma)^2} \cdot d_{\mathrm{dyn}}\left(T^L, T^E\right) \tag{55}$$

The transfer error can be bound as:

$$\begin{aligned}
\left| V^{\pi^{\mathrm{pl}}}_{\boldsymbol{\theta}^*, \alpha T^L + (1-\alpha)T^*} - V^{\pi^{\mathrm{pl}}}_{M^L_{\boldsymbol{\theta}^*}} \right| &\overset{\mathrm{a}}{\leq} \frac{\gamma \cdot |R_{\boldsymbol{\theta}^*}|^{\max}}{(1-\gamma)^2} \cdot d_{\mathrm{dyn}}\left(\alpha T^L + (1-\alpha)T^*, T^L\right) \\
&= \frac{\gamma \cdot |R_{\boldsymbol{\theta}^*}|^{\max}}{(1-\gamma)^2} (1-\alpha) \cdot d_{\mathrm{dyn}}\left(T^*, T^L\right) \\
&= \frac{\gamma \cdot |R_{\boldsymbol{\theta}^*}|^{\max}}{(1-\gamma)^2} 2(1-\alpha)^2
\end{aligned}$$

where in a, we used the Simulation Lemma [35, 36].

Finally, for the infeasibility error

$$\begin{aligned}
\left| V^{\pi^*_{M^E_{\boldsymbol{\theta}^*}}}_{M^E_{\boldsymbol{\theta}^*}} - V^{\pi^{\mathrm{pl}}}_{\boldsymbol{\theta}^*, \alpha T^L + (1-\alpha)T^*} \right| &\leq \left| V^{\pi^*_{M^E_{\boldsymbol{\theta}^*}}}_{M^E_{\boldsymbol{\theta}^*}} - V^{\pi^{\mathrm{soft}}_{M^E_{\boldsymbol{\theta}^*}}}_{M^E_{\boldsymbol{\theta}^*}} \right| + \left| V^{\pi^{\mathrm{soft}}_{M^E_{\boldsymbol{\theta}^*}}}_{M^E_{\boldsymbol{\theta}^*}} - V^{\pi^{\mathrm{pl}}}_{\boldsymbol{\theta}^*, \alpha T^L + (1-\alpha)T^*} \right| \\
&\overset{\mathrm{a}}{\leq} \frac{|R_{\boldsymbol{\theta}^*}|^{\max}}{(1-\gamma)^2} d_{\mathrm{pol}}\left(\pi^*_{M^E_{\boldsymbol{\theta}^*}}, \pi^{\mathrm{soft}}_{M^E_{\boldsymbol{\theta}^*}}\right) + \frac{2 \cdot \kappa^2_{\boldsymbol{\theta}^*} \cdot |R_{\boldsymbol{\theta}^*}|^{\max}}{(1-\gamma)^4} d_{\mathrm{dyn}}\left(T^E, \alpha T^L + (1-\alpha)T^*\right) \\
&\leq \frac{|R_{\boldsymbol{\theta}^*}|^{\max}}{(1-\gamma)^2} d_{\mathrm{pol}}\left(\pi^*_{M^E_{\boldsymbol{\theta}^*}}, \pi^{\mathrm{soft}}_{M^E_{\boldsymbol{\theta}^*}}\right) + \\
&\quad \frac{2 \cdot \kappa^2_{\boldsymbol{\theta}^*} \cdot |R_{\boldsymbol{\theta}^*}|^{\max}}{(1-\gamma)^4} \left[\alpha \cdot d_{\mathrm{dyn}}\left(T^E, T^L\right) + (1-\alpha) \cdot d_{\mathrm{dyn}}\left(T^E, T^*\right)\right]
\end{aligned}$$

where in a we used follow from [59, Lemma A.1] for the first term and Lemma 1 on the second term.

It can be seen that in case of MCE IRL $\alpha = 1$, the infeasibility term can be bounded adding an additional term scaling linearly with the mismatch $d_{\mathrm{dyn}}(T^E, T^L)$, however when $\alpha < 1$, the bound dependent on the linear combination of the mismatches $\alpha \cdot d_{\mathrm{dyn}}(T^E, T^L) + (1-\alpha) \cdot d_{\mathrm{dyn}}(T^E, T^*)$ where T^* is a minimizer of (7). Therefore the bound is tighter for problems such that $d_{\mathrm{dyn}}(T^E, T^*) < d_{\mathrm{dyn}}(T^E, T^L)$. However, our bounds also explains that for $\alpha < 1$, we have nonzero bound on the transfer error that arises from the fact that the matching policy $\alpha\pi^{\mathrm{pl}} + (1-\alpha)\pi^{\mathrm{op}}$ is not equal to the evaluated policy π^{pl}. □

The following corollary provides a value of α for which we can attain better bound on the performance gap of Robust MCE IRL.

Corollary 2. *When the condition in Theorem 2 does not hold, the upper bound on the performance gap between the policies π_1 and π^{pl} in the MDP $M^L_{\boldsymbol{\theta}^*}$ given in Theorem 9 is minimized for the following choice of α:*

$$\alpha = \min\left(1, 1 - \frac{\kappa^2_{\theta^*}}{(1-\gamma)^2\gamma}\left(\frac{d_{\mathrm{dyn}}(T^E, T^L)}{2} - \frac{d_{\mathrm{dyn}}(T^*, T^E)}{2}\right)\right),$$

where T^* minimizes (7).

The suggested choice of α follows the intuition of having a decreasing α as the distance $d_{\mathrm{dyn}}(T^E, T^L)$ increases. However, it should be closer to 1 as the distance $d_{\mathrm{dyn}}(T^E, T^*)$ increases, i.e., a less powerful opponent should work better if the expert transition dynamics are not close to the worst ones (the ones that minimize (7)).

