Inverse Reinforcement Learning with Multiple Planning Horizons

Jiayu Yao
jiayu.yao@gladstone.ucsf.edu
Gladstone Institutes

Weiwei Pan
weiweipan@g.harvard.edu
SEAS, Harvard University

Finale Doshi-Velez
finale@seas.harvard.edu
SEAS, Harvard University

Barbara E Engelhardt
bengelhardt@stanford.edu
Gladstone Institutes & Stanford University

Abstract

We study an inverse reinforcement learning (IRL) problem where the experts are planning *under a shared reward function but with different, unknown planning horizons*. Without the knowledge of discount factors, the reward function has a larger feasible solution set, which makes it harder for existing IRL approaches to identify a reward function. To overcome this challenge, we develop two algorithms that can learn a global multi-agent reward function with agent-specific discount factors that reconstruct the expert policies. We characterize the feasible solution space of the reward function and discount factors for both algorithms and demonstrate the generalizability of the learned reward function across multiple domains.

1 Introduction

Designing reward functions in reinforcement learning (RL) that appropriately capture key aspects of a real-world task can be difficult in domains such as healthcare (Riachi et al., 2021) and finance (Charpentier et al., 2021). Inverse reinforcement learning (IRL) addresses this challenge by learning a reward function from expert demonstrations for a given task. The learned reward serves as a succinct description of the task and then can be transferred to similar tasks. Recent research in IRL has focused on the additional challenge of learning the reward function from heterogeneous expert behaviors, for example, where expert demonstrations vary in quality (Shiarlis et al., 2016; Brown et al., 2019), or where each expert is optimizing a different reward function (Mendez et al., 2018; Gleave & Habryka, 2018; Yu et al., 2019).

In this work, we focus on the setting where the expert behaviors vary because each expert is optimizing for *a different planning horizon, using the same reward function.* In RL, the planning horizon is encoded in the discount factor, which discounts future (expected) rewards attained by a given policy. A small discount factor corresponds to a short planning horizon, implying that the expert prioritizes short-term goals, whereas a large discount factor corresponds to a long planning horizon. For example, in an intensive care unit, the ventilation weaning practice is influenced by specific unit protocols and team or individual physician preferences (Kapnadak et al., 2015). Here,

a more aggressive weaning practice implies a small discount factor, since it prioritizes immediate changes in the patient's state. In a mobile health application that helps users manage their own wellness, users may select the planning horizon by choosing to set up short-term ($<$ 1 year), long-term (1-2 years), or maintenance ($>$ 2 years) health goals (Dicianno et al., 2017).

Existing IRL work often first chooses discount factors based on domain knowledge, and then learns the reward function by fixing these discount factors (Ng et al., 2000; Ziebart, 2010; Ramachandran & Amir, 2007). For entropy-regularized Markov decision processes (MDPs), when observing multiple experts and when the true discount factors of the experts are known, prior work shows that IRL can identify the true reward function up to a constant (Cao et al., 2021; Rolland et al., 2022). However, in most applications, we do not know the set of true discount factors a priori – this set must be learned alongside the reward function. Unfortunately, when the discount factors are misspecified or unknown, current IRL literature does not address the inference, nor the identifiability of the reward function. In this work, we fill this gap in the literature and study settings where both the discount factors (one per expert) and the reward function are unknown.

In this work, we assume that a global reward function is shared among the experts as the first step in understanding how unknown discount factors pose challenges to existing IRL work. This setting corresponds to experts having a shared goal, but different attention to the time needed to achieve them. We first provide analyses on the hardness of adapting existing IRL approaches to our problem setting, where both the set of discount factors and the reward must be inferred from the data. In particular, we consider two classes of popular IRL approaches: linear programming IRL (LP-IRL; Ng et al. (2000)) and max causal entropy IRL (MCE-IRL; Ziebart (2010)). We show that naive extensions of LP-IRL admit undesirable feasible solutions such as the degenerate solution wherein multiple experts are assigned the same discount factor.

In the case of MCE-IRL, whose Lagrangian dual problem can be interpreted as maximum likelihood IRL (ML-IRL) (Zeng et al., 2022), we show that when discount factors are unknown, strong duality does not hold for MCE-IRL and ML-IRL. Thus, solving ML-IRL does not guarantee convergence towards a feasible solution that reconstructs the expert policies. Furthermore, we show that for MCE-IRL, when the discount factors are misspecified, then either: (1) there does not exist a reward function that recovers the expert policy, or (2) there exists a unique reward function (up to a constant), which recovers the expert policy but may not be the true reward function. Fortunately, in Section 6.2, we observe that, in practice, when there are more than two experts, the set of reward functions that recovers the expert policy is non-empty only for a small set of discount factors. That is, in many applications, both the discount factors and the reward function are identifiable.

Finally, to address the failure modes of naive adaptations of LP-IRL and MCE-IRL, we develop two novel algorithms to learn a global multi-agent reward function with agent-specific discount factors based on LP-IRL and MCE-IRL. We (1) characterize the feasible solution space of the reward function and discount factors for both algorithms, and (2) empirically demonstrate the generalizability of the learned reward function across multiple domains.

2 Related Work

IRL with homogeneous demonstrations. Previous IRL work focuses on identifying a reward function that explains expert behavior when the demonstrations are generated by a single expert. For example, max-margin IRL methods (Ng et al., 2000; Abbeel & Ng, 2004) seek a reward function

that maximally separates the optimal policy and the second-most optimal policy. Max entropy IRL methods (Ziebart et al., 2008; Ziebart, 2010) estimate a reward function that maximizes the likelihood of the expert demonstrations. Bayesian IRL methods (Ramachandran & Amir, 2007; Jin et al., 2010) use prior knowledge to infer a posterior distribution over all possible reward functions. However, when given trajectories from multiple experts, each of these IRL approaches learns a *separate* reward function for each expert by default, which is data inefficient. In our work, we focus on a common scenario where the *global* reward function is shared by all experts with discount factors specific to each expert; these different discount factors lead to different optimal policies.

IRL with heterogeneous demonstrations. Recent IRL work explores heterogeneous expert demonstrations. Some methods study the scenario where expert demonstrations vary in quality. For example, Shiarlis et al. (2016) learn from demonstrations of both optimal policies and policies with undesirable behaviors (e.g., violating safety constraints). Similarly, Brown et al. (2019) assume that one has access to a set of demonstrations ranked by their expected return. Other work studies the setting where each expert optimizes for a different task. For example, Babes et al. (2011) first identify the tasks by clustering the expert demonstrations and then identify a reward function for each task. In Yu et al. (2019), the authors use deep latent generative models to capture the shared reward structure of expert demonstrations. Finally, Mendez et al. (2018) consider a lifelong learning setting where the agent faces a sequence of similar tasks and optimizes overall performance. In contrast, we consider the scenario where experts share the same reward function but have different planning horizons. To the best of our knowledge, we are the first to study this type of IRL problem.

Identifiability in IRL with respect to the reward function and discount factors. Recent work provides identifiability analysis on the reward function when the true discount factors are known for entropy-regularized MDPs (Cao et al., 2021; Rolland et al., 2022). Specifically, the authors show that, given optimal policies from two distinct discount factors and a shared reward function, one can identify the true reward function up to a constant. However, their analysis assumes that the true discount factors are provided, which is not realistic. In this work, we show that given misspecified discount factors, the feasible reward function set may not include the true reward function under the same rank conditions in Cao et al. (2021). We develop algorithms to recover the set of discount factors and the reward function for settings where both are unknown.

3 Problem Setting

Markov decision processes (MDPs). Consider an MDP, $\mathcal{M}^* = (\mathcal{S}, \mathcal{A}, r^*, T, \gamma^*)$, where $\mathcal{S}$ is a finite state space, $\mathcal{A}$ is a set of discrete actions, $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition dynamics describing the probability of reaching the next state s' by taking action a in the current state s, $r^*(s)$ is an action-independent reward function, and $\gamma^* \in [0, 1]$ is the discount factor that controls the weight of the future reward. For standard MDPs, the optimal policy is defined as,

$$\pi^*(a|s) = \arg\max_a Q_\pi^{r^*,\gamma^*}(s,a) = \arg\max_a \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^{*t} r^*(S_{t+1}) \middle| S_t = s, A_t = a \right]. \tag{1}$$

For entropy-regularized MDPs, the optimal policy is defined as,

$$\tilde{\pi}^*(a|s) = \arg\max_\pi \tilde{Q}_\pi^{r^*,\gamma^*}(s,a) = \arg\max_\pi \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^{*t} \left[r^*(S_{t+1}) + \lambda \mathcal{H}(\pi(\cdot|S_t)) \right] \middle| S_t = s, A_t = a \right], \tag{2}$$

where λ is a temperature parameter with a larger λ leading to a more stochastic policy. When $\lambda \to 0$, we will recover the expert policy of standard MDPs ($\pi^* = \tilde{\pi}^*$).

Multi-planning horizon IRL (MP-IRL). We assume that we are given an MDP, $\mathcal{M}^* \backslash \{r^*, \gamma^*\}$, with an unknown reward function and discount factor. We observe demonstrations from a set of K expert policies, each optimized under a shared global reward function r^*, and transition dynamics T, but using distinct discount factors ($\gamma_i^* \neq \gamma_j^*$ for $i \neq j$). We denote the set of distinct discount factors by $\Gamma^* = \{\gamma_k^*\}_{k=1}^K$. We assume that the expert policies are solved using Eq. 1 or 2 given appropriate contexts, and denote the set of expert policies for standard or entropy regularized MDPs by $\Pi^* = \{\pi_k^*\}_{k=1}^K$ or $\tilde{\Pi}^* = \{\tilde{\pi}_k^*\}_{k=1}^K$, respectively.

In MP-IRL, we wish to find a reward function r and a set of distinct discount factors $\Gamma = \{\gamma_k\}_{k=1}^K$ such that each expert policy remains optimal under the reconstructed MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, r, T, \gamma_k)$.

4 Algorithms for MP-IRL: LP-IRL

In this section, we introduce a popular class of IRL algorithms, linear programming IRL (LP-IRL), for the single known discount factor IRL setting. We explain how naive extensions of LP-IRL to the MP-IRL setting fail. Finally, we present a novel algorithm, multi-planning horizon LP-IRL (MPLP-IRL), that extends LP-IRL to jointly learn a set of distinct discount factors and a global reward function.

In the following, we assume that the expert policies are obtained by solving standard MDPs, and are denoted by, $\Pi^* = \{\pi_k^*\}_{k=1}^K$. In Ng et al. (2000), the authors assume that expert demonstrations are obtained from a single known planning horizon. They learn a reward function from expert demonstrations by solving the following optimization problem,

$$\max_r \sum_{s \in S} \min_{a \in \mathcal{A} \backslash \pi^*(s)} \{Q_{\pi^*}^{r,\gamma}(s, \pi^*(s)) - Q_{\pi^*}^{r,\gamma}(s, a)\} - \lambda \|r\|_1 \tag{3a}$$

$$\text{subject to } Q_{\pi^*}^{r,\gamma}(s, \pi^*(s)) - Q_{\pi^*}^{r,\gamma}(s, a) \geq 0 \quad \text{for } s \in \mathcal{S}, a \in \mathcal{A} \setminus \pi^*(s) \tag{3b}$$

$$|r| \preceq r_{\max} \tag{3c}$$

where the l_1 norm penalty, $\lambda \|r\|_1$, regularizes the sparsity of the reward function. In the above optimization problem, constraints (3b) ensure that π^* is optimal under the inferred reward function r. The reward function solution to Eq. 3a maximizes the sum of the differences of the Q-functions between the best and the next-best action over all states. The LP-IRL problem in Eq. 3 can be solved with linear programming (LP) (Ng et al., 2000). We extend the LP-IRL approach to expert demonstrations with multiple planning horizons. As such, we make an additional assumption:

Assumption 1. *For any two distinct expert policies $\pi_i^*, \pi_j^* \in \Pi^*$, optimized under γ_i^*, γ_j^*, respectively, there exists a state $s \in \mathcal{S}$ such that $Q_{\pi_i^*}^{r^*,\gamma_i^*}(s, \pi_i^*(s)) > Q_{\pi_i^*}^{r^*,\gamma_i^*}(s, \pi_j^*(s))$.*

The above assumption ensures that each expert policy is uniquely optimal in at least one of the states, i.e., no two expert policies are equally optimal in all states. Since we assume a common reward function, this assumption implies that $\gamma_i \neq \gamma_j$ for $i \neq j$. In other words, our assumption states that the discount factors lead to distinct policies. This assumption is necessary because, otherwise, we cannot distinguish the expert policies from the observed data and the reward function is non-identifiable. In Appendix C.1, we also show that in practice, Assumption 1 is rarely violated.

4.1 Naive Extension of LP-IRL Fails

We first note that naive extensions of the formulation in Eq. 3 to the MP-IRL setting return optimal solutions that violate Assumption 1. The naive solution would be to simply maximize the sum of the differences of the Q-functions over all experts and states, i.e.,

$$\max_{\Gamma\in[0,1]^K} \max_{r} \sum_{k\in[K]} \sum_{s\in\mathcal{S}} \min_{a\in A\setminus\pi_k^*(s)} \{Q^{r,\gamma_k}_{\pi_k^*}(s,\pi_k^*(s)) - Q^{r,\gamma_k}_{\pi_k^*}(s,a)\} - \lambda\|r\|_1 \tag{4a}$$

$$\text{subject to } Q^{r,\gamma_k}_{\pi_k^*}(s,\pi_k^*(s)) - Q^{r,\gamma_k}_{\pi_k^*}(s,a) \geq 0 \quad \forall s\in\mathcal{S},\ a\in\mathcal{A}\setminus\pi_k^*(s), k\in[K] \tag{4b}$$

$$|r| \preceq r_{\max} \tag{4c}$$

In Appendix E.1, we show that without further constraining the discount factors, there are situations where the above optimization problem assigns the same discount factors to multiple experts. This implies that, under the learned discount factors, some expert policies are not distinguishable from each other under any reward function, which violates Assumption 1. The expert policies should be sufficiently different in terms of Q-functions under the learned reward function and discount factors. We modify this naive optimization problem to address this problem below in Section 4.2.

