Maximum-Likelihood Inverse Reinforcement Learning with Finite-Time Guarantees

Siliang Zeng
University of Minnesota, Twin Cities
Minneapolis, MN, USA
zeng0176@umn.edu

Chenliang Li
The Chinese University of Hong Kong,
Shenzhen, China
chenliangli@link.cuhk.edu.cn

Alfredo Garcia
Texas A&M University
College Station, TX, USA
alfredo.garcia@tamu.edu

Mingyi Hong
University of Minnesota, Twin Cities
Minneapolis, MN, USA
mhong@umn.edu

Abstract

Inverse reinforcement learning (IRL) aims to recover the reward function and the associated optimal policy that best fits observed sequences of states and actions implemented by an expert. Many algorithms for IRL have an inherently nested structure: the inner loop finds the optimal policy given parametrized rewards while the outer loop updates the estimates towards optimizing a measure of fit. For high dimensional environments such nested-loop structure entails a significant computational burden. To reduce the computational burden of a nested loop, novel methods such as SQIL [1] and IQ-Learn [2] emphasize policy estimation at the expense of reward estimation accuracy. However, without accurate estimated rewards, it is not possible to do counterfactual analysis such as predicting the optimal policy under different environment dynamics and/or learning new tasks. In this paper we develop a novel *single-loop* algorithm for IRL that does not compromise reward estimation accuracy. In the proposed algorithm, each policy improvement step is followed by a stochastic gradient step for likelihood maximization. We show that the proposed algorithm provably converges to a stationary solution with a finite-time guarantee. If the reward is parameterized linearly, we show the identified solution corresponds to the solution of the maximum entropy IRL problem. Finally, by using robotics control problems in MuJoCo and their transfer settings, we show that the proposed algorithm achieves superior performance compared with other IRL and imitation learning benchmarks.

1 Introduction

Given observed trajectories of states and actions implemented by an expert, we consider the problem of estimating the reinforcement learning environment in which the expert was trained. This problem is generally referred to as inverse reinforcement learning (IRL) (see [3] for a recent survey). Assuming the environment dynamics are known (or available online), the IRL problem consists of estimating the reward function and the expert's policy (optimizing such rewards) that best fits the data. While there are limitations on the identifiability of rewards [4], the estimation of rewards based upon expert trajectories enables important counterfactual analysis such as the estimation of optimal policies under *different* environment dynamics and/or reinforcement learning of *new* tasks.

In the seminal work [5], the authors developed an IRL formulation, in which the model for the expert's behavior is the policy that maximizes entropy subject to a constraint requiring that the expected

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

features under such policy match the empirical averages in the expert's observation dataset. The algorithms developed for MaxEnt-IRL [5–7] have a nested loop structure, alternating between an outer loop with a reward update step, and an inner loop that calculates the explicit policy estimates. The computational burden of this nested structure is manageable in tabular environments, but it becomes significant in high dimensional settings requiring function approximation.

Towards developing more efficient IRL algorithms, a number of works [8–12] propose to leverage the idea of adversarial training [13]. These algorithms learn a non-stationary reward function through training a discriminator, which is then used to guide the policy to match the behavior trajectories from the expert dataset. However, [14] pointed out that the resulting discriminator (hence the reward function) typically cannot be used in new learning tasks, since it is highly dependent on the corresponding policy and current environment dynamics. Moreover, due to the brittle approximation techniques and sensitive hyperparameter choice in the adversarial training, these IRL algorithms can be unstable. [15, 16].

More recent works [1, 2] have developed algorithms to alleviate the computational burden of the nested-loop training procedures. In [1], the authors propose to model the IRL using certain maximum entropy RL problem with specific reward function (which assigns $r = +1$ for matching expert demonstrations and $r = 0$ for all other behaviors). Then a soft Q imitation learning (SQIL) algorithm is developed. In [2], the authors propose to transform the standard formulation of IRL (discussed above) into a single-level problem, through learning a soft Q-function to implicitly represent the reward function and the policy. An inverse soft-Q learning (IQ-Learn) algorithm is then developed, which is shown to be effective in estimating the policy for the environment that it is trained on. Despite being computationally efficient, IQ-Learn sacrifices the accuracy in estimating the rewards since it indirectly recovers rewards from a soft Q-function approximator which is highly dependent upon the environment dynamics and does not strictly satisfy the soft-Bellman equation. Therefore it is not well-suited for counterfactual prediction or transfer learning setting.