F.6 Proof of Theorem 4

Proof. For any policy π acting in the expert environment M^E, we can compute the state occupancy measures, as follows:

$$\rho^{\pi}_{M^E}(s_0) \;=\; 1 - \gamma \tag{56}$$

$$\rho^{\pi}_{M^E}(s_1) \;=\; (1-\epsilon_E)\cdot\gamma\cdot\pi(a_1|s_0) \tag{57}$$
$$\rho^{\pi}_{M^E}(s_2) \;=\; \epsilon_E\cdot\gamma\cdot\pi(a_1|s_0)+\gamma\cdot\pi(a_2|s_0) \tag{58}$$

Then, for the MDP $M^E_{\boldsymbol{\theta}^*}$ endowed with the true reward function $R_{\boldsymbol{\theta}^*}$, we have:

$$V^{\pi}_{M^E_{\boldsymbol{\theta}^*}} \;=\; \frac{\gamma}{1-\gamma}\cdot\left\{2\cdot(1-\epsilon_E)\cdot\pi(a_1|s_0)-1\right\}, \tag{59}$$

which is maximized when $\pi(a_1|s_0) = 1$. Therefore, the optimal expert policy is given by: $\pi^*_{M^E_{\boldsymbol{\theta}^*}}(a_1|s_0) = 1$ and $\pi^*_{M^E_{\boldsymbol{\theta}^*}}(a_2|s_0) = 0$, with the corresponding optimal value $V^{\pi^*_{M^E_{\boldsymbol{\theta}^*}}}_{M^E_{\boldsymbol{\theta}^*}} = \frac{\gamma}{1-\gamma}\cdot(1-2\epsilon_E)$.

On the learner side (M^L), Algorithm 1 converges when the occupancy measure of the mixture policy $\alpha\pi^{\mathrm{pl}}+(1-\alpha)\pi^{\mathrm{op}}$ matches the expert's occupancy measure. First, we compute the occupancy measures for the mixture policy:

$$\begin{aligned}
\rho^{\alpha\pi^{\mathrm{pl}}+(1-\alpha)\pi^{\mathrm{op}}}_{M^L}(s_0) \;&=\; 1-\gamma\\
\rho^{\alpha\pi^{\mathrm{pl}}+(1-\alpha)\pi^{\mathrm{op}}}_{M^L}(s_1) \;&=\; \gamma\cdot\left\{\alpha\cdot\pi^{\mathrm{pl}}(a_1|s_0)+(1-\alpha)\cdot\pi^{\mathrm{op}}(a_1|s_0)\right\}\\
\rho^{\alpha\pi^{\mathrm{pl}}+(1-\alpha)\pi^{\mathrm{op}}}_{M^L}(s_2) \;&=\; \gamma\cdot\left\{\alpha\cdot\pi^{\mathrm{pl}}(a_2|s_0)+(1-\alpha)\cdot\pi^{\mathrm{op}}(a_2|s_0)\right\}
\end{aligned}$$

Here, the worst-case opponent is given by $\pi^{\mathrm{op}}(a_1|s_0) = 0$ and $\pi^{\mathrm{op}}(a_2|s_0) = 1$. Note that the choice of the opponent does not rely on the unknown reward function. Instead, we choose as opponent the policy that takes the action leading to the state where the demonstrated occupancy measure is lower. Then, the above expressions reduce to:

$$\begin{aligned}
\rho^{\alpha\pi^{\mathrm{pl}}+(1-\alpha)\pi^{\mathrm{op}}}_{M^L}(s_0) \;&=\; 1-\gamma\\
\rho^{\alpha\pi^{\mathrm{pl}}+(1-\alpha)\pi^{\mathrm{op}}}_{M^L}(s_1) \;&=\; \gamma\cdot\alpha\cdot\pi^{\mathrm{pl}}(a_1|s_0)\\
\rho^{\alpha\pi^{\mathrm{pl}}+(1-\alpha)\pi^{\mathrm{op}}}_{M^L}(s_2) \;&=\; \gamma\cdot\left\{\alpha\cdot\pi^{\mathrm{pl}}(a_2|s_0)+(1-\alpha)\right\}
\end{aligned}$$

Now, we match the above occupancy measures with the expert occupancy measures (Eqs. (56)-(58) with $\pi \leftarrow \pi^*_{M^E_{\boldsymbol{\theta}^*}}$):

$$\begin{aligned}
1-\epsilon_E \;&=\; \alpha\cdot\pi^{\mathrm{pl}}(a_1|s_0)\\
\epsilon_E \;&=\; \alpha\cdot\pi^{\mathrm{pl}}(a_2|s_0)+(1-\alpha)
\end{aligned}$$

Thus, we get: $\pi^{\mathrm{pl}}(a_1|s_0) = \frac{1-\epsilon_E}{\alpha}$ and $\pi^{\mathrm{pl}}(a_2|s_0) = \frac{\alpha-(1-\epsilon_E)}{\alpha}$. Note that π^{pl} is well-defined when $\alpha \geq 1-\epsilon_E$.

Given $\alpha \geq 1-\epsilon_E$, the state occupancy measure of π^{pl} in the MDP M^L is given by:

$$\begin{aligned}
\rho^{\pi^{\mathrm{pl}}}_{M^L}(s_0) \;&=\; 1-\gamma\\
\rho^{\pi^{\mathrm{pl}}}_{M^L}(s_1) \;&=\; \gamma\cdot\pi^{\mathrm{pl}}(a_1|s_0) \;=\; \gamma\cdot\frac{1-\epsilon_E}{\alpha}\\
\rho^{\pi^{\mathrm{pl}}}_{M^L}(s_2) \;&=\; \gamma\cdot\pi^{\mathrm{pl}}(a_2|s_0) \;=\; \gamma\cdot\frac{\alpha-(1-\epsilon_E)}{\alpha}
\end{aligned}$$