4.2 Multi-planning horizon LP-IRL (MPLP-IRL)

To avoid undesirable global optimum in Appendix E.1, we need to incorporate the constraints in Assumption 1. This is challenging because optimization problems that include strict inequality constraints may not have attainable optimal solutions. We avoid incorporating strict inequalities by first selecting states where the expert policies are distinguishable and then maximizing the differences of Q-functions only on those states. The proposed MPLP-IRL problem is as follows:

$$\max_{\Gamma\in[0,1]^K} \max_{r} \min_{k\in[K],(s,a)\in\Omega_k} \{Q^{r,\gamma_k}_{\pi_k^*}(s,\pi_k^*(s)) - Q^{r,\gamma_k}_{\pi_k^*}(s,a)\} - \lambda\|r\|_1 \tag{5a}$$

$$\text{subject to } Q^{r,\gamma_k}_{\pi_k^*}(s,\pi_k^*(s)) - Q^{r,\gamma_k}_{\pi_k^*}(s,a) \geq 0 \quad \forall s\in\mathcal{S},\ a\in\mathcal{A}\setminus\pi_k^*(s), k\in[K] \tag{5b}$$

$$|r| \preceq r_{\max}, \tag{5c}$$

where Ω_k is a set of state-action tuples ensuring that there exists a feasible reward solution r such that, for any $(s,a)\in\Omega_k$, $Q^{\gamma_k}_{\pi_k^*}(s,\pi_k^*(s)) - Q^{\gamma_k}_{\pi_k^*}(s,a) > 0$. Ω_k is constructed by solving another LP problem which distinguishes the optimal action from other actions on as many states as possible.

Theorem 1. *For a set of arbitrary distinct discount factors, Γ ($\gamma_i \neq \gamma_j$ for $i \neq j$), let $\{z_k^*\}_{k=1}^K$ ($z_k^* \in \mathbb{R}^{|\mathcal{S}|\times(|\mathcal{A}-1|)}$) be the optimal solution to the following LP problem,*

$$\min_{r,\{z_k\}} \sum_{k=1}^{K} \mathbf{1}^\intercal z_k \tag{6a}$$

$$\text{subject to } Q^{r,\gamma_k}_{\pi_k^*}(s,\pi_k^*(s)) - Q^{r,\gamma_k}_{\pi_k^*}(s,a) + z_k(s,a) \geq 1\ \forall s\in\mathcal{S},\ a\in\mathcal{A}\setminus\pi_k^*(s), k\in[K] \tag{6b}$$

$$Q^{r,\gamma_k}_{\pi_k^*}(s,\pi_k^*(s)) - Q^{r,\gamma_k}_{\pi_k^*}(s,a) \geq 0\ \forall s\in\mathcal{S},\ a\in\mathcal{A}\setminus\pi_k^*(s), k\in[K] \tag{6c}$$

$$z_k \geq 0\ \forall k\in[K], \tag{6d}$$

where $z_k(s,a)$ denotes the element of vector z_k corresponding to the state-action tuple (s,a). There exists a feasible reward solution r that satisfies Assumption 1 if, for any pair of expert policies π_i^*, π_j^* ($i \neq j$), there exists a state s such that $z_i^*(s,\pi_j^*(s)) = 0$.

Proof (Sketch). The optimization problem in Eq. 6 is equivalent to

$$\max_{r} \sum_{k=1}^{K} \sum_{s \in \mathcal{S}} \sum_{a \in \{\mathcal{A} \setminus \pi_k^*(s)\}} \mathbb{1}_{\{Q_{\pi_k^*}^{r,\gamma_k}(s,\pi_k^*(s)) - Q_{\pi_k^*}^{r,\gamma_k}(s,a) > 0\}},$$

which maximizes the number of state-action pairs where the Q-function difference is positive (where the expert policy is strictly better). $z_k^*(s,a) = 1$ if the optimal solution cannot distinguish the optimal action, $\pi_k^*(s)$, from action a on state s. If under the optimal solution, there is a pair of expert policies π_i^*, π_j^* that cannot be distinguished on any states (i.e., $(s, \pi_i^*(s)) \notin \Omega_j$ for $\forall s \in \mathcal{S}$), we fail to find a reward solution that does not violate Assumption 1, and we claim the inner optimization problem in Eq. 5 to be infeasible under Γ. Otherwise, we have that $\Omega_k = \{(s,a) | z_k^*(s,a) = 0\}$. See full proof in Appendix A.1. □

Intuitively, our proposed optimization problem seeks a reward function that maximizes the minimal non-zero difference of Q-functions over states where expert policies are distinguishable, thus encouraging expert policies to be sufficiently different and ensuring the satisfaction of Assumption 1.

4.3 Inference for MPLP-IRL

While the objective function in 5a gives us the desired solutions, it is not convex with respect to the discount factors Γ. To solve for the global optima, we perform a bi-level optimization. Denote the lower-level objective function as

$$g(\Gamma, r) = \min_{k \in [K], (s,a) \in \Omega_k} \{Q_{\pi_k^*}^{r,\gamma_k}(s, \pi_k^*(s)) - Q_{\pi_k^*}^{r,\gamma_k}(s,a)\} - \lambda \|r\|_1.$$

We rewrite the optimization problem in Eq. 5 as,

$$\max_{\Gamma \in [0,1]^K} g^*(\Gamma), \quad \text{where} \quad g^*(\Gamma) = \max_{r} g(\Gamma, r) \text{ subject to constraints as in 5b, 5c.} \tag{7}$$

Given any $\Gamma \in [0,1]^K$, the lower-level objective function g can be solved analytically with LP. The upper-level optimization problem can be solved by performing a grid search over the space $[0,1]^K$. However, the computation complexity of grid search increases exponentially with respect to the total number of expert policies. To improve computation efficiency, we use Bayesian optimization (BO) techniques, which are suitable for nonconvex objective functions that are expensive to evaluate. See full algorithm details in Appendix B.

5 Algorithms for multi-planning horizon IRL: MCE-IRL

Although LP-IRL is easy to solve, we cannot apply it to domains with continuous state spaces as this will result in an infinite number of constraints. In this section, we switch our focus to MCE-IRL, which is a popular class of IRL algorithms for continuous domains. As is standard in this setting, we assume that the expert policies, $\tilde{\Pi}^* = \{\tilde{\pi}_k^*\}_{k=1}^K$, are solved with entropy-regularized MDPs and are optimizing a linear reward function,

$$r_{\theta^*}(s) = \theta^{*\intercal} \phi^s, \theta^*, \phi^s \in \mathbb{R}^{|S|}.$$

We first introduce notations used for MCE-IRL. For any state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$, given an initial state distribution ρ_0, the expected discounted state-action visitation count, and the expected discounted state visitation count of policy $\tilde{\pi}$ are defined as,

$$\mu^{\gamma}_{\tilde{\pi}}(s,a) = \mathbb{E}_{\tilde{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t \mathbb{1}_{\{S_t=s, A_t=a\}}\right], \text{ and } \mu^{\gamma}_{\tilde{\pi}}(s) = \mathbb{E}_{\tilde{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t \mathbb{1}_{\{S_t=s\}}\right] = \sum_a \mu^{\gamma}_{\tilde{\pi}}(s,a),$$

respectively. We further define the expected feature count of policy $\tilde{\pi}$ as

$$f^{\gamma}_{\tilde{\pi}} = \mathbb{E}_{\tilde{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t \phi(S_t = s)\right] \in \mathbb{R}^{|S|}.$$

Note that, the expected feature count can also be written as $f^{\gamma}_{\tilde{\pi}} = \sum_s \mu^{\gamma}_{\tilde{\pi}}(s)\phi^s$. Assuming that the discount factor γ is given, MCE-IRL solves the following constrained optimization problem:

$$\max_{\mu^{\gamma}_{\tilde{\pi}}(s,a)} \quad \mathcal{H}^{\gamma}_{\tilde{\pi}} = \mathbb{E}_{\tilde{\pi}}\left[\sum_{t=0}^{\infty} -\gamma^t \log \tilde{\pi}(A_t|S_t)\right] = \sum_{(s,a)} -\log\left(\frac{\mu^{\gamma}_{\tilde{\pi}}(s,a)}{\sum_a \mu^{\gamma}_{\tilde{\pi}}(s,a)}\right)\mu^{\gamma}_{\tilde{\pi}}(s,a) \tag{8a}$$

$$\text{subject to} \quad f^{\gamma}_{\tilde{\pi}} = f^{\gamma}_{\tilde{\pi}^*} \tag{8b}$$

$$\sum_a \mu^{\gamma}_{\tilde{\pi}}(s',a) = \rho_0(s') + \gamma \sum_s \sum_a T(s'|s,a)\mu^{\gamma}_{\tilde{\pi}}(s,a) \quad \forall s' \in \mathcal{S} \tag{8c}$$

$$\mu^{\gamma}_{\tilde{\pi}}(s,a) \geq 0 \quad \forall (s,a) \in (\mathcal{S} \times \mathcal{A}). \tag{8d}$$

MCE-IRL (Ziebart, 2010) identifies a reward function by the principle of maximum causal entropy (Eq. 8a) while matching the feature expectations between the expert and the learned policy (Eq. 8b). The Bellman flow constraints in Eqs. 8c-8d ensure that $\mu^{\gamma}_{\tilde{\pi}}(s,a)$ are state-action visitation count of a valid stochastic policy $\tilde{\pi}$ with discount factor γ. When the reward function is linear, Zeng et al. (2022) establish a strong duality between the MCE-IRL problem (Eq. 8) and its Lagrangian dual problem, which is equivalent to the following ML-IRL problem:

$$\max_{\theta} \quad \mathcal{L}(\theta) = \mathbb{E}_{\tilde{\pi}^*}\left[\sum_{t=0}^{\infty} \gamma^t \log \tilde{\pi}^{\theta}(A_t|S_t)\right] \tag{9a}$$

$$\text{subject to} \quad \tilde{\pi}^{\theta} = \arg\max_{\pi} \tilde{Q}^{\theta,\gamma}_{\pi}(s,a), \tag{9b}$$

where $\mathcal{L}(\theta)$ is the expectation of the discounted likelihood of expert trajectories under policy $\tilde{\pi}^{\theta}$. In practice, one often solves the above ML-IRL problem instead because it is more tractable.

5.1 Strong duality does not hold for multi-planning horizon MCE-IRL (MPMCE-IRL)

Although extending the formulation in Eq. 8 to the MP-IRL setting is straightforward, we cannot solve the Lagrangian dual or the ML-IRL problem as alternative optimization problems. In this section, we show that when the discount factors are unknown, strong duality does not hold between the MCE-IRL problem and its Lagrangian dual, which makes the inference less tractable. We extend the MCE-IRL formulation to the MP-IRL setting by maximizing the sum of causal entropy

while matching the expected feature counts for all expert policies:

$$\max_{\Gamma\in[0,1]^K} \max_{\{\mu^{\gamma_k}_{\tilde{\pi}_k}(s,a)\}} \quad \sum_{k=1}^{K} \mathcal{H}^{\gamma_k}_{\tilde{\pi}_k} = \sum_{k=1}^{K} \sum_{(s,a)} -\log\left(\frac{\mu^{\gamma_k}_{\tilde{\pi}_k}(s,a)}{\sum_a \mu^{\gamma_k}_{\tilde{\pi}_k}(s,a)}\right) \mu^{\gamma_k}_{\tilde{\pi}_k}(s,a) \tag{10a}$$

$$\text{subject to} \quad f^{\gamma_k}_{\tilde{\pi}_k} = f^{\gamma_k}_{\tilde{\pi}^*_k} \quad \forall k \in [K] \tag{10b}$$

$$\sum_a \mu^{\gamma_k}_{\tilde{\pi}_k}(s',a) = \rho_0(s') + \gamma_k \sum_s \sum_a T(s'|s,a)\mu^{\gamma_k}_{\tilde{\pi}_k}(s,a) \quad \forall s' \in \mathcal{S},\ k \in [K] \tag{10c}$$

$$\mu^{\gamma_k}_{\tilde{\pi}_k}(s,a) \geq 0 \quad \forall (s,a) \in (\mathcal{S}\times\mathcal{A}),\ k \in [K]. \tag{10d}$$

The multi-planning horizon ML-IRL (MPML-IRL) formulation is as follows:

$$\max_{\Gamma\in[0,1]^K} \max_{\theta} \quad \mathcal{L}(\theta,\Gamma) = \sum_{k=1}^{K} \mathbb{E}_{\tilde{\pi}^*_k}\left[\sum_{t=0}^{\infty} \gamma^t \log \tilde{\pi}^{\theta}_k(A_t|S_t)\right] \tag{11a}$$

$$\text{subject to} \quad \tilde{\pi}^{\theta}_k = \arg\max_{\pi} \tilde{Q}^{\theta,\gamma_k}_{\pi}(s,a) \quad \forall k \in [K] \tag{11b}$$

We first show that strong duality does not hold between the MPMCE-IRL problem and its Lagrangian dual problem.