Finally, in f-IRL [14] the authors consider an approach for estimating rewards based on the minimization of several measures of divergence with respect to the expert's state visitation measure. The approach is limited to estimating rewards that only depend on state. Moreover, while the results reported are based upon a single-loop implementation, the paper does not provide a convergence guarantee to support performance. We refer the readers to Appendix A for other related works.

Our Contributions. The goal of this work is to develop an algorithm for IRL which is capable of producing high-quality estimates of *both* rewards and behavior policies with finite-time guarantees. The major contributions of this work are listed below.

• We consider a formulation of IRL based on maximum likelihood (ML) estimation over optimal (entropy-regularized) policies, and prove that a strong duality relationship with maximum entropy IRL holds if rewards are represented by a *linear* combination of features. [1] The ML formulation is a *bi-level* optimization problem, where the upper-level problem maximizes the likelihood function, while the lower-level finds the optimal policy under the current reward parameterization. Such a bi-level structure is not only instrumental to the subsequent algorithm design, but is also flexible to incorporate the use of state-only, as well as the regular reward function (which depends on the state and action pair). The former is suitable for transfer learning since it is *insensitive* to the changes of the environment dynamics, while the latter can be used to efficiently imitate the expert policy.

• Based on the ML-IRL formulation, we develop an efficient algorithm. To avoid the computational burden of repeatedly solving the lower-level policy optimization problem, the proposed algorithm has a single-loop structure where the policy improvement step and reward optimization step are performed alternatingly so that each step can be performed relatively cheaply. Further, we show that the algorithm has strong theoretical guarantees: to achieve certain ϵ-approximate stationary solution for a non-linearly parameterized problem, it requires $\mathcal{O}(\epsilon^{-2})$ steps of policy and reward updates each. To our knowledge, it is the first algorithm which has finite-time guarantee for the IRL problem under nonlinear parameterization of reward functions.

• We conduct extensive experiments to demonstrate that the proposed algorithm outperforms many state-of-the-art IRL algorithms in *both* policy estimation and reward recovery. In particular, when

[1] Heuristic arguments for this duality result are discussed in [5] wherein the distribution of state-action paths is approximated (see equation (4) in [5]) and the equivalence between maximum entropy estimation and maximum likelihood (over the class of exponential distributions) [17] is invoked.

transferring to a new environment, RL algorithms using rewards recovered by the proposed algorithm outperform those that use rewards recovered from existing IRL and imitation learning benchmarks.

2 Preliminaries

In this section, we review the fundamentals of the maximum entropy inverse reinforcement learning (MaxEnt-IRL). We consider an MDP defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \eta, r, \gamma)$; $\mathcal{S}$ and $\mathcal{A}$ denote the state space and the action space respectively; $\mathcal{P}(s'|s,a) : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$ denotes the transition probability; $\eta(\cdot)$ denotes the distribution for the initial state; $r(s,a) : \mathcal{S} \times \mathcal{A} \to R$ is the reward function and γ is a discount factor.

The MaxEnt-IRL formulation [6, 18–20] consists of finding a policy maximizing entropy subject to the expected features under such policy matching the empirical averages in the expert's observation dataset. Specifically, the MaxEnt-IRL formulation is given by:

$$\begin{aligned} \max_{\pi} \quad & H(\pi) := \mathbb{E}_{\tau\sim\pi}\Big[\sum_{t=0}^{\infty} -\gamma^t \log \pi(a_t|s_t)\Big] \\ \text{s.t.} \quad & \mathbb{E}_{\tau\sim\pi}\Big[\sum_{t=0}^{\infty} \gamma^t \phi(s_t,a_t)\Big] = \mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\Big[\sum_{t=0}^{\infty} \gamma^t \phi(s_t,a_t)\Big] \end{aligned} \tag{MaxEnt-IRL}$$

where $\tau = \{(s_t,a_t)\}_{t=0}^{\infty}$ denotes a trajectory, $\phi(s,a)$ is the feature vector of the state-action pair (s,a) and π^{E} denotes the expert policy. Let θ denote the dual variable for the linear constraint, then the Lagrangian of (MaxEnt-IRL) is given by

$$\mathcal{L}(\pi,\theta) := H(\pi) + \Big\langle \theta, \mathbb{E}_{\tau\sim\pi}\Big[\sum_{t=0}^{\infty}\gamma^t\phi(s_t,a_t)\Big] - \mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\Big[\sum_{t=0}^{\infty}\gamma^t\phi(s_t,a_t)\Big]\Big\rangle. \tag{1}$$