Then, the expected return of π^{pl} in the MDP $M^L_{\boldsymbol{\theta}^*}$ is given by:

$$V^{\pi^{\mathrm{pl}}}_{M^L_{\boldsymbol{\theta}^*}} \;=\; \frac{\gamma}{1-\gamma}\cdot\frac{2\cdot(1-\epsilon_E)-\alpha}{\alpha}.$$

Consider the MCE IRL learner receiving the expert occupancy measure $\boldsymbol{\rho}$ from the learner environment M^L itself, i.e., $\boldsymbol{\rho} = \boldsymbol{\rho}^{\pi^*_{M^L_{\boldsymbol{\theta}^*}}}_{M^L}$. Note that $\pi^*_{M^L_{\boldsymbol{\theta}^*}}(a_1|s_0) = 1$, and $\pi^*_{M^L_{\boldsymbol{\theta}^*}}(a_2|s_0) = 0$. In this case, the learner recovers a policy $\pi_1 := \pi^{\mathrm{soft}}_{M^L_{\boldsymbol{\theta_L}}}$ such that $\boldsymbol{\rho}^{\pi_1}_{M^L} = \boldsymbol{\rho}^{\pi^*_{M^L_{\boldsymbol{\theta}^*}}}_{M^L}$. Thus, we have $V^{\pi_1}_{M^L_{\boldsymbol{\theta}^*}} = V^{\pi^*_{M^L_{\boldsymbol{\theta}^*}}}_{M^L_{\boldsymbol{\theta}^*}} = \frac{\gamma}{1-\gamma}$. Consequently, for this example, the performance gap is given by:

$$\left|V^{\pi_1}_{M^L_{\boldsymbol{\theta}^*}} - V^{\pi^{\mathrm{pl}}}_{M^L_{\boldsymbol{\theta}^*}}\right| \;=\; \left|\frac{\gamma}{1-\gamma}\cdot\left\{1-\frac{2\cdot(1-\epsilon_E)-\alpha}{\alpha}\right\}\right| \;=\; \frac{2\cdot\gamma}{1-\gamma}\cdot\left|\frac{\alpha-(1-\epsilon_E)}{\alpha}\right|.$$

The following two cases are of particular interest:

- For $\alpha = 1 - \epsilon_E = 1 - \frac{d_{\mathrm{dyn}}\left(T^L, T^E\right)}{2}$, the performance gap vanishes. This indicates that our Algorithm 1 can recover the optimal performance even under dynamics mismatch.
- For $\alpha = 1$ (corresponding to the standard MCE IRL), the performance gap is given by:

$$\left| V^{\pi_1}_{M^L_{\boldsymbol{\theta}^*}} - V^{\pi^{\mathrm{p1}}}_{M^L_{\boldsymbol{\theta}^*}} \right| \;=\; \frac{2 \cdot \gamma \cdot \epsilon_E}{1-\gamma} \;=\; \frac{\gamma}{1-\gamma} \cdot d_{\mathrm{dyn}}\left(T^L, T^E\right).$$

□

G Further Details of Section 5

G.1 Hyperparameter Details and Additional Results

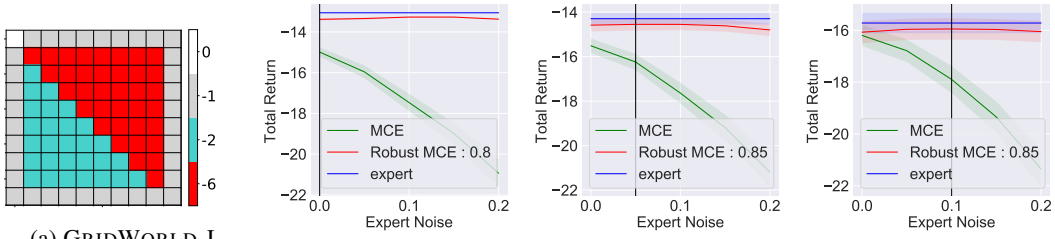
Here, we present the Figures 10, and 11, mentioned in the main text. All the hyperparameter details are reported in Tables 2, 3 and 4. We consider a uniform initial distribution P_0. For the performance evaluation of the learned policies, we compute the average reward of $1000 \times |\mathcal{S}|$ trajectories; along with this mean, we have reported the SD as well.

G.2 Low Dimensional Features

We consider a GRIDWORLD-L environment with a low dimensional (of dimension 3) binary feature mapping $\phi : \mathcal{S} \to \{0,1\}^3$. For any state $s \in \mathcal{S}$, the first two entries of the vector $\phi(s)$ are defined as follows:

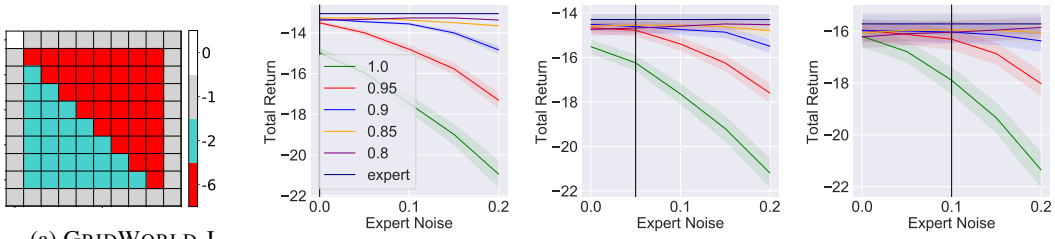
$$\phi(s)_i = \begin{cases} 1 & \text{the danger is of type-i in the state } s \\ 0 & \text{otherwise} \end{cases}$$

Whereas, the last entry of the vector $\phi(s) = 1$ for non-terminal states. The true reward function is given by $R_{\mathbf{w}}(s) = \langle \mathbf{w}, \phi(s) \rangle$, where $\mathbf{w} = [-2, -6, -1]$. In this low dimensional setting, our Algorithm 1 significantly outperforms the standard MCE IRL algorithm (see Figures 6, and 7).