Theorem 2. *Let $\mathcal{H}^*, \mathcal{G}^*$ be the optimal value of the MPMCE-IRL problem in Eq. 10 and its Lagrangian dual problem, respectively. Let $\mathcal{L}^*$ be the optimal value of the MPML-IRL problem in Eq. 11. Then, we have that $\mathcal{G}^* \geq \mathcal{H}^* \geq -\mathcal{L}^*$.*

Proof (Sketch). The MPMCE-IRL problem in Eq. 10 is nonconcave because the constraints in Eq. 10c are not affine. In Appendix A.2, we show that there are no saddle points for the Lagrangian dual function. Thus, strong duality does not hold. □

Theorem 2 also implies that, for MP-IRL problems, solving the MPMCE-IRL problem is not equivalent to solving the MPML-IRL problem. In Proposition 1, we further show that an optimal solution to the MPML-IRL problem may not reconstruct the expert policies. Thus, we cannot solve the MPML-IRL as an alternative even though it is more computationally convenient.

Proposition 1. *Let Γ, θ be an optimal solution to the MPML-IRL in Eq. 11 and $\tilde{\pi}^{\theta}_k$ be the optimal policy for the reward parameters θ and the discount factor γ_k. Then $\tilde{Q}^{\theta,\gamma_k}_{\tilde{\pi}^{\theta}_k} \geq \tilde{Q}^{\theta,\gamma_k}_{\tilde{\pi}^*_k}$ (i.e., $\tilde{\pi}^*_k$ may not be optimal under the optimal solution Γ, θ).*

Proof (Sketch). Let $\hat{\theta}$, $\hat{\Gamma}$ be any reward parameters and discount factors in the parameter space such that for all $k \in [K]$, $\tilde{\pi}^{\hat{\theta}}_k = \tilde{\pi}^*_k$. In Appendix A.3, we show that such a feasible solution that reconstructs expert policies Π^* may not be a critical point of $\mathcal{L}(\theta,\Gamma)$. □

5.2 Inference for multi-planning horizon MCE-IRL

Our theory in Section 5.1 above implies that for MP-IRL problems, we cannot solve the MPMCE-IRL problem by solving the MPML-IRL problem. To solve the MPMCE-RL problem, similar to

MPLP-IRL (Sec. 4.3), we propose a bi-level optimization in Eq. 10. Given a fixed set of discount factors, $\Gamma \in [0,1]^K$, let the lower-level optimization problem be

$$g^*(\Gamma) = \max_{\{\mu_{\tilde{\pi}_k}^{\gamma_k}(s,a)\}} \sum_{k=1} \mathcal{H}_{\tilde{\pi}_k}^{\gamma_k} \quad \text{subject to constraints as in Eq. 10b-10d.} \tag{12}$$

In practice, we observe that the above optimization problem is feasible for only a small set of discount factors (see the full discussion in Sec. 5.3). To ensure convergence to a feasible solution, for each lower-level optimization problem, we calculate the duality gap between the primal problem and its Lagrangian dual problem. For the upper-level optimization problem, we use BO to quickly identify discount factors with feasible reward functions. The full algorithm is given in Algorithm 2.

5.3 Feasibilty and Identifiability Analysis for Inference

Although the optimization problem in Eq. 12 is concave and can be solved with the Lagrangian duality method, it may not be feasible when the discount factors are misspecified. In this section, we study the effect of misspecified or unknown discount factors on the identifiability of the true reward function, which further determines the feasible solution space of the MPMCE-IRL problem.

Similar to Rolland et al. (2022), we first give matrix rank conditions that determine the set of reward functions that reconstruct optimal policies, $\tilde{\Pi}^*$.

Proposition 2. *Consider an MP-IRL problem defined in Sec. 3 with K expert policies, $\tilde{\Pi}^* = \{\tilde{\pi}_k^*\}_{k=1}^K$, solved with entropy-regularized RL. Assume that we are given a set of arbitrary discount factors, $\Gamma = \{\gamma_k\}_{k=1}^K$ ($\gamma_i \neq \gamma_j$ for $i \neq j$). T_{a_i} is the transition dynamics under action a_i. Let*

$$T_{\mathcal{A}} = \begin{bmatrix} T_{a_1} \\ \vdots \\ T_{a_{|\mathcal{A}|}} \end{bmatrix}, \quad \Phi = \begin{bmatrix} T_{\mathcal{A}} & -(\mathbf{1} \otimes I - \gamma_1 T_{\mathcal{A}}) & 0 & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ T_{\mathcal{A}} & 0 & \cdots & -(\mathbf{1} \otimes I - \gamma_k T_{\mathcal{A}}) \end{bmatrix}$$

$$\text{and} \quad b^\intercal = \lambda \left[\log \pi_1^*(a_1|\cdot), \cdots, \log \pi_1^*(a_{|\mathcal{A}|}|\cdot), \cdots, \log \pi_k^*(a_1|\cdot), \cdots, \log \pi_k^*(a_{|\mathcal{A}|}|\cdot)\right]^\intercal.$$

Then there does not exist any reward function that reconstructs optimal policies if and only if $rank(\Phi|b) > rank(\Phi)$. There exists a unique reward function (up to a constant factor) that reconstructs optimal policies if $rank(\Phi|b) = rank(\Phi) = K|\mathcal{S}| + |\mathcal{S}| - 1$.

Proof. The proof of Proposition 2, follows in a similar manner to that of Theorem 3 in Rolland et al. (2022). To find a reward function that reconstructs optimal policies, the linear system $\Phi x = b$, with $K|\mathcal{A}||\mathcal{S}|$ equations and $(K+1)|\mathcal{S}|$ variables, needs to be consistent.

Unlike the proof of Theorem 3 in Rolland et al. (2022), we do not assume the linear system is consistent because when the discount factors, Γ, are misspecified, the expert policies may not be optimal for the true reward function r^*. According to the Rouché-Capelli theorem, the above system of equations is inconsistent if $rank(\Phi|b) > rank(\Phi)$. The linear system has a unique solution (up to a constant factor) if $rank(\Phi|b) = rank(\Phi) = (K+1)|\mathcal{S}| - 1$. □

Proposition 2 implies that, without correctly specified discount factors, the reward set that reconstructs optimal policies may not include the true reward function, which emphasizes the importance of inferring the discount factors (and learning them correctly) rather than fixing them arbitrarily.

But is it possible to identify the true discount factors and the true reward function in practice? By the proof of Proposition 2, with more expert policies, the number of constraints of the linear system grows faster than the number of variables, which makes the set of feasible solutions of the linear system smaller. In Section 6.2, we empirically show that when the number of experts is sufficiently large, the reward set is non-empty for only a small set of discount factors. Thus, in highly heterogenous settings, both the reward function and the discount factors are more identifiable.

The rank conditions in Proposition 2 also determine if the optimization problem in Eq. 12 is feasible.

Corollary 1. *Given a set of discount factors $\Gamma = \{\gamma_k\}_{k=1}^K$, the optimization problem in Eq. 12 is feasible if and only if $rank(\Phi|b) = rank(\Phi)$ where Φ, b is defined in Proposition 2. Additionally, if the optimization problem is feasible, the optimal solution is achieved at $\mu_{\widetilde{\pi}_k}^{\gamma_k}(s,a) = \mu_{\widetilde{\pi}_k^*}^{\gamma_k}(s,a)$.*

Proof (Sketch). We show that, when the optimization problem is feasible, an optimal solution of the Lagrangian multipliers is a reward function that induces the same visitation variables as the expert policies, $\{\mu_{\widetilde{\pi}_k^*}^{\gamma_k}(s,a)\}$. Thus, for the primal problem to be feasible, it is sufficient to find a reward function that reconstructs the expert policies. See full details in Appendix A.4. □

6 Experiments and Results

In this section, we provide details of our designed domains and experiment setup. We first study the identifiability and generalizability of both the reward function and discount factors of each domain. We then study the properties of the learned reward function and the set of discount factors of MPLP-IRL and MPMCE-IRL algorithms, and investigate how well the learned reward functions generalize to similar tasks. Last, we demonstrate the fast convergence of our algorithms.

6.1 Domains

We test the MPLP-IRL (Algorithm 1) and MPMCE-IRL (Algorithm 2) on three domains: (1) the toy domain (Fig. 4a), in which the experts trade off between the probability of getting the reward and the reward magnitude; (2) the big-small domain (Ankile et al., 2023), in which experts choose between a small reward close by or a large one that is far away; and (3) the cliff domain, in which the experts trade off the risk of falling off the cliff with a large reward. For each domain, we provide expert demonstrations from 3 distinct expert policies. See full details in Appendix C.

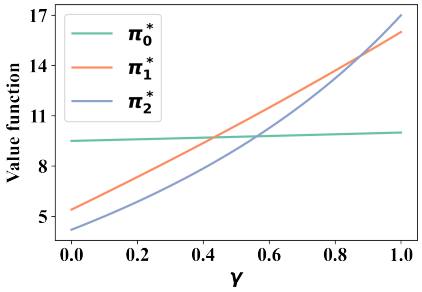
6.2 Identifiability and Generalizability Analysis

For each domain, we solve the linear system $\Phi x = b$ (Proposition 2) by performing a grid search over the space of $\Gamma \in [0,1]^K$ with an interval of 0.01. Across all designed domains, we observe that there exist reward functions that reconstruct expert policies only when $\Gamma = \Gamma^*$, which implies that the MPMCE-IRL problem (Eq. 10) has a small feasible solution space.

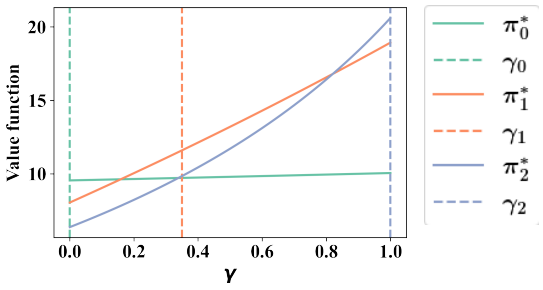
For the toy domain, if we only provide expert demonstrations from 2 expert policies, for all the given discount factors obtained form $[0,1]^2$, there always exists a unique reward function (up to a constant) that reconstructs the optimal policy, which may not include the true reward function. We further conduct a generalizability analysis of these reward functions with the full procedure described in Appendix D. In Appendix Fig. 11, we see that $\sim 40\%$ of these feasible reward functions do not

generalize well to new tasks with generalization errors (Eq. 27) larger than 0.5, which emphasizes the importance of learning the discount factors correctly.

6.3 Results



(a) The value function $V_{\pi_k^*}^{\gamma, r*}(s_0)$ of expert policies π_0^*, π_1^*, π_2^* under the true reward r^*.



(b) The value function $V_{\pi_k}^{\tilde{r}, \tilde{\gamma}}(s_0)$ of reconstructed optimal policies π_0, π_1, π_2 under the learned reward function of MPLP-IRL, $\tilde{r}$.

Figure 1: Plots of the value function of the initial state under (a) the true reward function r^*, (b) the learned reward function of MPLP-IRL, $\tilde{r}$: x, y-axes represent the discount factor $\gamma \in [0, 1]$ and the value function of expert policies or reconstructed optimal policies, respectively. Each color represents a different policy. The dashed lines in (b) represent the learned discount factors, $\tilde{\Gamma}$. We see that MPLP-IRL recovers the order of true discount factors.

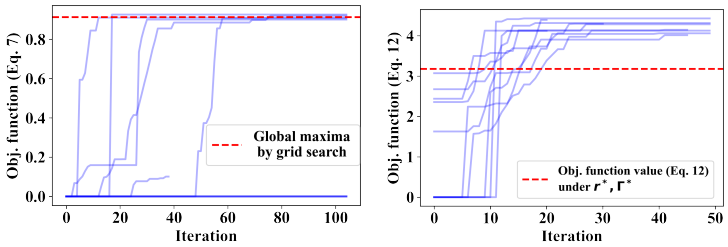
	Toy	Big Small	Cliff
MPLP-IRL	0.009 ± 0.040	0.039 ± 0.117	0.0 ± 0.0
MPMCE-IRL	0.21 ± 0.25	0.040 ± 0.065	0.0 ± 0.0

Figure 2: The table of the generalization error (Eq. 27) with one standard deviation of the learned reward function: each row and column represents a different algorithm and domain, respectively.

The learned discount factors recover the order of the true discount factors. From Fig. 1, we see that although the learned discount factor ($\tilde{\Gamma} \approx \{0,\ 0.35,\ 1\}$) does not exactly align with the ground truth, they follow the same order of the true discount factors ($\gamma_0^* \leq \gamma_1^* \leq \gamma_2^*$). This is also true for the big-small domain and the cliff domain (see full comparison of the true discount factors and the learned discount factors in Appendix Table 1). This property allows us to interpret the bias of each expert's goal—a small discount factor implies that the expert cares more about short-term outcomes and vice versa.

MPLP-IRL and MPMCE-IRL can appropriately capture key aspects of the task and the learned reward functions generalize well to similar new tasks. For the toy domain, we see that both MPLP-IRL and MPMCE-IRL learn a larger reward at state s_2 than at state s_1 (Appendix Fig. 4). When the discount factor is large, this reward structure encourages the agent to collect the large reward even in the face of more stochasticity. For the big-small domain, the learned reward functions have a small reward for the bottom left grid and a large reward for the bottom right grid (Appendix Fig. 5). For the cliff domain, the learned reward functions have large penalties for the top rows and a large reward for the upper right grid (Appendix Fig. 7). Thus,

all the learned reward functions have similar structures to the true reward function, which allows us to transfer the reward function to new RL tasks. The setup of the generalizability analysis is described in Appendix Section D. In Table 2, we see that our learned reward functions have good generalizability (all the generalization errors are below 0.04 except for that of the toy domain of MPMCE-IRL, which fails to learn $r(s_0)$ correctly (Appendix Fig. 4c)).



(a) MPLP-IRL: the toy domain (b) MPMCE-IRL: the toy domain

Figure 3: Trace plots of the best observed objective value of BO: x, y-axis represent the iteration and the best observed objective value, respectively. The red dashed line represents an approximate global maximum or the objective value under the ground truth.