In [6, 18, 19], the authors proposed a "dual descent" algorithm, which alternates between *i)* solving $\max_\pi \mathcal{L}(\pi,\theta)$ for fixed θ, and *ii)* a gradient descent step to optimize the dual variable θ. It is shown that the optimizer π^*_θ in step *i)* can be recursively defined as $\pi^*_\theta(a_t|s_t) = \frac{Z_{a_t|s_t,\theta}}{Z_{s_t,\theta}}$, where $\log Z_{a_t|s_t,\theta} = \phi(s_t,a_t)^T\theta + \gamma \mathbb{E}_{s_{t+1}\sim\mathcal{P}(\cdot|s_t,a_t)}\left[\log Z_{s_{t+1},\theta}\right]$ and $\log Z_{s_t,\theta} = \log\big(\sum_{a\in\mathcal{A}} Z_{a|s_t,\theta}\big)$.

From a computational perspective, the above algorithm is not efficient: it has a nested-loop structure, which repeatedly computes the optimal policy π^*_θ under each variable θ. It is known that when the underlying MDP is of high-dimension, such an algorithm can be computationally prohibitive [9, 10].

Recent work [2] proposed an algorithm called IQ-Learn to improve upon the MaxEnt-IRL by considering a saddle-point formulation:

$$\min_r \max_\pi \ \Big\{ H(\pi) + \mathbb{E}_{\tau\sim\pi}\big[\sum_{t=0}^{\infty}\gamma^t \cdot r(s_t,a_t)\big] - \mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\big[\sum_{t=0}^{\infty}\gamma^t\cdot r(s_t,a_t)\big]\Big\} \tag{2}$$

where $r(s_t,a_t)$ is the reward associated with state-action pair (s_t,a_t). The authors show that this problem can be transformed into an optimization problem *only* defined in terms of the soft Q-function, which implicitly represents both reward and policy. IQ-Learn is shown to be effective in imitating the expert behavior while only relying on the estimation of the soft Q-function. However, the implicit reward estimate obtained is not necessarily accurate since its soft Q-function estimate depends on the environment dynamics and does not strictly satisfy the soft-Bellman equation. Hence, it is difficult to transfer the recovered reward function to new environments.

3 Problem Formulation

In this section, we consider a ML formulation of the IRL problem and formalize a duality relationship with maximum entropy-based formulation (MaxEnt-IRL).

Maximum Log-Likelihood IRL (ML-IRL)

A model of the expert's behavior is a randomized policy $\pi_\theta(\cdot|s)$, where π_θ is a specific policy corresponding to the reward parameter θ. With the state dynamics $\mathcal{P}(s_{t+1}|s_t,a_t)$, the discounted log-likelihood of observing the expert trajectory τ under model π_θ can be written follows:

$$\mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\big[\log\prod_{t\geq 0}(\mathcal{P}(s_{t+1}|s_t,a_t)\pi_\theta(a_t|s_t))^{\gamma^t}\big] = \mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\big[\sum_{t\geq 0}\gamma^t\log\pi_\theta(a_t|s_t)\big] + \mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\big[\sum_{t\geq 0}\gamma^t\log\mathcal{P}(s_{t+1}|s_t,a_t)\big].$$

Then we consider the following maximum log-likelihood IRL formulation:

$$\max_\theta \quad L(\theta) := \mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\big[\sum_{t=0}^{\infty}\gamma^t\log\pi_\theta(a_t|s_t)\big] \tag{ML-IRL}$$

$$s.t \quad \pi_\theta := \arg\max_\pi\ \mathbb{E}_{\tau\sim\pi}\bigg[\sum_{t=0}^{\infty}\gamma^t\bigg(r(s_t,a_t;\theta)+\mathcal{H}(\pi(\cdot|s_t))\bigg)\bigg], \tag{3a}$$

where $r(s,a;\theta)$ is the reward function and $\mathcal{H}(\pi(\cdot|s)) := -\sum_{a\in\mathcal{A}}\pi(a|s)\log\pi(a|s)$.