(a) GRIDWORLD-L (b) M^{L,ϵ_L} with $\epsilon_L = 0$ (c) M^{L,ϵ_L} with $\epsilon_L = 0.05$ (d) M^{L,ϵ_L} with $\epsilon_L = 0.1$

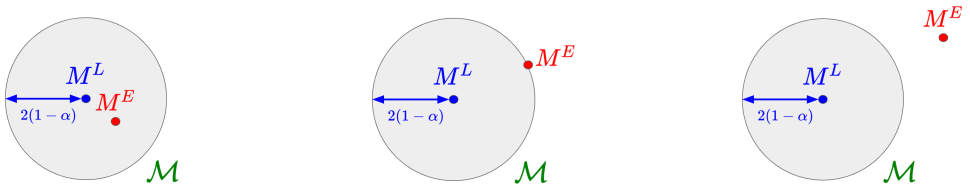
Figure 6: Comparison of the performance our Algorithm 1 against the baselines, under different levels of mismatch: $(\epsilon_E, \epsilon_L) \in \{0.0, 0.05, 0.1, 0.15, 0.2\} \times \{0.0, 0.05, 0.1\}$. Each plot corresponds to a fixed leaner environment M^{L,ϵ_L} with $\epsilon_L \in \{0.0, 0.05, 0.1\}$. The values of α used for our Algorithm 1 are reported in the legend. The vertical line indicates the position of the learner environment in the x-axis.



(a) GRIDWORLD-L (b) M^{L,ϵ_L} with $\epsilon_L = 0$ (c) M^{L,ϵ_L} with $\epsilon_L = 0.05$ (d) M^{L,ϵ_L} with $\epsilon_L = 0.1$

Figure 7: Comparison of the performance our Algorithm 1 with different values of α, under different levels of mismatch: $(\epsilon_E, \epsilon_L) \in \{0.0, 0.05, 0.1, 0.15, 0.2\} \times \{0.0, 0.05, 0.1\}$. Each plot corresponds to a fixed leaner environment M^{L,ϵ_L} with $\epsilon_L \in \{0.0, 0.05, 0.1\}$. The values of α used for our Algorithm 1 are reported in the legend. The vertical line indicates the position of the learner environment in the x-axis.

G.3 Impact of the Opponent Strength Parameter $1-\alpha$ on Robust MCE IRL



(a) Overestimating $1-\alpha$ (b) Perfect estimation of $1-\alpha$ (c) Underestimating $1-\alpha$

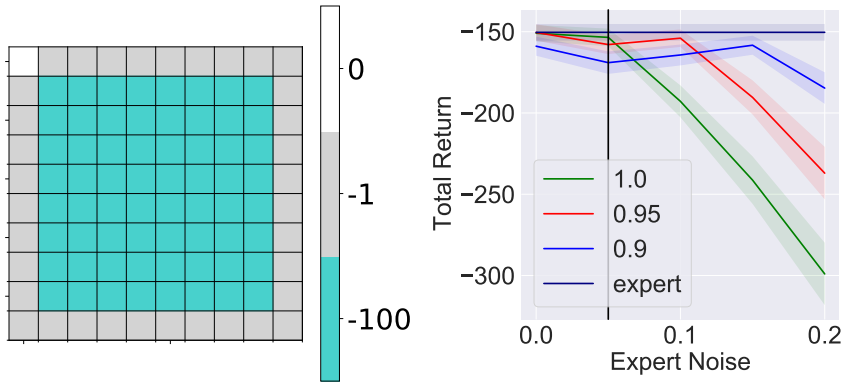
Figure 8: Illustration of the three cases related to the choice of the opponent strength parameter $1-\alpha$.

Here, we study the effect of the opponent strength parameter $(1-\alpha)$ on the performance of our Algorithm 1. Consider the uncertainty set associated with our Algorithm 1:

$$\mathcal{T}^{L,\alpha} = \left\{T : d_{\text{dyn}}\left(T, T^L\right) \leq 2(1-\alpha)\right\}.$$

Ideally, we prefer to choose the smallest set $\mathcal{T}^{L,\alpha}$ s.t. $T^E \in \mathcal{T}^{L,\alpha}$. To this end, we consider the following three cases (see Figure 8):

1. overestimating the opponent strength, i.e., $1-\alpha > \frac{d_{\text{dyn}}\left(T^E,T^L\right)}{2}$.
2. perfect estimation of the opponent strength, i.e., $1-\alpha = \frac{d_{\text{dyn}}\left(T^E,T^L\right)}{2}$.
3. underestimating the opponent strength, i.e., $1-\alpha < \frac{d_{\text{dyn}}\left(T^E,T^L\right)}{2}$.



(a) GRIDWORLD-1 (b) M^{L,ϵ_L} with $\epsilon_L = 0.05$

Figure 9: Comparison of the performance our Algorithm 1 with different values of the player strength parameter $\alpha \in \{0.9, 0.95, 1.0\}$, under different levels of mismatch: $(\epsilon_E, \epsilon_L) \in \{0.0, 0.05, 0.1, 0.15, 0.2\} \times \{0.05\}$. The values of α used for our Algorithm 1 are reported in the legend. Every point in the x-axis denotes an expert environment M^{E,ϵ_E} with the corresponding ϵ_E. The vertical line indicates the position of the learner environment M^{L,ϵ_L} in the x-axis. Note that moving away from the vertical line increases the mismatch between the learner and the expert, i.e., $|\epsilon_L - \epsilon_E|$.