MPLP-IRL and MPMCE-IRL converge quickly. For MPLP-IRL, we approximate the global optimum by performing a grid search over $[0, 1]^K$ with an interval of 0.01. For MPMCE-IRL, it is computationally heavy to solve the optimization problem in Eq. 12 10^6 times. We instead compare the best current objective value of BO to the objective value evaluated under the true reward function and discount factors. In Fig. 3, we see that on the toy domain, MPLP-IRL converges to the global maximum within 100 iterations while MPMCE-IRL converges within 50, reducing the computational burden by a factor of $\sim 10^4$ compared to grid search. See Appendix E.3 for details.

7 Discussion and Future Work

In this work, we study the MP-IRL setting where each expert is planning under different planning horizons but the same reward function. We provide theoretical and empirical evidence that highlights the importance of learning correct discount factors.

We develop two novel algorithms, MPLP-IRL and MPMCE-IRL, that learn the reward function and the discount factors jointly. Although MPLP-IRL is more computationally efficient (LP problems are faster to solve), it only applies to discrete domains. Additionally, MPLP-IRL does not guarantee identifying the true reward function and discount factors (in fact, for standard MDPs, identifying the set of discount factors for which the policy is optimal is nontrivial (Denis, 2019)). In contrast, we show that when there is a sufficiently large number of experts, MPMCE-IRL can identify both the reward function and discount factors. However, in practice, MPMCE-IRL has a larger feasible solution set than Corollary 1 suggests because we only require the algorithm to match the feature expectation in Eq. 10b within some threshold. Moreover, if the optimization problem in Eq. 12 is feasible for any $\Gamma \in [0, 1]^K$, MPMCE-IRL does not have attainable optimal solutions.

Interesting future work includes studying when the reward function is identifiable for an MP-IRL problem and extending to an IRL setting where both planning horizons and reward functions vary.

8 Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. IIS-2007076. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. JY and BEE were funded in part by Helmsley Trust grant AWD1006624, NIH NCI 5U2CCA233195, and CZI. BEE is a CIFAR Fellow in the Multiscale Human Program.

References

Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 1, 2004.

Lars L Ankile, Brian S Ham, Kevin Mao, Eura Shin, Siddharth Swaroop, Finale Doshi-Velez, and Weiwei Pan. Discovering user types: Mapping user traits by task-specific behaviors in reinforcement learning. *arXiv preprint arXiv:2307.08169*, 2023.

Monica Babes, Vukosi Marivate, Kaushik Subramanian, and Michael L Littman. Apprenticeship learning about multiple intentions. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 897–904, 2011.

Daniel Brown, Wonjoon Goo, Prabhat Nagarajan, and Scott Niekum. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. In *International conference on machine learning*, pp. 783–792. PMLR, 2019.

Haoyang Cao, Samuel Cohen, and Lukasz Szpruch. Identifiability in inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 34:12362–12373, 2021.

Arthur Charpentier, Romuald Elie, and Carl Remlinger. Reinforcement learning in economics and finance. *Computational Economics*, pp. 1–38, 2021.

Nicholas Denis. Issues concerning realizability of blackwell optimal policies in reinforcement learning. *arXiv preprint arXiv:1905.08293*, 2019.

Brad Edward Dicianno, Geoffrey Henderson, and Bambang Parmanto. Design of mobile health tools to promote goal achievement in self-management tasks. *JMIR mHealth and uHealth*, 5(7): e7335, 2017.

Adam Gleave and Oliver Habryka. Multi-task maximum entropy inverse reinforcement learning. *arXiv preprint arXiv:1805.08882*, 2018.

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *International conference on machine learning*, pp. 1352–1361. PMLR, 2017.

Zhuo-Jun Jin, Hui Qian, and Miao-Liang Zhu. Gaussian processes in inverse reinforcement learning. In *2010 International Conference on Machine Learning and Cybernetics*, volume 1, pp. 225–230. IEEE, 2010.

Siddhartha G Kapnadak, Steve E Herndon, Suzanne M Burns, Y Michael Shim, Kyle Enfield, Cynthia Brown, Jonathon D Truwit, and Ajeet G Vinayak. Clinical outcomes associated with high, intermediate, and low rates of failed extubation in an intensive care unit. *Journal of Critical Care*, 30(3):449–454, 2015.

Jorge Mendez, Shashank Shivkumar, and Eric Eaton. Lifelong inverse reinforcement learning. *Advances in neural information processing systems*, 31, 2018.

Andrew Y Ng, Stuart Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, pp. 2, 2000.

Deepak Ramachandran and Eyal Amir. Bayesian inverse reinforcement learning. In *IJCAI*, volume 7, pp. 2586–2591, 2007.

Elsa Riachi, Muhammad Mamdani, Michael Fralick, and Frank Rudzicz. Challenges for reinforcement learning in healthcare. *arXiv preprint arXiv:2103.05612*, 2021.

Paul Rolland, Luca Viano, Norman Schürhoff, Boris Nikolov, and Volkan Cevher. Identifiability and generalizability from multiple experts in inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 35:550–564, 2022.

Kyriacos Shiarlis, Joao Messias, and Shimon Whiteson. Inverse reinforcement learning from failure. 2016.

Umar Syed, Michael Bowling, and Robert E Schapire. Apprenticeship learning using linear programming. In *Proceedings of the 25th international conference on Machine learning*, pp. 1032–1039, 2008.

Lantao Yu, Tianhe Yu, Chelsea Finn, and Stefano Ermon. Meta-inverse reinforcement learning with probabilistic context variables. *Advances in neural information processing systems*, 32, 2019.

Siliang Zeng, Chenliang Li, Alfredo Garcia, and Mingyi Hong. Maximum-likelihood inverse reinforcement learning with finite-time guarantees. *Advances in Neural Information Processing Systems*, 35:10122–10135, 2022.

Zhengyuan Zhou, Michael Bloem, and Nicholas Bambos. Infinite time horizon maximum causal entropy inverse reinforcement learning. *IEEE Transactions on Automatic Control*, 63(9):2787–2802, 2017.

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

Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pp. 1433–1438. Chicago, IL, USA, 2008.

A Theorems

A.1 Proof of Theorem 1

In this section, we prove that the pre-computed state-action tuple Ω_k allows us to find a feasible reward function that satisfies the following assumptions:

Assumption 1. *For any two distinct expert policies $\pi_i^*, \pi_j^* \in \Pi^*,\ i \neq j$, optimized under γ_i^*, γ_j^* ($\gamma_i^* \neq \gamma_j^*$), respectively, there is at least one state $s \in \mathcal{S}$ such that $Q_{\pi_i^*}^{r^*,\gamma_i^*}(s, \pi_i^*(s)) > Q_{\pi_i^*}^{r^*,\gamma_i^*}(s, \pi_j^*(s))$.*

Theorem 1. *(Restated) For a set of arbitrary distinct discount factors, Γ ($\gamma_i \neq \gamma_j$ for $i \neq j$), let $\{z_k^*\}_{k=1}^K$ ($z_k^* \in \mathbb{R}^{|\mathcal{S}|\times(|\mathcal{A}-1|)}$) be the optimal solution to the following problem,*

$$\min_{r,z_k} \sum_{k=1}^{K} \mathbf{1}^\intercal z_k \tag{13a}$$

$$\text{subject to } Q_{\pi_k^*}^{r,\gamma_k}(s, \pi_k^*(s)) - Q_{\pi_k^*}^{r,\gamma_k}(s,a) + z_k(s,a) \geq 1\ \forall s \in \mathcal{S},\ a \in \mathcal{A}\backslash\pi_k^*(s), k \in [K] \tag{13b}$$

$$Q_{\pi_k^*}^{r,\gamma_k}(s, \pi_k^*(s)) - Q_{\pi_k^*}^{r,\gamma_k}(s,a) \geq 0\ \forall s \in \mathcal{S},\ a \in \mathcal{A}\backslash\pi_k^*(s), k \in [K] \tag{13c}$$

$$z_k \geq 0\ \forall k \in [K], \tag{13d}$$

where $z_k(s,a)$ denotes the element of vector z_k corresponding to the state-action tuple (s,a). There exists a feasible reward solution r that satisfies Assumption 1 if, for any pair of policies $\pi_i^*,\ \pi_j^*$ ($i \neq j$), there exists a state s such that $z_i^*(s, \pi_j^*(s)) = 0$.

Proof. Given an optimization problem in the following:

$$\max_{r} \sum_{(s,a)} \mathbb{1}_{\{Q_{\pi_k^*}^{r,\gamma_k}(s,\pi_k^*(s)) - Q_{\pi_k^*}^{r,\gamma_k}(s,a)\}\geq 0} \tag{14a}$$

$$\text{subject to } Q_{\pi_k^*}^{r,\gamma_k}(s, \pi_k^*(s)) - Q_{\pi_k^*}^{r,\gamma_k}(s,a) \geq 0 \quad \forall s \in \mathcal{S},\ a \in \mathcal{A}\backslash\pi_k^*(s), k \in [K] \tag{14b}$$

Let $\hat{r}$ be an optimal solution to the optimization problem in Eq. 14. It is easy to see that if, for any pair of policies $\pi_i^*,\ \pi_j^*$ ($i \neq j$), there exists a state s such that $Q_{\pi_i^*}^{\hat{r},\gamma_i}(s, \pi_i^*(s)) - Q_{\pi_i^*}^{\hat{r},\gamma_i}(s, \pi_j^*(s)) > 0$, then Assumption 1 is satisfied.

We start the proof by showing that an optimal solution to optimization problem 13 is also an optimal solution to optimization problem 14.

With Bellman Equations, the difference of the Q-functions can be written as,

$$Q_{\pi_k^*}^{r,\gamma_k}(s, \pi_k^*(s)) - Q_{\pi_k^*}^{r,\gamma_k}(s,a) = (T(\cdot|s, \pi_k^*(s)) - T(\cdot|s,a))(I - \gamma_k T^{\pi_k^*})^{-1} r.$$

Let $W \in \mathbb{R}^{(|\mathcal{S}|\times(|\mathcal{A}|-1)\times K)\times|\mathcal{S}|}$ be a matrix where each row

$$w_{s,a,k}^\intercal = (T(\cdot|s, \pi_k^*(s)) - T(\cdot|s,a))(I - \gamma_k T^{\pi_k^*})^{-1},\ \forall s \in \mathcal{S}, a \in \mathcal{A} \setminus \pi_k^*(s), k \in [K].$$

The optimization problem in Eq. 13 then can be rewritten as

$$\max_{r} |Wr|_0 \tag{15a}$$

$$\text{subject to } Wr \geq 0 \tag{15b}$$

The optimization problem in Eq. 14 can be rewritten as

$$\min_{r,z} \mathbf{1}^\intercal z \tag{16a}$$
$$\text{subject to } Wr + z \geq 1 \tag{16b}$$
$$Wr \geq 0 \tag{16c}$$
$$z \geq 0 \tag{16d}$$

Let r be a reward function that satisfies Constraint 16c. Then for any constant $c > 0$, cr also satisfies Constraint 16c. Now, let c be a constant such that any positive element in cWr is larger than 1. Denote the i-th element of any vector x by x_i. An optimal solution, z^* and $\hat{r}$. of optimization problem in 16 has the following form:

$$z_i^* = 1 \text{ if } (cW\hat{r})_i = 0,\ z_i^* = 0 \text{ if } (cW\hat{r})_i \geq 1. \tag{17}$$

We show that $\hat{r}$ is also an optimal solution to optimization problem 15 with proof by contradiction.

Let $z^*, \hat{r}$ be an optimal solution to optimization problem 16, but $\hat{r}$ is not an optimal solution to optimization problem 15. Then there exists another $\tilde{r}$ such that $W\tilde{r} \geq 0$ and $|W\tilde{r}|_0 > |W\hat{r}|_0$. Now, let c be a constant such that any positive element in $W(c\tilde{r})$ is larger than or equal 1 and construct a vector $\tilde{z}$ according to Eq. 17. Because $|cW\tilde{r}|_0 = |W\tilde{r}|_0 > |W\hat{r}|_0$, $W\tilde{r} \geq 0$ and $W\hat{r} \geq 0$, $\tilde{z}$ will have more zero elements than z^*. Thus, $\mathbf{1}^\intercal \tilde{z} < \mathbf{1}^\intercal z^*$, which contradicts the assumption that z^* is optimal.

Additionally, with the definition of z^* in Eq. 17, we can see that

$$w_{s,a,k}^\intercal = Q_{\pi_k^*}^{\hat{r},\gamma_k}(s, \pi_k^*(s)) - Q_{\pi_k^*}^{\hat{r},\gamma_k}(s,a) > 0$$

if and only if $z_k^*(s,a) = 0$. Thus, there exists a feasible reward solution r that satisfies Assumption 1 if, for any pair of policies $\pi_i^*,\ \pi_j^*$, there exists a state s such that $z_i^*(s, \pi_j^*(s)) = 0$. □

A.2 Proof of Theorem 2

In this section, we show that under our targetted IRL setting in 3, strong duality does not hold between the MCE-IRL problem and its Lagrangian dual. Furthermore, solving the MCE-IRL problem is not equivalent to solving the ML-IRL problem.