We now make some remarks about ML-IRL. First, the problem takes the form of a *bi-level* optimization problem, where the *upper-level* problem (ML-IRL) optimizes the reward parameter θ, while the *lower-level* problem describes the corresponding policy π_θ as the solution to an entropy-regularized MDP ([21, 22]). In what follows we will leverage recently developed (stochastic) algorithms for bi-level optimization [23–25], that avoid the high complexity resulted from nested loop algorithms. Second, it is reasonable to use the ML function as the loss, because it searches for a reward function which generates a behavior policy that can best fit the expert demonstrations. While the ML function has been considered in [26, 27], they rely on heuristic algorithms with nested-loop computations to solve their IRL formulations, and the theoretical properties are not studied. Finally, the lower-level problem has been well-studied in the literature [21, 22, 28–30]. The entropy regularization in (3a) ensures the uniqueness of the optimal policy π_θ under the fixed reward function $r(s,a;\theta)$ [21, 28]. Even when the underlying MDP is high-dimensional and/or complex, the optimal policy could still be obtained; see recent developments in [21, 22]. We close this section by formally establishing a connection between (MaxEnt-IRL) and (ML-IRL).

Theorem 1. **(Strong Duality)** *Suppose that the reward function is given as: $r(s,a;\theta) := \phi(s,a)^T\theta$, for all $s\in\mathcal{S}$ and $a\in\mathcal{A}$. Then* (ML-IRL) *is the Lagrangian dual of* (MaxEnt-IRL)*. Furthermore, strong duality holds, that is: $L(\theta^*) = H(\pi^*)$, where θ^* and π^* are the global optimal solutions for problems* (ML-IRL) *and* (MaxEnt-IRL)*, respectively.*

The proof of Theorem 1 is relegated to Appendix G. To our knowledge this result which specifically addresses the (MaxEnt-IRL) formulation is novel. Under finite horizon, a duality between ML estimation and maximum causal entropy is obtained in [18, Theorem 3]. However, the problem considered in that paper is not in RL nor IRL setting, therefore they cannot be directly used in the context of the present paper.

The above duality result reveals a strong connection between the two formulations under linear reward parameterization. Due to the duality result, we know that (ML-IRL) is a concave problem under linear reward parameterization. In this case, any stationary solution to (ML-IRL) is a global optimal estimator of the reward parameter.

4 The Proposed Algorithm

In this section, we design algorithms for (ML-IRL). Recall that one major drawback of algorithms for (MaxEnt-IRL) is that, they repeatedly solve certain policy optimization problem in the inner loop. Even though the recently proposed algorithm IQ-Learn [2] tries to improve the computational efficiency through implicitly representing the reward function and the policy by a Q-function approximator, it has sacrificed the estimation accuracy of the recovered reward. Therefore, one important goal of our design is to find provably efficient algorithms that can avoid high-complexity operations and accurately recover the reward function. Specifically, it is desirable that the resulting algorithm only uses a finite number of reward and policy updates to reach certain high-quality solutions.

To proceed, we will leverage the special *bi-level* structure of the ML-IRL problem. The idea is to alternate between one step of policy update to improve the solution of the lower-level problem, and

Algorithm 1 *Maximum Likelihood Inverse Reinforcement Learning (ML-IRL)*

Input: Initialize reward parameter θ_0 and policy π_0. Set the reward parameter's stepsize as α.
for $k = 0, 1, \ldots, K-1$ **do**
Policy Evaluation: Compute $Q^{\text{soft}}_{r_{\theta_k},\pi_k}(\cdot,\cdot)$ under reward function $r(\cdot,\cdot;\theta_k)$
Policy Improvement: $\pi_{k+1}(\cdot|s) \propto \exp(Q^{\text{soft}}_{r_{\theta_k},\pi_k}(s,\cdot)), \forall s \in \mathcal{S}$.
Data Sampling I: Sampling an expert trajectory $\tau_k^{\text{E}} := \{s_t, a_t\}_{t\geq 0}$
Data Sampling II: Sampling a trajectory $\tau_k^{\text{A}} := \{s_t, a_t\}_{t\geq 0}$ from the current policy π_{k+1}
Estimating Gradient: $g_k := h(\theta_k; \tau_k^{\text{E}}) - h(\theta_k; \tau_k^{\text{A}})$ where $h(\theta;\tau) := \sum_{t\geq 0}\gamma^t \nabla_\theta r(s_t, a_t; \theta)$
Reward Parameter Update: $\theta_{k+1} := \theta_k + \alpha g_k$
end for

one step of the parameter update which improves the upper-level loss function. At each iteration k, given the current policy π_k and the reward parameter θ_k, a new policy π_{k+1} is generated from the policy improvement step, and θ_{k+1} is generated by the reward optimization step.