Now, consider the experimental setup described in Section 5. Recall that, in this setup, the distance between the learner and the expert environment is given by $d_{\text{dyn}}\left(T^{L,\epsilon_L}, T^{E,\epsilon_E}\right) = 2\left(1 - \frac{1}{|\mathcal{S}|}\right)|\epsilon_L - \epsilon_E|$. Thus, a reasonable choice for the opponent strength would be $1-\alpha \approx |\epsilon_L - \epsilon_E|$. We note the following behavior in Figure 9:

- For $\alpha = 1.0$ (—), we observe a linear decay in the performance when moving away from the vertical line, i.e, with the increase of mismatch. Note that this curve corresponds to the MCE IRL algorithm.
- For $\alpha = 0.95$ (—), we observe a linear decay in the performance when moving away from the vertical line, after $\epsilon_E = 0.10$. Note that, for $1 - \alpha \approx 0.05$, beyond $\epsilon_L \pm 0.05$ is underestimation region (here, $\epsilon_L = 0.05$).
- For $\alpha = 0.9$ (—), we observe a linear decay in the performance when moving away from the vertical line, after $\epsilon_E = 0.15$. Note that, for $1-\alpha \approx 0.1$, beyond $\epsilon_L \pm 0.1$ is underestimation region (here, $\epsilon_L = 0.05$).
- Within the overestimation region, choosing the larger value of $1-\alpha$ hinders the performance. For example, the region $\epsilon_L \pm 0.05$ is overestimation region for both $1 - \alpha \approx 0.05$ (—) and $1 - \alpha \approx 0.1$ (—). Within this region, the performance of (—) curve is lower than that of (—) curve.

In addition, in Figure 11, we note the following:

- In general, the curves $\alpha = 1.0$ (—), $\alpha = 0.95$ (—), and $1-\alpha \approx 0.1$ demonstrated the above discussed behavior on the right hand side of the vertical line. Note that the right hand side of the vertical line represents the setting where the expert environment is more stochastic/noisy than the learner environment.
- In general, the curves $\alpha = 1.0$ (—), $\alpha = 0.95$ (—), and $1 - \alpha \approx 0.1$ demonstrated a stable and good performance on the left hand side of the vertical line. Note that the left hand side of the vertical line represents the setting where the expert environment is more deterministic than the learner environment.

To choose the right value of α, that depends on $d_{\text{dyn}}\left(T^E, T^L\right)$, we need to have an estimate $\widehat{T}^E$ of the expert environment T^E. A few recent works [20, 21, 31] attempt to infer the expert's transition dynamics from the demonstration set or via additional information. Our robust IRL approach can be incorporated into this research vein to improve the IRL agent's performance further.

Table 2: Hyperparameters for the GRIDWORLD experiments

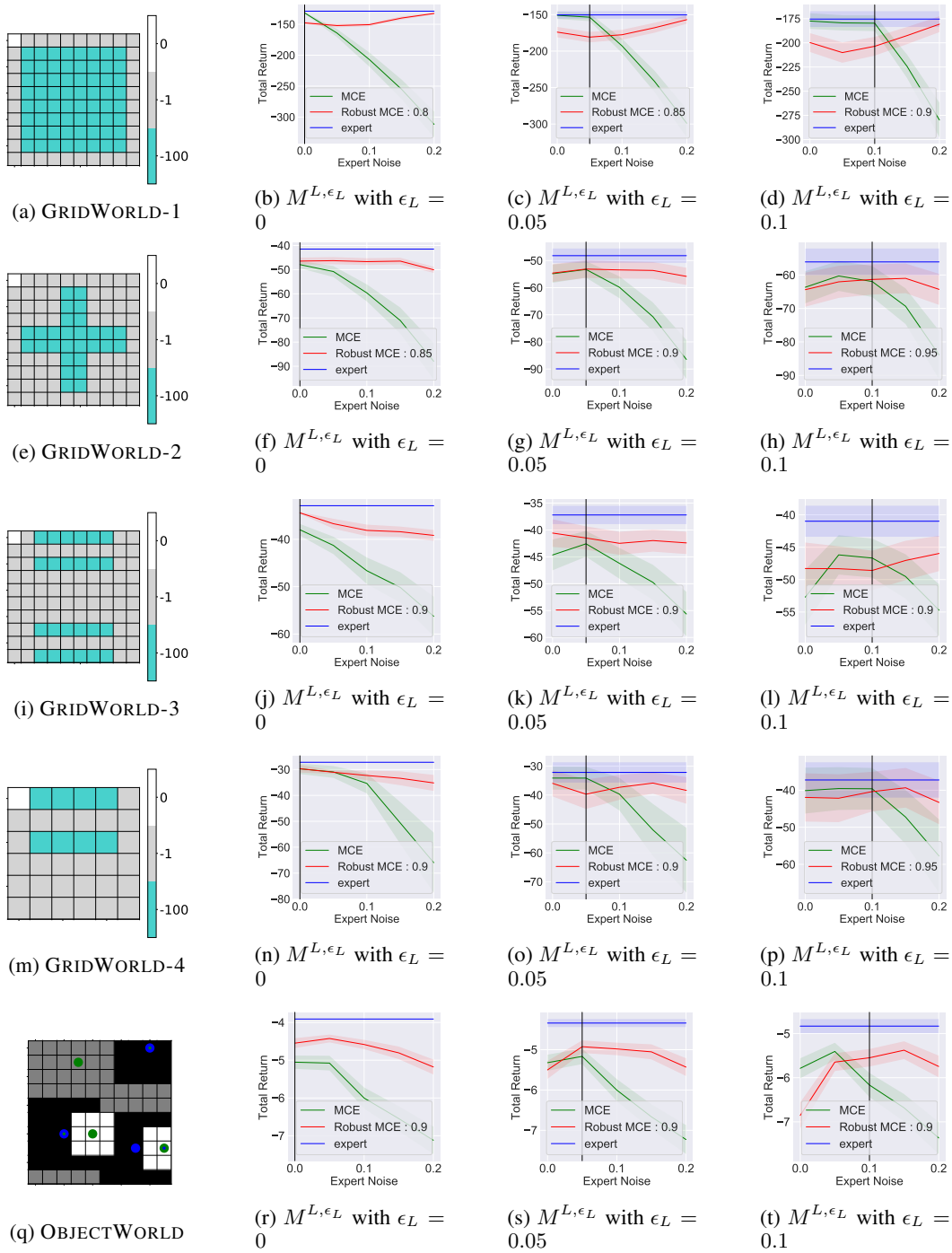
Hyperparameter	**Value**
IRL Optimizer	Adam
Learning rate	0.5
Weight decay	0.0
First moment exponential decay rate	0.9
Second moment exponential decay rate	0.99
Numerical stabilizer	$1e-7$
Number of steps	200
Discount factor γ	0.99