Theorem 2. *(Restated) Let the multi-planning horizon MCE-IRL problem be*

$$\max_{\Gamma \in [0,1]^K} \max_{\{\mu_{\tilde{\pi}_k}^{\gamma_k}(s,a)\}} \sum_{k=1}^K \mathcal{H}_{\tilde{\pi}_k}^{\gamma_k} = \sum_{k=1}^K \sum_{(s,a)} -\log\left(\frac{\mu_{\tilde{\pi}_k}^{\gamma_k}(s,a)}{\sum_a \mu_{\tilde{\pi}_k}^{\gamma_k}(s,a)}\right) \mu_{\tilde{\pi}_k}^{\gamma_k}(s,a) \tag{18a}$$
$$\textit{subject to} \quad f_{\tilde{\pi}_k}^{\gamma_k} = f_{\tilde{\pi}_k^*}^{\gamma_k} \quad \forall k \in [K] \tag{18b}$$
$$\sum_a \mu_{\tilde{\pi}_k}^{\gamma_k}(s',a) = \rho_0(s') + \gamma_k \sum_s \sum_a T(s'|s,a)\mu_{\tilde{\pi}_k}^{\gamma_k}(s,a) \quad \forall s' \in \mathcal{S},\ k \in [K] \tag{18c}$$
$$\mu_{\tilde{\pi}_k}^{\gamma_k}(s,a) \geq 0 \quad \forall (s,a) \in (\mathcal{S} \times \mathcal{A}),\ k \in [K]. \tag{18d}$$

Let the multi-planning horizon ML-IRL problem be

$$\max_{\Gamma\in[0,1]^K} \max_{\theta} \quad \mathcal{L}(\theta,\Gamma) = \sum_{k=1}^{K} \mathbb{E}_{\tilde{\pi}_k^*}\left[\sum_{t=0}^{\infty} \gamma^t \log \tilde{\pi}_k^{\theta}(A_t|S_t)\right] \tag{19a}$$

$$= \sum_{k=1}^{K} \sum_{(s,a)} \log\left(\frac{\mu_{\tilde{\pi}_k}^{\gamma_k}(s,a)}{\sum_a \mu_{\tilde{\pi}_k}^{\gamma_k}(s,a)}\right) \mu_{\tilde{\pi}_k^*}^{\gamma}(s,a) \tag{19b}$$

$$\text{subject to} \quad \tilde{\pi}_k^{\theta} = \arg\max_{\pi} \tilde{Q}_{\pi}^{\theta,\gamma_k}(s,a) \quad \forall k \in [K] \tag{19c}$$

Let $\mathcal{H}^*, \mathcal{G}^*$ be the optimal value of the multi-planning horizon MCE-IRL problem in Eq. 10 and its Lagrangian dual, respectively. Let $\mathcal{L}^*$ be the optimal value of the ML-IRL problem in Eq. 11. Then, we have $\mathcal{G}^* \geq \mathcal{H}^* \geq -\mathcal{L}^*$.

Proof.
Part I – Proof of $\mathcal{G}^* \geq \mathcal{H}^*$:

The first inequality holds $\mathcal{G}^* \geq \mathcal{H}^*$ by the weak duality. We further show that when the expert policies are stochastic, the optimal solutions to the primal problem are not critical points of the Lagrangian dual function. Thus, strong duality may not always hold.

The Lagrangian dual problem of the primal problem in Eq. 18 is

$$\min_{\Theta} \max_{\Gamma, M} \mathcal{G}(\Theta,\Gamma,M) \quad \text{,where}$$

$$\mathcal{G}(\Theta,\Gamma,M) = \sum_{k=1}^{K} \sum_{(s,a)} -\log\left(\frac{\mu_{\tilde{\pi}_k}^{\gamma_k}(s,a)}{\sum_a \mu_{\tilde{\pi}_k}^{\gamma_k}(s,a)}\right) \mu_{\tilde{\pi}_k}^{\gamma_k}(s,a) + \theta_k^{\intercal}(f_{\tilde{\pi}_k}^{\gamma_k} - f_{\tilde{\pi}_k^*}^{\gamma_k})$$

$$+ \sum_{s',k} x_{s',k}\left(\mu_{\tilde{\pi}_k}^{\gamma_k}(s') - \rho_0(s') - \gamma_k \sum_s \sum_a T(s'|s,a)\mu_{\tilde{\pi}_k}^{\gamma_k}(s,a)\right)$$

$$M = \{\mu_{\tilde{\pi}_k}^{\gamma_k}(s,a)\}, \Theta = \{\theta_k, x_{s,k}\}, \Gamma = \{\gamma_k\}_{k=1}^{K}.$$

We treat the constraints in Eq. 18d as implicit because the objective function in Eq. 18a is not defined under non-positive $\mu_{\tilde{\pi}_k}^{\gamma_k}(s,a)$.

For any Lagrangian multipliers Θ, we find the critical points of $\mathcal{G}(\Theta,\Gamma,M)$ by setting the gradient to zero.

$$\frac{\partial \mathcal{G}}{\partial \mu_{\tilde{\pi}_k}^{\gamma_k}(s,a)} = -\log\left(\frac{\mu_{\tilde{\pi}_k}^{\gamma_k}(s,a)}{\sum_a \mu_{\tilde{\pi}_k}^{\gamma_k}(s,a)}\right) + \theta_k^{\intercal}\phi^s + x_{s,k} - \gamma_k \sum_{s'} x_{s',k} T(s'|s,a) = 0$$

$$\log\left(\frac{\mu_{\tilde{\pi}_k}^{\gamma_k}(s,a)}{\sum_a \mu_{\tilde{\pi}_k}^{\gamma_k}(s,a)}\right) = \theta_k^{\intercal}\phi^s + x_{s,k} - \gamma_k \sum_{s'} x_{s',k} T(s'|s,a) \tag{20}$$

By Theorem 1 in Cao et al. (2021), Eq. 20 is satisfied by the state-action visitation count of the optimal policy under the reward parameters θ_k and discount factor γ_k with $-x_{s,k}$ being its value function. Under our targetted IRL setting, the state-action visitation counts need to be induced by a global reward function. Thus, we have $\theta_1 = \cdots = \theta_k = \theta$.

Denote the optimal policy under the reward parameters θ and discount factor γ_k as $\tilde{\pi}_k^\theta$ and plug it into $\mathcal{G}$, the Lagrangian dual problem becomes

$$\min_{\theta} \max_{\Gamma} \; f(\theta, \gamma) = \sum_{k=1}^{K} \sum_{(s,a)} -\log \left(\frac{\mu^{\gamma_k}_{\tilde{\pi}_k^\theta}(s,a)}{\sum_a \mu^{\gamma_k}_{\tilde{\pi}_k^\theta}(s,a)} \right) \mu^{\gamma_k}_{\tilde{\pi}_k^\theta}(s,a) + \theta_k^\intercal (f^{\gamma_k}_{\tilde{\pi}_k^\theta} - f^{\gamma_k}_{\tilde{\pi}_k^*}) \tag{21a}$$

$$\text{subject to} \quad \tilde{\pi}_k^\theta = \arg\max_{\pi} \tilde{Q}_\pi^{\theta,\gamma_k}(s,a) \quad \forall k \in [K] \tag{21b}$$

Both the primal and the dual problems have an optimal solution because both of them feasible and bounded. Strong duality holds iff there exists a saddle point for function $f(\theta, \Gamma)$. That is, there exists some $\tilde{\theta}, \tilde{\Gamma}$ such that

$$\forall \theta, \Gamma \in [0,1]^K, \quad f(\tilde{\theta}, \Gamma) \leq f(\tilde{\theta}, \tilde{\Gamma}) \leq f(\theta, \tilde{\Gamma}).$$

Furthermore, if strong duality holds, $\tilde{\Gamma}$ is a global maximum point of the primal problem and $\tilde{\theta}$ is a global minimum point of the Lagrangian dual. By Corollary 1, we know that $\tilde{\Gamma}$ is a global maximum point iff there exist feasible reward functions that reconstruct the expert policies. Corollary 1 also tells us that $\tilde{\theta} \in \arg\min f(\theta, \tilde{\Gamma})$ are the reward parameters such that $\mu^{\gamma_k}_{\tilde{\pi}_k^{\tilde{\theta}}} = \mu^{\gamma_k}_{\tilde{\pi}_k^*}$.

We now rewrite the Lagrangian dual function f as follows:

$$f(\theta, \Gamma) = \sum_{k=1}^{K} \mathcal{H}^{\gamma_k}_{\tilde{\pi}_k^\theta} + \theta_k^\intercal (f^{\gamma_k}_{\tilde{\pi}_k^\theta} - f^{\gamma_k}_{\tilde{\pi}_k^*}) = \sum_{k=1}^{K} \tilde{V}^{\theta,\gamma}_{\tilde{\pi}_k^\theta} - \tilde{V}^{\theta,\gamma}_{\tilde{\pi}_k^*} + \mathcal{H}^{\gamma_k}_{\tilde{\pi}_k^*}$$

When $\mu^{\gamma_k}_{\tilde{\pi}_k^{\tilde{\theta}}} = \mu^{\gamma_k}_{\tilde{\pi}_k^*}$, we have that $f(\tilde{\theta}, \tilde{\Gamma}) = \mathcal{H}^{\tilde{\gamma}_k}_{\tilde{\pi}_k^*}$. However, for such $\tilde{\theta}$, we can find another $\gamma_k > \tilde{\gamma}_k$ for some $k \in [K]$ such that

$$f(\tilde{\theta}, \Gamma) = \sum_{k=1}^{K} \tilde{V}^{\tilde{\theta},\gamma}_{\tilde{\pi}_k^{\tilde{\theta}}} - \tilde{V}^{\tilde{\theta},\gamma}_{\tilde{\pi}_k^*} + \mathcal{H}^{\gamma_k}_{\tilde{\pi}_k^*} \overset{(i)}{\geq} 0 + \mathcal{H}^{\gamma_k}_{\tilde{\pi}_k^*} \overset{(ii)}{>} \mathcal{H}^{\tilde{\gamma}_k}_{\tilde{\pi}_k^*}$$

where (i) follows from the fact that $\tilde{\pi}_k^{\tilde{\theta}}$ is optimal for reward parameter $\tilde{\theta}$ and discount fact γ_k and (ii) follows from the fact that the entropy is monotonically increasing on $\gamma_k \in [0,1]$. Thus, strong duality does not hold.

Part II – Proof of $\mathcal{H}^* \geq -\mathcal{L}^*$:

By negating the objective function in Eq. 19a, we have that

$$\min_{\Gamma \in [0,1]^K} \min_{\theta} \quad -\mathcal{L}(\theta, \Gamma) = \sum_{k=1}^{K} \sum_{(s,a)} \log \left(\frac{\mu^{\gamma_k}_{\tilde{\pi}_k^\theta}(s,a)}{\sum_a \mu^{\gamma_k}_{\tilde{\pi}_k^\theta}(s,a)} \right) \mu^{\gamma}_{\tilde{\pi}_k^*}(s,a)$$

$$\text{subject to} \quad \tilde{\pi}_k^\theta = \arg\max_{\pi} \tilde{Q}_\pi^{\theta,\gamma_k}(s,a) \quad \forall k \in [K]$$

Because an optimal solution to the multi-planing horizon MCE-IRL problem in Eq. 18 is a feasible solution to the multi-planing horizon ML-IRL problem in Eq. 19. Thus, $\mathcal{H}^* \geq \mathcal{H} \geq -\mathcal{L}^*$. □

A.3 Proof of Proposition 1

In this section, we show that an optimal solution to the naive multi-planning horizon ML-IRL problem may not reconstruct the expert policies. Thus, we cannot use this formulation to learn a feasible reward function.

Proposition 3. *(Restated) Let Γ, θ be an optimal solution to the ML-IRL in Eq. 11 and $\tilde{\pi}_k^\theta$ be the optimal policy for the reward parameters θ and the discount factor γ_k. Then $\tilde{Q}^{\theta,\gamma_k}_{\tilde{\pi}_k^\theta} \geq \tilde{Q}^{\theta,\gamma_k}_{\tilde{\pi}_k^*}$ (i.e., $\tilde{\pi}_k^*$ may not be optimal under the optimal solution Γ, θ).*

Proof. Given any reward parameters θ, the optimal value function and Q-value function satisfy,

$$\tilde{V}^{\theta,\gamma}_{\tilde{\pi}^\theta}(s) = \lambda \log \sum_a exp(\tilde{Q}^{\theta,\gamma}_{\tilde{\pi}^\theta}(s,a)/\lambda) \tag{22}$$

$$\pi^\theta(a|s) = exp\left(\frac{\tilde{Q}^{\theta,\gamma}_{\tilde{\pi}^\theta}(s,a) - \tilde{V}^{\theta,\gamma}_{\tilde{\pi}^\theta}(s))}{\lambda}\right) \tag{23}$$

Given the above equations, we can rewrite the expectation of the discount likelihood of expert trajectories as,

$$\begin{aligned}
\mathcal{L}(\theta,\Gamma) &= \sum_{k=1}^K \mathbb{E}_{(S_t,A_t)\sim\tilde{\pi}_k^*}\left[\sum_{t=0}^\infty \gamma_k^t \log \tilde{\pi}_k^\theta(A_t|S_t)\right] \\
&\overset{(i)}{=} \frac{1}{\lambda}\sum_{k=1}^K \mathbb{E}_{(S_t,A_t)\sim\tilde{\pi}_k^*}\left[\sum_{t=0}^\infty \gamma_k^t \left(\tilde{Q}^{\theta,\gamma_k}_{\tilde{\pi}_k^\theta}(S_t,A_t) - \tilde{V}^{\theta,\gamma_k}_{\tilde{\pi}_k^\theta}(S_t))\right)\right] \\
&= \frac{1}{\lambda}\sum_{k=1}^K \mathbb{E}_{(S_t,A_t)\sim\tilde{\pi}_k^*}\left[\sum_{t=0}^\infty \gamma_k^t \left(r(S_{t+1}) + \gamma_k \tilde{V}^{\theta,\gamma_k}_{\tilde{\pi}_k^\theta}(S_{t+1}) - \tilde{V}^{\theta,\gamma_k}_{\tilde{\pi}_k^\theta}(S_t)\right)\right] \\
&= \frac{1}{\lambda}\sum_{k=1}^K \left(\mathbb{E}_{(S_t,A_t)\sim\tilde{\pi}_k^*}\left[\sum_{t=0}^\infty \gamma_k^t r(S_{t+1})\right] + \mathbb{E}_{(S_t,A_t)\sim\tilde{\pi}_k^*}\left[\sum_{t=1}^\infty \gamma_k^t \tilde{V}^{\theta,\gamma_k}_{\tilde{\pi}_k^\theta}(S_{t+1})\right] - \mathbb{E}_{(S_t,A_t)\sim\tilde{\pi}_k^*}\left[\sum_{t=0}^\infty \gamma_k^t \tilde{V}^{\theta,\gamma_k}_{\tilde{\pi}_k^\theta}(S_t)\right]\right) \\
&= \frac{1}{\lambda}\sum_{k=1}^K \left(\mathbb{E}_{(S_t,A_t)\sim\tilde{\pi}_k^*}\left[\sum_{t=0}^\infty \gamma_k^t r(S_{t+1})\right] - \mathbb{E}_{S_0\sim\rho}\left[\sum_{t=0}^\infty \gamma_k^t \tilde{V}^{\theta,\gamma_k}_{\tilde{\pi}_k^\theta}(S_0)\right]\right) \\
&\overset{(ii)}{=} \sum_{k=1}^K \left(\frac{1}{\lambda}\mathbb{E}_{(S_t,A_t)\sim\tilde{\pi}_k^*}\left[\sum_{t=0}^\infty \gamma_k^t r(S_{t+1})\right] - \mathbb{E}_{S_0\sim\rho}\left[\log \sum_a exp\left(\frac{\tilde{Q}^{\theta,\gamma_k}_{\tilde{\pi}_k^\theta}(S_0,A_0)}{\lambda}\right)\right]\right),
\end{aligned}$$

where (i) and (ii) follow from Eq. 22 and 23, respectively. Let $\hat{\theta}$, $\hat{\Gamma}$ be any reward parameters and discount factors in the parameter space such that for all $k \in [K]$, $\tilde{\pi}_k^{\hat{\theta}} = \tilde{\pi}_k^*$. We further map γ_k to an unconstrained variable space: $\delta_k = \text{logit}(\gamma_k)$ and $\Delta = \{\delta_k\}_{k=1}^K$.