This kind of alternating update is efficient, because there is no need to completely solve the policy optimization subproblem, before updating the reward parameters. It has been used in many other RL related settings as well. For example, the well-known actor-critic (AC) algorithm for policy optimization [31, 32, 23] alternates between one step of policy update, and one step of critic parameter update. Below we present the details of our algorithm at a given iteration k.

Policy Improvement Step. Let us consider optimizing the lower-level problem, when the reward parameter θ_k is held fixed. Towards this end, define the so-called soft Q and soft value functions for a given policy π_k and a reward parameter θ_k:

$$V^{\text{soft}}_{r_k,\pi_k}(s) = \mathbb{E}_{\pi_k}\left[\sum_{t=0}^{\infty}\gamma^t\bigg(r(s_t,a_t;\theta_k) + \mathcal{H}(\pi_k(\cdot|s_t))\bigg)\bigg| s_0 = s\right] \tag{4a}$$

$$Q^{\text{soft}}_{r_k,\pi_k}(s,a) = r(s,a;\theta_k) + \gamma\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)}\left[V^{\text{soft}}_{r_k,\pi_k}(s)\right] \tag{4b}$$

We will adopt the well-known *soft policy iteration* [21] to optimize the lower-level problem (3a). Under the current reward parameter θ_k and the policy π_k, the soft policy iteration generates a new policy π_{k+1} as follows

$$\pi_{k+1}(a|s) \propto \exp\left(Q^{\text{soft}}_{r_{\theta_k},\pi_k}(s,a)\right), \quad \forall s \in \mathcal{S}, a \in \mathcal{A}. \tag{5}$$

Under a fixed reward function, it can be shown that the new policy π_{k+1} monotonically improves π_k, and it converges linearly to the optimal policy; see [21, Theorem 4] and [28, Thoerem 1].

Note that in practice, we usually do not have direct access to the exact soft Q-function in (4b). In order to perform the policy improvement, a few stochastic update steps in soft Q-learning [21] or soft Actor-Critic (SAC) [22] could be used to replace the one-step soft policy iteration (5). In the appendix, we present Alg. 2 to demonstrate such practical implementation of our proposed algorithm.

Reward Optimization Step. We propose to use a stochastic gradient-type algorithm to optimize θ. Towards this end, let us first derive the exact gradient $\nabla L(\theta)$. See Appendix D for detailed proof.

Lemma 1. *The gradient of the likelihood function $L(\theta)$ can be expressed as follows:*

$$\nabla L(\theta) = \mathbb{E}_{\tau\sim\pi^{\text{E}}}\left[\sum_{t\geq 0}\gamma^t\nabla_\theta r(s_t,a_t;\theta)\right] - \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t\geq 0}\gamma^t\nabla_\theta r(s_t,a_t;\theta)\right]. \tag{6}$$

To obtain stochastic estimators of the exact gradient $\nabla L(\theta_k)$, we take two approximation steps: 1) approximate the optimal policy π_{θ_k} by π_{k+1} in (5), since the optimal policy π_{θ_k} is not available throughout the algorithm; 2) sample one expert trajectory τ_k^{E} which is already generated by the expert policy π^{E}; 3) sample one trajectory τ_k^{A} from the current policy π_{k+1}.

Following the approximation steps mentioned above, we construct a stochastic estimator g_k to approximate the exact gradient $\nabla L(\theta_k)$ in (6) as follows:

$$g_k := h(\theta_k;\tau_k^{\text{E}}) - h(\theta_k;\tau_k^{\text{A}}), \quad \text{where} \quad h(\theta;\tau) := \sum_{t\geq 0}\gamma^t\nabla_\theta r(s_t,a_t;\theta). \tag{7}$$

With the stochastic gradient estimator g_k, the reward parameter θ_k is updated as:

$$\theta_{k+1} = \theta_k + \alpha g_k. \tag{8}$$

where α is the stepsize in updating the reward parameter.

In summary, the proposed algorithm for solving the ML-IRL problem (ML-IRL) is given in Alg. 1.

5 Theoretical Analysis

In this section, we present finite-time guarantees for the proposed algorithm.