Table 3: Hyperparameters for the OBJECTWORLD experiments

Hyperparameter	**Value**
IRL Optimizer	Adam
Learning rate	$1e-3$
Weight decay	0.01
First moment exponential decay rate	0.9
Second moment exponential decay rate	0.999
Numerical stabilizer	$1e-8$
Number of steps	200
Reward network	two 2D-CNN layers; layers size = number of input features; ReLu
Discount factor γ	0.7

Table 4: Hyperparameters for the MDP solvers

Hyperparameter	**Value**
Two-Player soft value iteration tolerance	$1e-10$
Soft value iteration tolerance	$1e-10$
Value iteration tolerance	$1e-10$
Policy propagation tolerance	$1e-10$



(a) GRIDWORLD-1

(b) M^{L,ϵ_L} with $\epsilon_L = 0$

(c) M^{L,ϵ_L} with $\epsilon_L = 0.05$

(d) M^{L,ϵ_L} with $\epsilon_L = 0.1$

(e) GRIDWORLD-2

(f) M^{L,ϵ_L} with $\epsilon_L = 0$

(g) M^{L,ϵ_L} with $\epsilon_L = 0.05$

(h) M^{L,ϵ_L} with $\epsilon_L = 0.1$

(i) GRIDWORLD-3

(j) M^{L,ϵ_L} with $\epsilon_L = 0$

(k) M^{L,ϵ_L} with $\epsilon_L = 0.05$

(l) M^{L,ϵ_L} with $\epsilon_L = 0.1$

(m) GRIDWORLD-4

(n) M^{L,ϵ_L} with $\epsilon_L = 0$

(o) M^{L,ϵ_L} with $\epsilon_L = 0.05$

(p) M^{L,ϵ_L} with $\epsilon_L = 0.1$

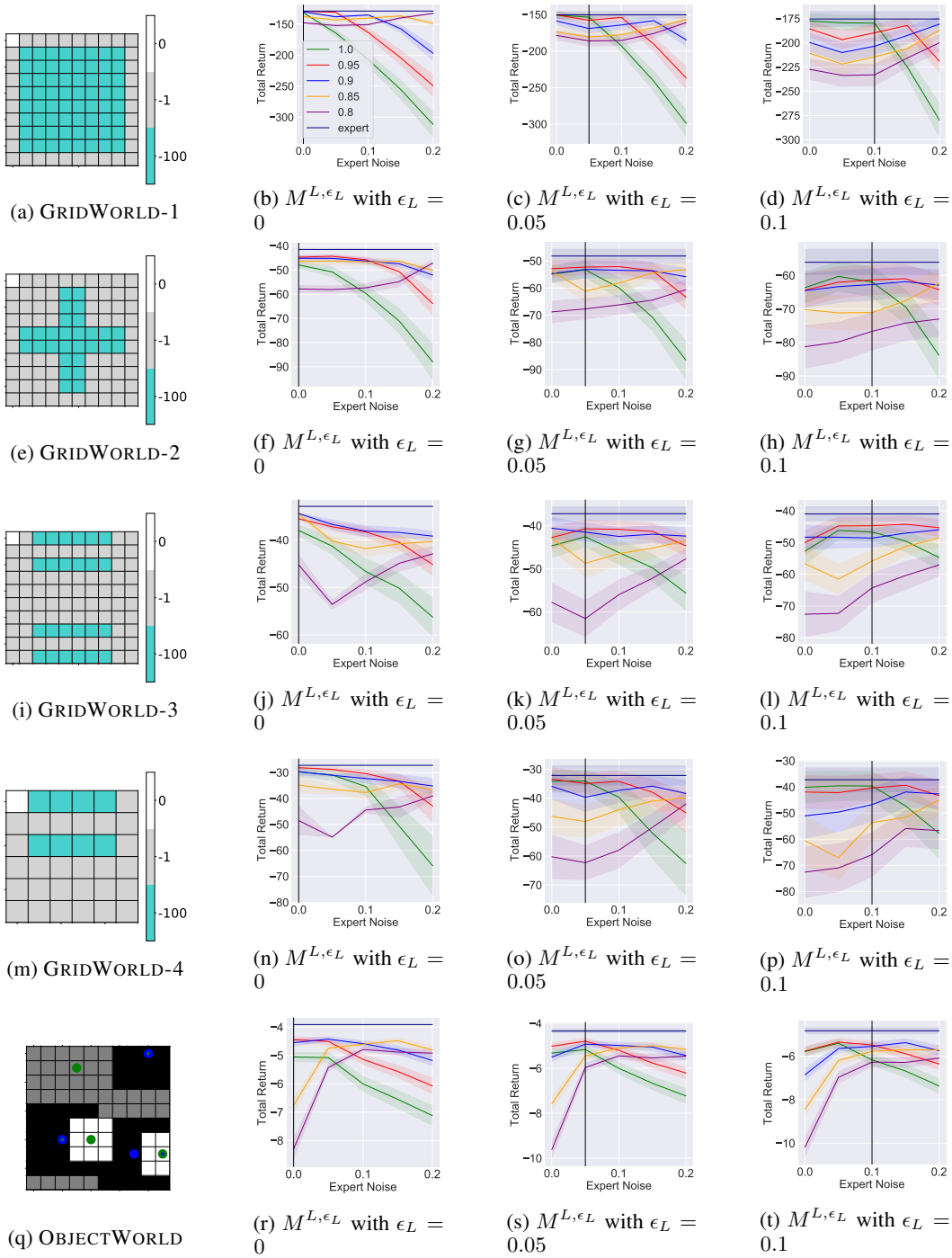
(q) OBJECTWORLD

(r) M^{L,ϵ_L} with $\epsilon_L = 0$

(s) M^{L,ϵ_L} with $\epsilon_L = 0.05$

(t) M^{L,ϵ_L} with $\epsilon_L = 0.1$

Figure 10: Comparison of the performance our Algorithm 1 against the baselines, under different levels of mismatch: $(\epsilon_E, \epsilon_L) \in \{0.0, 0.05, 0.1, 0.15, 0.2\} \times \{0.0, 0.05, 0.1\}$. Each plot corresponds to a fixed leaner environment M^{L,ϵ_L} with $\epsilon_L \in \{0.0, 0.05, 0.1\}$. The values of α used for our Algorithm 1 are reported in the legend. The vertical line indicates the position of the learner environment in the x-axis.



(a) GRIDWORLD-1 (b) M^{L,ϵ_L} with $\epsilon_L = 0$ (c) M^{L,ϵ_L} with $\epsilon_L = 0.05$ (d) M^{L,ϵ_L} with $\epsilon_L = 0.1$

(e) GRIDWORLD-2 (f) M^{L,ϵ_L} with $\epsilon_L = 0$ (g) M^{L,ϵ_L} with $\epsilon_L = 0.05$ (h) M^{L,ϵ_L} with $\epsilon_L = 0.1$

(i) GRIDWORLD-3 (j) M^{L,ϵ_L} with $\epsilon_L = 0$ (k) M^{L,ϵ_L} with $\epsilon_L = 0.05$ (l) M^{L,ϵ_L} with $\epsilon_L = 0.1$

(m) GRIDWORLD-4 (n) M^{L,ϵ_L} with $\epsilon_L = 0$ (o) M^{L,ϵ_L} with $\epsilon_L = 0.05$ (p) M^{L,ϵ_L} with $\epsilon_L = 0.1$

(q) OBJECTWORLD (r) M^{L,ϵ_L} with $\epsilon_L = 0$ (s) M^{L,ϵ_L} with $\epsilon_L = 0.05$ (t) M^{L,ϵ_L} with $\epsilon_L = 0.1$

Figure 11: Comparison of the performance our Algorithm 1 with different values of α, under different levels of mismatch: $(\epsilon_E, \epsilon_L) \in \{0.0, 0.05, 0.1, 0.15, 0.2\} \times \{0.0, 0.05, 0.1\}$. Each plot corresponds to a fixed leaner environment M^{L,ϵ_L} with $\epsilon_L \in \{0.0, 0.05, 0.1\}$. The values of α used for our Algorithm 1 are reported in the legend. The vertical line indicates the position of the learner environment in the x-axis.