We now show that $\frac{\partial \mathcal{L}}{\partial \delta_k}(\hat{\delta}_k) \leq 0$. Calculating the gradient of $\mathcal{L}(\theta, \Gamma)$ with respect to δ_k gives us the following:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \delta_k} =& \frac{1}{\lambda} \frac{\partial}{\partial \delta_k} \mathbb{E}_{(S_t,A_t)\sim\tilde{\pi}_k^*} \left[\sum_{t=0}^{\infty} \frac{\partial \gamma_k^t}{\partial \delta_k} r(S_{t+1}) \right] \\
& - \frac{1}{\lambda} \mathbb{E}_{S_0\sim\rho} \left[\sum_a exp\left(\frac{\tilde{Q}^{\theta,\gamma_k}_{\tilde{\pi}_k^\theta}(S_0,A_0)}{\lambda} - \log \sum_a exp\left(\frac{\tilde{Q}^{\theta,\gamma_k}_{\tilde{\pi}_k^\theta}(S_0,A_0)}{\lambda} \right) \right) \frac{\partial \tilde{Q}^{\theta,\gamma_k}_{\tilde{\pi}_k^\theta}(S_0,A_0)}{\partial \delta_k} \right] \\
\overset{(i)}{=} & \frac{1}{\lambda} \left(\mathbb{E}_{(S_t,A_t)\sim\tilde{\pi}_k^*} \left[\sum_{t=0}^{\infty} \frac{\partial \gamma_k^t}{\partial \delta_k} r(S_{t+1}) \right] - \mathbb{E}_{S_0\sim\rho} \left[\sum_a \tilde{\pi}_k^\theta(A_0|S_0) \frac{\partial \tilde{Q}^{\theta,\gamma_k}_{\tilde{\pi}_k^\theta}(S_0,A_0)}{\partial \delta_k} \right] \right), \qquad (24)
\end{aligned}$$

where (i) follows from Eq. 22 and 23.

The gradient of $\tilde{Q}^{\theta,\gamma_k}_{\tilde{\pi}_k^\theta}(S_0,A_0)$ can be further expanded as,

$$\frac{\partial \tilde{Q}^{\theta,\gamma_k}_{\tilde{\pi}_k^\theta}(S_0,A_0)}{\partial \delta_k} = \mathbb{E}_{(S_t,A_t)\sim\tilde{\pi}_k^\theta} \left[\sum_{t=0}^{\infty} \frac{\partial \gamma_k^t}{\partial \delta_k} r(S_{t+1}) \right] + \lambda \frac{\partial}{\partial \delta_k} \mathbb{E}_{(S_t,A_t)\sim\tilde{\pi}_k^\theta} \left[\sum_{t=0}^{\infty} \gamma^t \tilde{\pi}_k^\theta(\cdot|S_t) \log \tilde{\pi}_k^\theta(\cdot|S_t) \right].$$

Plugging the gradient of $\tilde{Q}^{\theta,\gamma_k}_{\tilde{\pi}_k^\theta}(S_0,A_0)$ into Eq. 24 and evaluating the gradient at $\tilde{\pi}_k^\theta = \tilde{\pi}_k^*$ gives us,

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \delta_k}(\hat{\delta}_k) \propto & \cancel{\mathbb{E}_{(S_t,A_t)\sim\tilde{\pi}_k^*} \left[\sum_{t=0}^{\infty} \frac{\partial \gamma_k^t}{\partial \delta_k} r(S_{t+1}) \right]} - \cancel{\mathbb{E}_{(S_t,A_t)\sim\tilde{\pi}_k^*} \left[\sum_{t=0}^{\infty} \frac{\partial \gamma_k^t}{\partial \delta_k} r(S_{t+1}) \right]} \\
& - \lambda \frac{\partial}{\partial \delta_k} \mathbb{E}_{(S_t,A_t)\sim\tilde{\pi}_k^\theta} \left[\sum_{t=0}^{\infty} \gamma_k^t \tilde{\pi}_k^\theta(\cdot|S_t) \log \tilde{\pi}_k^\theta(\cdot|S_t) \right] \\
\overset{(i)}{=} & -\lambda \mathbb{E}_{(S_t,A_t)\sim\tilde{\pi}_k^*} \left[\sum_{t=0}^{\infty} t \hat{\gamma}_k^{t-1} \hat{\gamma}_k (1-\hat{\gamma}_k) \tilde{\pi}_k^*(\cdot|S_t) \log \tilde{\pi}_k^*(\cdot|S_t) \right] \\
= & \sum_{t=1}^{\infty} -\lambda (1-\hat{\gamma}_k) \mathbb{E}_{(S_t,A_t)\sim\tilde{\pi}_k^*} \left[\sum_{t'=t}^{\infty} \hat{\gamma}_k^t \tilde{\pi}_k^*(\cdot|S_t) \log \tilde{\pi}_k^*(\cdot|S_t) \right] \geq 0,
\end{aligned}$$

where (i) follows from the fact $\frac{\partial \gamma_k}{\partial \delta_k} = \frac{\partial \sigma(\delta_k)}{\partial \delta_k} = \sigma(\delta_k)(1-\sigma(\delta_k)) = \gamma_k(1-\gamma_k)$. We see that when $\hat{\gamma}_k \in (0,1)$ $\frac{\partial \mathcal{L}}{\partial \delta_k}(\hat{\delta}_k) > 0$ and $\tilde{Q}^{\theta,\gamma_k}_{\tilde{\pi}_k^*} < \mathcal{L}^*$. □

A.4 Proof of Corollary 1

In this section, we provide a characterization of the feasible solution space of the reward function and discount factors for the lower-level optimization problem in Eq. 10.

Corollary 1. *(Restated) Given a set of discount factors $\Gamma = \{\gamma_k\}_{k=1}^K$, the following optimization problem*

$$\max_{\{\mu_{\tilde{\pi}_k}^{\gamma_k}(s,a)\}} \quad \sum_{k=1}^K \mathcal{H}_{\tilde{\pi}_k}^{\gamma_k} = \sum_{k=1}^K \sum_{(s,a)} -\log\left(\frac{\mu_{\tilde{\pi}_k}^{\gamma_k}(s,a)}{\sum_a \mu_{\tilde{\pi}_k}^{\gamma_k}(s,a)}\right)\mu_{\tilde{\pi}_k}^{\gamma_k}(s,a) \tag{25a}$$

$$\textit{subject to} \quad f_{\tilde{\pi}_k}^{\gamma_k} = f_{\tilde{\pi}_k^*}^{\gamma_k} \quad \forall k \in [K] \tag{25b}$$

$$\sum_a \mu_{\tilde{\pi}_k}^{\gamma_k}(s',a) = \rho_0(s') + \gamma_k \sum_s \sum_a T(s'|s,a)\mu_{\tilde{\pi}_k}^{\gamma_k}(s,a) \quad \forall s' \in \mathcal{S},\ k \in [K] \tag{25c}$$

$$\mu_{\tilde{\pi}_k}^{\gamma_k}(s,a) \geq 0 \quad \forall (s,a) \in (\mathcal{S}\times\mathcal{A}),\ k \in [K], \tag{25d}$$

is feasible if and only if $rank(\Phi|b) = rank(\Phi)$ where Φ, b is defined in Proposition 2. Additionally, if the optimization problem is feasible, the optimal solution is achieved at $\mu_{\tilde{\pi}_k}^{\gamma_k}(s,a) = \mu_{\tilde{\pi}_k^*}^{\gamma_k}(s,a)$.

Proof. In Zeng et al. (2022), the authors show that when the optimization problem in Eq. 25 is feasible, strong duality holds. We first show that when the optimization problem is feasible, the optimal solution to the optimization problem in Eq. 25 are the visitation variables induced by the expert policies (i.e. $\mu_{\tilde{\pi}_k}^{\gamma_k}(s,a) = \mu_{\tilde{\pi}_k^*}^{\gamma_k}(s,a)$). Thus, Theorem 2 in Syed et al. (2008), $\tilde{\pi}_k(a|s) = \tilde{\pi}_k^*(a|s)$.

By strong duality, if the primal problem has an optimal solution, the Lagrangian dual problem also has an optimal solution, which we denote by $\hat{\theta}, \{\mu_{\hat{\pi}_k}^{\gamma_k}(s,a)\}_{k=1}^K$. Previous work shows that, $\mu_{\hat{\pi}_k}^{\gamma_k}(s,a)$ are the visitation variables induced by the entropy-regularized optimal policy, $\hat{\pi}_k$, under reward parameters $\hat{\theta}$ and the discount factor γ_k (Zhou et al., 2017).

By proof by contradiction, assume that there exists a k such that $\hat{\pi}_k \neq \tilde{\pi}_k^*$. By strong duality, an optimal solution to the dual problem satisfies the constraints of the primal problem. Thus, $f_{\hat{\pi}_k}^{\gamma_k} = f_{\tilde{\pi}_k^*}^{\gamma_k}$. Because $\hat{\pi}_k$ is a unique optimal policy for reward parameters $\hat{\theta}$ (Haarnoja et al., 2017), we have that

$$\begin{aligned}
\tilde{V}_{\hat{\pi}_k}^{\hat{\theta},\gamma_k} &> \tilde{V}_{\tilde{\pi}_k^*}^{\hat{\theta},\gamma_k} \\
\hat{\theta}^\intercal f_{\hat{\pi}_k}^{\gamma_k} + \mathcal{H}_{\hat{\pi}_k}^{\gamma_k} &> \hat{\theta}^\intercal f_{\tilde{\pi}_k}^{\gamma_k} + \mathcal{H}_{\tilde{\pi}_k}^{\gamma_k} \\
\mathcal{H}_{\hat{\pi}_k}^{\gamma_k} &> \mathcal{H}_{\tilde{\pi}_k}^{\gamma_k}.
\end{aligned}$$

This implies that for any reward parameters θ, $\tilde{\pi}_k^*$ cannot be optimal because

$$\tilde{V}_{\hat{\pi}_k}^{\theta,\gamma_k} = \theta^\intercal f_{\hat{\pi}_k}^{\gamma_k} + \mathcal{H}_{\hat{\pi}_k}^{\gamma_k} = \theta^\intercal f_{\tilde{\pi}_k}^{\gamma_k} + \mathcal{H}_{\hat{\pi}_k}^{\gamma_k} > \tilde{V}_{\tilde{\pi}_k^*}^{\theta,\gamma_k}. \tag{26}$$

However, we can construct a reward function as

$$r' = r^* + (\gamma_k^* - \gamma_k)\tilde{V}_{\tilde{\pi}_k}^{r^*,\gamma_k^*} = \left(\theta + \frac{(\gamma_k^* - \gamma_k)\tilde{V}_{\tilde{\pi}_k}^{r^*,\gamma_k^*}}{\sum_s \phi^s(s)}\mathbf{1}\right)^\intercal \phi^s.$$

By Theorem 5 in Cao et al. (2021), we have that

$$\begin{aligned}
T^a r' &= T^a(r^* + (\gamma_k^* - \gamma_k)\tilde{V}_{\tilde{\pi}_k}^{r^*,\gamma_k^*}) \\
&= T^a(r^* + \gamma_k^*\tilde{V}_{\tilde{\pi}_k}^{r^*,\gamma_k^*}) - \gamma_k T^a \tilde{V}_{\tilde{\pi}_k}^{r^*,\gamma_k^*} \\
&= \lambda \log \tilde{\pi}_k^*(a|\cdot) - \gamma_k T^a \tilde{V}_{\tilde{\pi}_k}^{r^*,\gamma_k^*} + \tilde{V}_{\tilde{\pi}_k}^{r^*,\gamma_k^*}
\end{aligned}$$

Thus, policy $\tilde{\pi}_k$ is optimal for reward function r' and the discount factor γ_k, which conflicts Eq. 26.