H Further Details of Section 6

Algorithm 3 Robust RE IRL via Markov Game

Input: opponent strength $1-\alpha$, the expert's empirical feature occupancy measure $\bar{\phi}^E$
Initialize: player policy parameters $\boldsymbol{w}^{\mathrm{pl}}$, opponent policy parameters $\boldsymbol{w}^{\mathrm{op}}$, reward parameters $\boldsymbol{\theta}$
Initialize: uniform sampling policy π
while not converged **do**
  collect trajectories dataset $\mathcal{D}^\pi$ with the sampling policy π.
  estimate the features occupancy measure for each trajectory $\tau \in \mathcal{D}^\pi$ as $\bar{\phi}^\tau = \frac{1}{|\tau|}\sum_{s\in\tau}\phi(s)$.
  for $t = 1, \dots, N^\theta$ **do**
    update the distribution over trajectories as:

$$P(\tau|\boldsymbol{\theta}) \propto \exp\left(\left\langle \boldsymbol{\theta}, \bar{\phi}^\tau \right\rangle\right)$$

    compute the gradient estimate for updating $\boldsymbol{\theta}$ as proposed in [15] (to tackle the unknown transition dynamics case):

$$\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}) = \bar{\phi}^E - \sum_{\tau\in\mathcal{D}^\pi} P(\tau|\boldsymbol{\theta}) \cdot \bar{\phi}^\tau$$

    update the reward parameter $\boldsymbol{\theta}$ with Adam [39] using the gradient estimate $\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta})$.
  end for
  use Algorithm 4 with $R = R_{\boldsymbol{\theta}}$ to update π^{pl} and π^{op} s.t. they solve the following Markov Game approximately with policy gradient:

$$\max_{\pi^{\mathrm{pl}}\in\Pi} \min_{\pi^{\mathrm{op}}\in\Pi} \mathbb{E}\left[G \mid \pi^{\mathrm{pl}}, \pi^{\mathrm{op}}, M^{\mathrm{two},L,\alpha}\right]$$

  update the sampling policy:

$$\pi = \alpha\pi^{\mathrm{pl}} + (1-\alpha)\pi^{\mathrm{op}}$$

end while
Output: player policy π^{pl}

Algorithm 4 Policy Gradient Method for Two-Player Markov Game

Input: reward parameters $\boldsymbol{\theta}$
Initialize: player policy parameters $\boldsymbol{w}^{\mathrm{pl}}$, opponent policy parameters $\boldsymbol{w}^{\mathrm{op}}$
for $s = 1, \dots, N^\pi$ **do**
  $\mathcal{D} = \{\}$
  for $i = 1, \dots, N^{\mathrm{traj}}$ **do**
    collect trajectory a with $a_t^{\mathrm{pl}} \sim \pi^{\mathrm{pl}}(\cdot|s_t)$, $a_t^{\mathrm{op}} \sim \pi^{\mathrm{op}}(\cdot|s_t)$, $s_{t+1} \sim T^{\mathrm{two},L,\alpha}(\cdot|s_t, a_t^{\mathrm{pl}}, a_t^{\mathrm{op}})$.
    store the trajectory $\tau^i := \left\{(s_t, a_t^{\mathrm{pl}}, a_t^{\mathrm{op}})\right\}_t$ in $\mathcal{D}$.
    compute the return-to-go at each step of the trajectory τ^i as $G_t^i = \sum_{k=t+1}^{T}\gamma^{k-t-1}R(s_k)$.
  end for
  update the policy parameters (player and opponent) with the following gradient estimates:

$$\widehat{\nabla}_{\boldsymbol{w}^{\mathrm{pl}}} J(\boldsymbol{w}^{\mathrm{pl}}, \boldsymbol{w}^{\mathrm{op}}) = \frac{1}{|\mathcal{D}|}\sum_{\tau_i\in\mathcal{D}}\sum_t \gamma^t \nabla_{\boldsymbol{w}^{\mathrm{pl}}} \log \pi^{\mathrm{pl}}(a_t^{\mathrm{pl}}|s_t) G_t^i$$

$$\widehat{\nabla}_{\boldsymbol{w}^{\mathrm{op}}} J(\boldsymbol{w}^{\mathrm{pl}}, \boldsymbol{w}^{\mathrm{op}}) = -\frac{1}{|\mathcal{D}|}\sum_{\tau_i\in\mathcal{D}}\sum_t \gamma^t \nabla_{\boldsymbol{w}^{\mathrm{op}}} \log \pi^{\mathrm{op}}(a_t^{\mathrm{op}}|s_t) G_t^i$$

end for
Output: player policy $\pi^{\mathrm{pl}} \leftarrow \pi_{\boldsymbol{w}^{\mathrm{pl}}}$, opponent policy $\pi^{\mathrm{op}} \leftarrow \pi_{\boldsymbol{w}^{\mathrm{op}}}$

GAUSSIANGRID Environment. We consider a 2D environment, where we denote the horizontal coordinate as $x \in [0,1]$ and vertical one as $y \in [0,1]$. The agent starts in the upper left corner, i.e., the coordinate $(0,1)$, and the episode ends when the agent reaches the lower right region defined by the indicator function $\mathbf{1}\{x \in [0.95, 1], y \in [-1, -0.95]\}$. The reward function is given by:

$R(s) = R(x, y) = -(x-1)^2 - (y+1)^2 - 80 \cdot e^{-8(x^2+y^2)} + 10 \cdot \mathbf{1}\{x \in [0.95, 1], y \in [-1, -0.95]\}$. Note that the central region of the 2D environment represents a low reward area that should be avoided. The action space for the agent is given by $\mathcal{A} = [-0.5, 0.5]^2$, and the transition dynamics are given by:

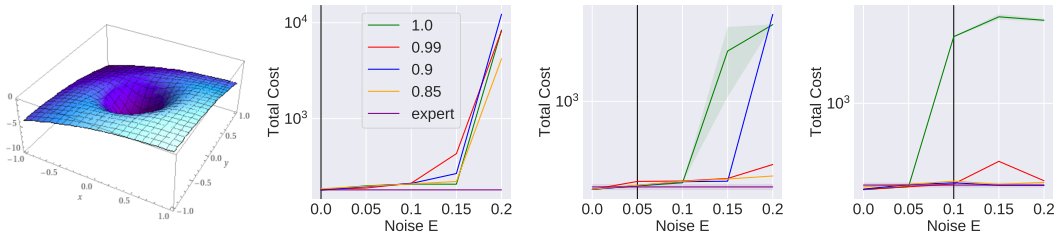
$$s_{t+1} = \begin{cases} s_t + \frac{a_t}{10} & \text{w.p.} \quad 1-\epsilon \\ s_t - \frac{s_t}{10\|s_t\|_2} & \text{w.p.} \quad \epsilon \end{cases}$$

Thus, with probability ϵ, the environment does not respond to the action taken by the agent, but it takes a step towards the low reward area centered at the origin, i.e., $-\frac{s_t}{10\|s_t\|_2}$. The agent should therefore pass far enough from the origin. The parameter ϵ can be varied to create a dynamic mismatch, e.g., higher ϵ corresponds to a more difficult environment. We investigate the performance of our Robust RE IRL method with different choices of the parameter α under various mismatches given by pairs (ϵ_E, ϵ_L). Let $\boldsymbol{\phi}(s) = \boldsymbol{\phi}(x, y) = \left[x^2, y^2, x, y, e^{-8(x^2+y^2)}, \mathbf{1}\left\{x \in [0.95, 1], y \in [-1, -0.95]\right\}, 1\right]^T$. The parameterization for both the player and opponent policies are given by:

$$a_t^{\text{pl}} \sim \mathcal{N}\left((\boldsymbol{w}^{\text{pl}})^T \phi(s_t), \Sigma^{\text{pl}}\right)$$

$$a_t^{\text{op}} \sim \mathcal{N}\left((\boldsymbol{w}^{\text{op}})^T \phi(s_t), \Sigma^{\text{op}}\right)$$

The covariance matrices $\Sigma^{\text{pl}}, \Sigma^{\text{op}}$ are constrained to be diagonal, and the diagonal elements are included as part of the policy parameterization.



(a) GAUSSIANGRID (b) M^{L,ϵ_L} with $\epsilon_L = 0$ (c) M^{L,ϵ_L} with $\epsilon_L = 0.05$ (d) M^{L,ϵ_L} with $\epsilon_L = 0.1$

Figure 12: Ablation of α in Algorithm 3 under different levels of mismatch: $(\epsilon_E, \epsilon_L) \in \{0.0, 0.05, 0.1, 0.15, 0.2\} \times \{0.0, 0.05, 0.1\}$. Each plot corresponds to a fixed leaner environment M^{L,ϵ_L} with $\epsilon_L \in \{0.0, 0.05, 0.1\}$. The values of α used in our Algorithm 3 are reported in the legend. The vertical line indicates the position of the learner environment in the x-axis. The results are averaged across 5 seeds.