Thus, if the optimization problem is feasible, $\mu^{\gamma_k}_{\hat{\pi}_k}(s,a) = \mu^{\gamma_k}_{\tilde{\pi}^*_k}(s,a)$ and $\hat{\pi}_k = \tilde{\pi}^*_k$. By Proposition 2, there exists Lagrangian multipliers θ such that $\hat{\pi}_k = \tilde{\pi}^*_k$ if $rank(\Phi|b) = rank(\Phi)$.

On the other hand, if $rank(\Phi|b) = rank(\Phi)$, then there exist reward parameters θ such that $\tilde{\pi}^*_k$ is optimal for θ and γ_k. We can see that the solution $\mu^{\gamma_k}_{\tilde{\pi}^*_k}(s,a)$ satisfy the KKT conditions. By strong duality, the primal problem is feasible and has an optimal solution, $\mu^{\gamma_k}_{\tilde{\pi}^*_k}(s,a)$. □

B Algorithm

In Algorithm 1 and 2, we give the pseudocode of MPLP-IRL and MPMCE-IRL algorithms.

Algorithm 1 Mutli-planning horizon LP-IRL (MPLP-IRL)

Given a set of K observed policies $\Pi^* = \{\pi^*_k\}_{k=1}^K$, the transition dynamics T.
Place a Gaussian process prior on $g^*(\Gamma)$ (Eq. 7).
Observe g^* at a set of m points, $\{\Gamma^{(i)}\}_{i=1}^m$, with $\Gamma^{(i)} \sim \texttt{Unif}(0,1)^K$
while $m \leq$ MaxIter **do**
    Update the posterior distribution of g^* given all observed points.
    Query a new point $\Gamma^{(m)}$ based on the acquisition function.
    Observe $g^*_m = \max_r g(\Gamma^{(m)}, r)$ with constraints in Eq. 5b, 5c.
    increment m
end while
Return the reward function r and the set of discount factors Γ with the largest observed g^*.

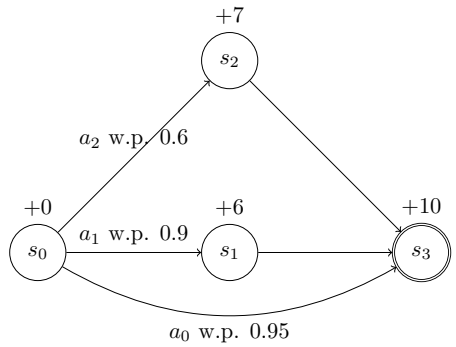
C Domains

C.1 Toy Domain with a Discrete MDP

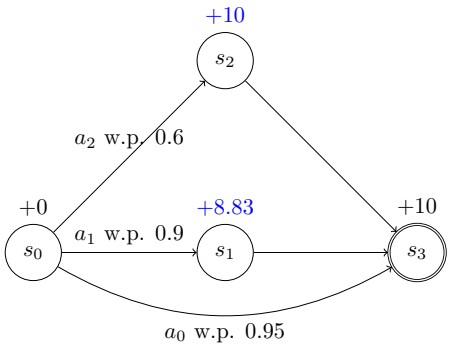
The toy domain is designed such that the optimal policy when solving standard MDPs is a 3-step piecewise function with respect to the discount factor $\gamma \in [0,1]$.

The MDP is described in Fig. 4a. Specifically, s_0 is the initial state, s_3 is the absorbing state. The agent gets a large positive reward when getting to the absorbing state ($r^*(s_3) = 10$). The agent gets some small rewards when getting to s_1, s_2 ($r^*(s_1) = 6$, $r^*(s_2) = 7$). With action a_0, the agent moves to the absorbing state s_3 with probability (w.p.) 0.95. With action a_1, the agent moves to s_1 w.p. 0.9. With action a_2, the agent moves to s_2 w.p. 0.6. The agent stays otherwise. Note that although s_2 has a larger reward, the agent faces more stochasticity. Thus, there is a trade-off when the discount factor γ varies.

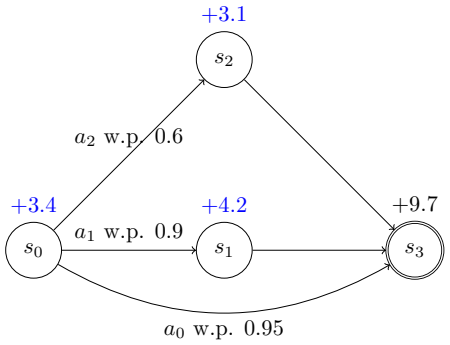
Expert policies for MPLP-IRL. Let $\pi^*_0, \pi^*_1, \pi^*_2$ denote the optimal policies where the optimal actions are a_0, a_1, a_2 ($\pi^*_i = \mathbb{1}_{a_i}$), respectively. We note that for this toy example, in state s_1, s_2, s_3, $\pi^*_0, \pi^*_1, \pi^*_2$ are equally optimal. Thus, we focus on s_0 in the later analysis.



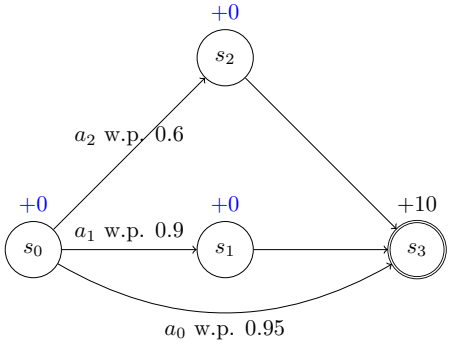
(a) Discrete MDP with the true reward function.



(b) Discrete MDP with the learned function of MPLP-IRL.



(c) Discrete MDP with the learned function of MPMCE-IRL.



(d) Discrete MDP with the learned function of naive LP-IRL.

Figure 4: Toy Domain

Algorithm 2 Mutli-planning horizon MCE-IRL (MCELP-IRL)

Given a set of K observed policies $\tilde{\Pi}^* = \{\tilde{\pi}_k^*\}_{k=1}^K$ and the transition dynamics T.
Place a Gaussian process prior on $g^*(\Gamma)$ (Eq. 12).
Observe g^* at a set of m points, $\{\Gamma^{(i)}\}_{i=1}^m$, with $\Gamma^{(i)} \sim \texttt{Unif}(0,1)^K$
while $m \leq$ MaxIter **do**
  Update the posterior distribution of g^* given all observed points.
  Query a new point $\Gamma^{(m)}$ based on the acquisition function.
  Solve the Lagrangian dual of the constrained optimization problem in Eq. 12 using the ML-IRL algorithm in Zeng et al. (2022).
  Denote the optimal solution to the Lagrangian dual problem as $\tilde{\theta}$. Calculate the duality gap as $l = \sum_{k=1}^K \tilde{\theta}^\intercal (f^{\gamma_k}_{\tilde{\pi}_k^{\tilde{\theta}}} - f^{\gamma_k}_{\tilde{\pi}_k^*})$
  if $l \leq \epsilon$ **then**
    $g_m^*(\Gamma) = \sum_{k=1}^K \mathcal{H}^{\gamma_k}_{\tilde{\pi}_k^{\tilde{\theta}}}$
  else
    $g_m^*(\Gamma) = -|l|$
  end if
  increment m
end while
Return the reward parameters θ and the set of discount factors Γ with the largest observed g^*.

$$\begin{cases} V_\gamma^{\pi_0}(s_0) & = \frac{0.95 \cdot 10}{1-0.05\gamma} \\ V_\gamma^{\pi_1}(s_0) & = \frac{0.9(6+10\gamma)}{1-0.1\gamma} \\ V_\gamma^{\pi_2}(s_0) & = \frac{0.6(7+10\gamma)}{1-0.4\gamma} \end{cases}$$

Solve pairwise difference between value functions and let $\gamma_0 \approx 0.432$, $\gamma_1 \approx 0.876$. When $\gamma < \gamma_0$, $\pi^* = \pi_0$. When $\gamma_0 < \gamma < \gamma_1$, $\pi^* = \pi_1$. When $\gamma > \gamma_1$, $\pi^* = \pi_2$. The value functions of these 3 expert policies, Π^*, under different discount values are shown in 1a.

Satisfaction of Assumption 1. Figure 9a plots the value function of all three expert policies of s_0 under the true reward function r^* with respect to discount factors $\gamma \in [0,1]$. Assumption 1 is violated ($Q^{r^*,\gamma_i^*}_{\pi_i^*}(s, \pi_i^*(s)) = Q^{r^*,\gamma_i^*}_{\pi_i^*}(s, \pi_j^*(s))$) if and only if any of the two lines intersect. We see that across the entire domain ($[0,1]^3$), there are only three sets of discount factors that violate Assumption 1. Thus in practice, the likelihood of violating Assumption 1 is low.

Expert policies for MPMCE-IRL. For MPMCE-IRL, we observe expert demonstrations from 3 expert policies under discount factors $\Gamma^* = \{0.3, 0.5, 0.95\}$.

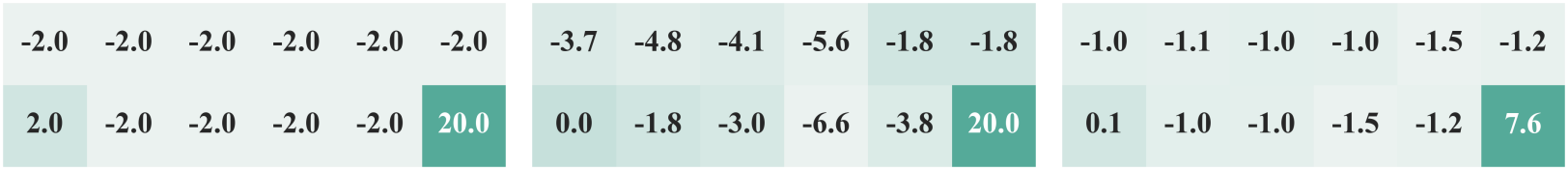
C.2 Big-Small Domain

In the big-small domain, there are two absorbing states: one at the left bottom cell with a small reward (+2) and one at the right bottom cell with a large reward (+20). Each step costs −2 if the

agent does not reach the absorbing state. The true reward function is plotted in 5a. The agent can start from anywhere in the grid world except for the absorbing states. The agent can choose to move {`right, down, left, up`} at each state. If the agent comes across the wall by taking an action, the agent stays where it is and moves to the corresponding state otherwise. The transition dynamics of the grid domain are deterministic.

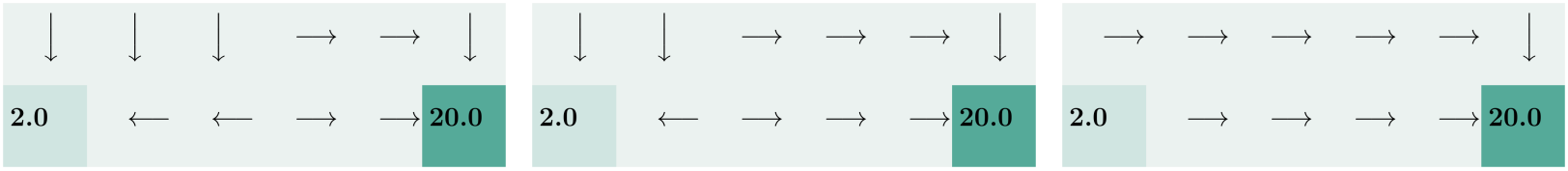
Expert policies for MPLP-IRL. We observe expert demonstrations from 3 distinct expert policies whose optimal actions are visualized in Fig. 6. We see that when the discount factor is small, the agent will go for the closest reward regardless of its magnitude (Fig. 6a). When the discount factor is large, the agent prefers the large reward regardless of how far the reward is (Fig. 6c).

Expert policies for MPMCE-IRL. For MPMCE-IRL, we observe expert demonstrations from 3 expert policies under discount factors $\Gamma^* = \{0.1, 0.45, 0.9\}$



(a) The true reward function r^*. (b) The learned reward function of MPLP-IRL. (c) The learned reward function of MPMCE-IRL.

Figure 5: The reward function of the big-small domain of each state.



(a) The optimal actions of expert policy π_0^*, ($\gamma_0^* \in [0, 0.2]$) (b) The optimal actions of expert policy π_1^*, ($\gamma_1^* \in (0.2, 0.68]$) (c) The optimal actions of expert policy π_2^*, ($\gamma_2^* \in [0.87, 1]$)

Figure 6: Visualization of the optimal policies for standard MDPs of the big-small domain.

C.3 Cliff Domain

In the cliff domain, the trajectory ends whenever the agent falls off the cliff (the top row in Fig 7a. There is one rewarding state at the upper right (+20). The agent gets a large penalty whenever it falls off the cliff (−10). There is a small cost for each step (−2 when the agent is close to the cliff and −1 when the agent is far away from the cliff). For each action, the agent moves in the intended direction with probability 0.9, and moves in a random other direction with probability 0.1.

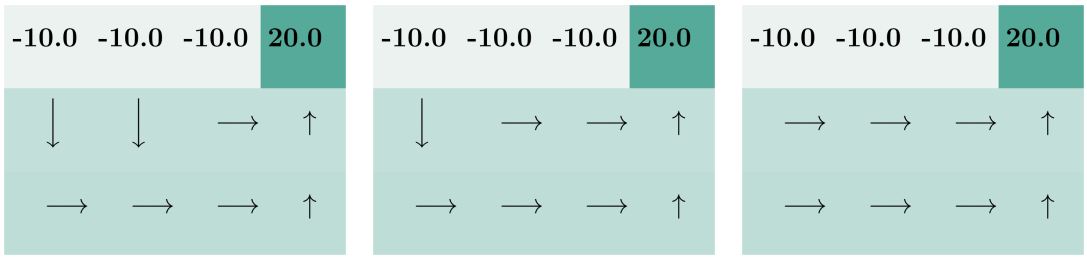
Expert policies for MPLP-IRL. We observe 3 distinct expert policies whose optimal actions are visualized in Fig. 8. We see that when the discount factor is small, the agent is not willing to

risk and tries to avoid walking along the cliff (Fig. 8a). When the discount factor is large, the agent risks falling off the cliff for good returns (Fig. 8c).

Expert policies for MPMCE-IRL. For MPMCE-IRL, we observe expert demonstrations from 3 expert policies under discount factors $\Gamma^* = \{0, 0.2, 0.52\}$.

(a) The true reward function r^*. (b) The learned reward function of MPLP-IRL. (c) The learned reward function of MPMCE-IRL.

Figure 7: The reward function of the cliff domain of each state.



(a) The optimal actions of expert policy π_0^*, $(\gamma_0^* \in [0.06, 0.28])$ (b) The optimal actions of expert policy π_1^*, $(\gamma_1^* \in (0.28, 0.52])$ (c) The optimal actions of expert policy π_2^*, $(\gamma_2^* \in (0.52, 0.95])$

Figure 8: Visualization of the optimal policies for standard MDPs of the cliff domain.

D Generalization Error

D.1 Generation of random environments

To test the generalizability of the learned reward function of our algorithms, for each domain environment, we generate $N = 100$ random transition dynamics $\{T^{(n)}\}_{n=1}^N$ and discount factors $\{\gamma^{(n)}\}_{n=1}^N$. For the toy domain, we randomize the probability of the agent' moving to the next state by taking actions from state s_0. From state s_1, s_2, the agent still moves to the absorbing state with probability 1 regardless of its actions. For the big-small and cliff domains, we generate random environments by adding noise to the agent's intended direction. For each action, the agent moves in the intended direction with probability $\epsilon \sim \text{Unif}(0, 1)$, and moves in a random other direction with probability $1 - \epsilon$.

D.2 Evaluation Metrics

For each randomized transition dynamics and discount factor generated from Appendix D.1, we first solve the optimal policies under the true reward function r^* and the learned reward function $\hat{r}$, denoted as $\pi^{*(n)}, \hat{\pi}^{(n)}$ respectively. We evaluate the generalizability of the learned reward function $\hat{r}$ by computing the normalized difference of the value function under the true r^* and randomly generated environments,

$$\triangle V_{\hat{r}} = \frac{1}{N} \sum_{n=1}^{N} \left(V_{\pi^{*(n)}}^{r^*, T^{(n)}, \gamma^{(n)}} - V_{\hat{\pi}^{(n)}}^{r^*, T^{(n)}, \gamma^{(n)}} \right) / V_{\pi^{*(n)}}^{r^*, T^{(n)}, \gamma^{(n)}} \tag{27}$$

E Additional Results

E.1 Undesirable Global Maxima of Naive Extensions of LP-IRL

In this Section, we demonstrate that naive extensions of LP-IRL in Eq. 4 give an undesirable global optimum. In Fig. 9a, we see that under the true reward function, r^*, each expert policy is optimal (when the corresponding line is on the top) for a continuous interval. However, in Fig. 9b, we see that naive LP-IRL algorithm identifies a set of discount factors, $\tilde{\Gamma} = \{0, 1, 1\}$ (the orange and the purple dashed lines intersect at $\gamma = 1$). Under the learned discount factors, expert policies π_1^* and π_2^* are not distinguishable from each other under any reward function, which violates assumption 1.

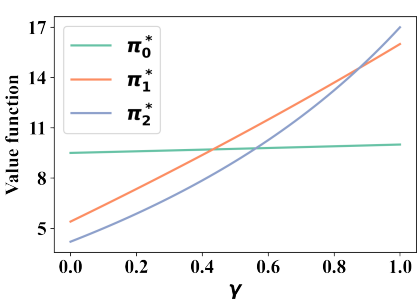
Furthermore, the naive LP-IRL cannot recover the structure of the true reward function In Fig. 4d, we see that the naive LP-IRL assigns a reward only to the absorbing state with $r(s_1) = r(s_2) = 0$. This reward function gives a generalization error (Eq. 27) of 0.213 ± 0.218. In Fig. 9b, we see that under this learned reward function, when $\gamma < 1$, the expert policy π_0^* performs substantially better than π_1^*, π_2^* while expert policies π_1^*, π_2^* perform similarly to each other. In contrast, under the true reward function (Fig. 9a), we can find a set of discount factors such that the expert policies are distinguishable from each other.

E.2 Comparison of the order of true and learned discount factors

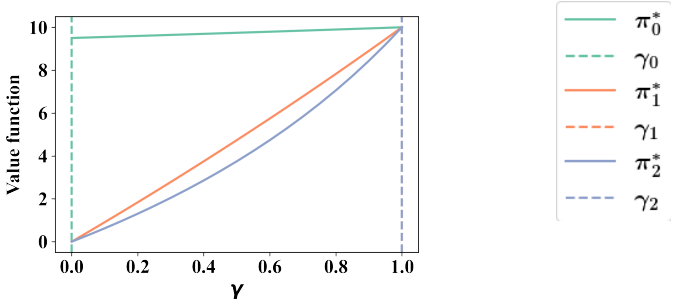
In Table 1, we provide a full comparison of the true discount factors Γ^* and the learned discount factors $\tilde{\Gamma}$. Both MPLP-IRL and MPMCE-IRL recover the order of the true discount factors.

E.3 Convergence of MPLP-IRL and MPMCE-IRL

For MPLP-IRL, we approximate the global optimum by performing a grid search on the range $[0, 1]^K$. For MPMCE-IRL, it is computationally heavy to solve the optimization problem in Eq. 12 for 10^K times. We instead compare the best current objective value of BO to the objective value evaluated under the true reward function and discount factors. From Fig. 3, we see across all domains, MPLP-IRL converges to the global maximum within 100 iterations while MPMCE-IRL converges within 50 iterations,



(a) The value function $V_{\pi_k^*}^{\gamma, r*}(s_0)$ of expert policies π_0^*, π_1^*, π_2^* under the true reward r^*.

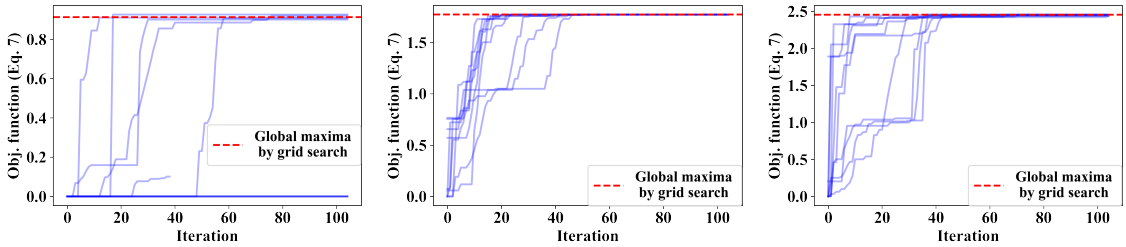


(b) The value function $V_{\pi_k}^{\tilde{\gamma}, \tilde{r}}(s_0)$ of reconstructed optimal policies π_0, π_1, π_2 under the learned reward function of the naive extension of LP-IRL (Eq. 4), $\tilde{r}$.

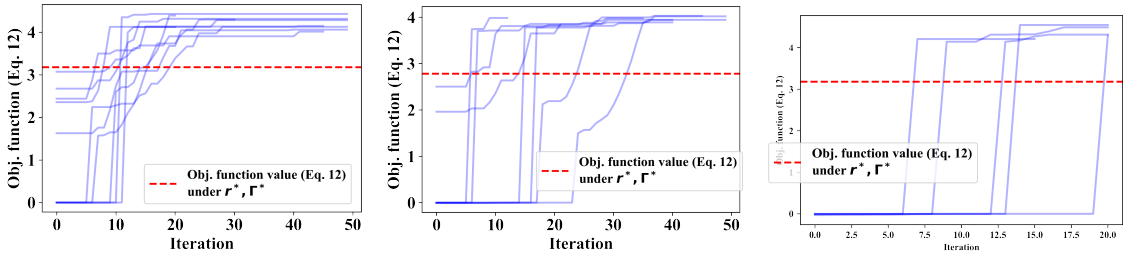
Figure 9: Plots of the value function of the initial state under (a) the true reward function r^*, (b) the learned reward function of the naive extensions of LP-IRL, $\tilde{r}$: x, y-axes represent the discount factor $\gamma \in [0, 1]$ and the value function of expert policies or reconstructed optimal policies, respectively. Each color represents a different policy. The dashed lines in (b) represent the learned discount factors, $\tilde{\Gamma}$. Each policy is optimal when the corresponding line is on the top. Policies are equally optimal when the lines interest.

	True Γ^* of MPLP-IRL	Learned $\tilde{\Gamma}$ of MPLP-IRL	True Γ^* of MPMCE-IRL	Learned $\tilde{\Gamma}$ of MPMCE-IRL
Toy	π_1^*:[0,0.43] π_2^*:(0.43, 0.87] π_3^*:[0.87,1)	$\pi_1 : 0$ $\pi_2 : 0.35$ $\pi_3 : 1$	$\tilde{\pi}_1^* : 0.3$ $\tilde{\pi}_2^* : 0.5$ $\tilde{\pi}_3^* : 0.95$	$\tilde{\pi}_1 : 0.53$ $\tilde{\pi}_2 : 0.82$ $\tilde{\pi}_3 : 0.98$
Big-Small	π_1^*:[0,0.2] π_2^*:(0.2,0.68] π_3^*:[0.87,1]	$\pi_1 : 0.38$ $\pi_3 : 0.74$ $\pi_3 : 1$	$\tilde{\pi}_1^* : 0$ $\tilde{\pi}_2^* : 0.45$ $\tilde{\pi}_3^* : 0.9$	$\tilde{\pi}_1 : 0.30$ $\tilde{\pi}_2 : 0.62$ $\tilde{\pi}_3 : 0.98$
Cliff	π_1^*:[0.06, 0.28] π_2^*:(0.28,0.52] π_3^*:(0.52,0.95]	$\pi_1 : 0.04$ $\pi_2 : 0.44$ $\pi_3 : 0.84$	$\tilde{\pi}_1^* : 0$ $\tilde{\pi}_2^* : 0.2$ $\tilde{\pi}_3^* : 0.52$	$\tilde{\pi}_1 : 0$ $\tilde{\pi}_2 : 0.31$ $\tilde{\pi}_3 : 0.55$

Table 1: Table of the true discount factors Γ^* and the learned discount factors $\tilde{\Gamma}$. Both MPLP-IRL and MPMCE-IRL recover the order of the true discount factors.



(a) MPLP-IRL: the toy domain (b) MPLP-IRL: the big-small domain (c) MPLP-IRL: the cliff domain



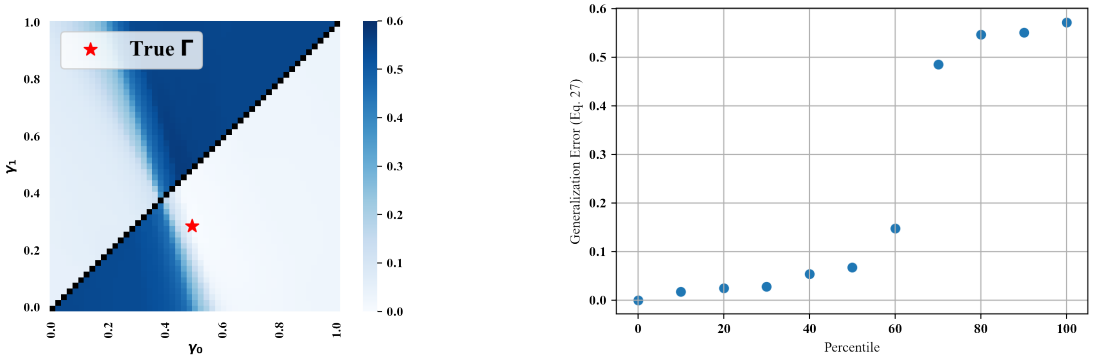
(d) MPMCE-IRL: the toy domain (e) MPMCE-IRL: the big-small domain (f) MPMCE-IRL: the cliff domain

Figure 10: Trace plots of the best observed objective value of BO: x, y-axis represent the iteration and the best observed objective value, respectively. The red dashed line represents an approximate global maximum from performing a grid search over $[0, 1]^K$ or the objective value under the ground truth.

E.4 Identifiability and Generalizability Analysis for the Toy Domain

In this section, we study the identifiability of the reward function for the toy domain, when the discount factors are misspecified. We only provide expert demonstrations from 2 expert policies. We obtain Γ on a grid space of $\Gamma \in [0, 1]^K$ with an interval of 0.01. For each set of assigned discount factors, we calculate $rank(\Phi|b)$ and $rank(\Phi)$. We find out that for all the given discount factors, $rank(\Phi|b) = rank(\Phi) = 3|\mathcal{S}| - 1$. By Proposition 2, this result implies that there exists reward functions that reconstruct expert policies for a large set of discount factors.

We further study the generalizability of these reward functions. In Fig. 11, we see that $\sim 40\%$ of these feasible reward functions do not generalize well to new tasks with generalization errors (Eq. 27) larger than 0.5, which emphasizes the importance of learning the discount factors correctly.



(a) Heatmap of the generalization error of the reward functions (Eq. 27) given (mis)specified discount factors.

(b) Percentile plot of the generalization error of the reward functions (Eq. 27) given (mis)specified discount factors.)

Figure 11: The (a) heatmap and (2) percentile plot of the generalization error of the reward functions (Eq. 27) given (mis)specified discount factors: In (a) x, y-axis represents the given γ_0, γ_1 respectively. The red star represents the true discount factors γ_0^*, γ_1^*. There are no feasible solutions along the diagonal. Lighter blue indicates a smaller error and vice versa. In (b) x, y-axis represents the percentile and the generalization error, respectively.