To begin with, first recall that in Sec. 3, we have mentioned that (ML-IRL) is a bi-level problem, where the upper level (resp. the lower level) problem optimizes the reward parameter (resp. the policy). In order to solve (ML-IRL), our algorithm 1 has a *singe-loop* structure, which alternates between one step of policy update and one step of the reward parameter update. Such a single-loop structure indeed has computational benefit, but it also leads to potential unstableness, since the lower level problem can stay far away from its true solutions. Specifically, at each iteration k, the potential unstableness is induced by the distribution mismatch between the policy π_{k+1} and π_{θ_k}, when we use estimator g_k (7) to approximate the exact gradient $\nabla L(\theta_k)$ (6) in updating the reward parameter θ_k.

Towards stabilizing the algorithm, we adopt the so-called *two-timescale* stochastic approximation (TTSA) approach [33, 23], where the lower-level problem updates in a faster time-scale (i.e., converges faster) compared with its upper-level counterpart. Intuitively, the TTSA enables the π_{k+1} tracks the optimal π_{θ_k}, leading to a stable algorithm. In the proposed Algorithm 1, the policy (lower-level variable) is continuously updated by the soft policy iteration (5), and it is 'fast' because it converges linearly to the optimal policy under a fixed reward function [28, Theorem 1]. On the other hand, the reward parameter update (8) does not have such linear convergence property, therefore it works in a 'slow' timescale. To begin our analysis, let us first present a few technical assumptions.

Assumption 1 (Ergodicity)**.** *For any policy π, assume the Markov chain with transition kernel $\mathcal{P}$ is irreducible and aperiodic under policy π. Then there exist constants $\kappa > 0$ and $\rho \in (0,1)$ such that*

$$\sup_{s\in\mathcal{S}} \|\mathbb{P}(s_t \in \cdot | s_0 = s, \pi) - \mu_\pi(\cdot)\|_{TV} \leq \kappa\rho^t, \quad \forall\, t \geq 0$$

where $\|\cdot\|_{TV}$ is the total variation (TV) norm; μ_π is the stationary state distribution under π.

Assumption 1 assumes the Markov chain mixes at a geometric rate. It is a common assumption in the iterature of RL [34, 35, 32], which holds for any time-homogeneous Markov chain with finite-state space or any uniformly ergodic Markov chain with general state space.

Assumption 2. *For any $s \in \mathcal{S}$, $a \in \mathcal{A}$ and any reward parameter θ, the following holds:*

$$\big\|\nabla_\theta r(s,a;\theta)\big\| \leq L_r, \tag{9a}$$

$$\big\|\nabla_\theta r(s,a;\theta_1) - \nabla_\theta r(s,a;\theta_2)\big\| \leq L_g\|\theta_1 - \theta_2\| \tag{9b}$$

where L_r and L_g are positive constants.

Assumption 2 assumes that the parameterized reward function has bounded gradient and is Lipschitz smooth. Such assumption in Lipschitz property are common in the literature of min-max / bi-level optimization [36, 23, 37, 25, 38].

Based on Assumptions 1 - 2, we next provide the following Lipschitz properties:

Lemma 2. *Suppose Assumptions 1 - 2 hold. For any reward parameter θ_1 and θ_2, the following results hold:*

$$|Q^{\textit{soft}}_{r_{\theta_1},\pi_{\theta_1}}(s,a) - Q^{\textit{soft}}_{r_{\theta_2},\pi_{\theta_2}}(s,a)| \leq L_q\|\theta_1 - \theta_2\|, \quad \forall s \in \mathcal{S}, a \in \mathcal{A} \tag{10a}$$

$$\|\nabla L(\theta_1) - \nabla L(\theta_2)\| \leq L_c\|\theta_1 - \theta_2\| \tag{10b}$$

where $Q^{\textit{soft}}_{r_\theta,\pi_\theta}(\cdot,\cdot)$ denotes the soft Q-function under the reward function $r(\cdot,\cdot;\theta)$ and the policy π_θ. The positive constants L_q and L_c are defined in Appendix E.

The Lipschitz properties identified in Lemma 2 are vital for the convergence analysis. Then we present the main results, which show the convergence speed of the policy $\{\pi_k\}_{k\geq 0}$ and the reward parameter $\{\theta_k\}_{k\geq 0}$ in the Alg. 1. Please see Appendix E for the detailed proof.

Theorem 2. *Suppose Assumptions 1 - 2 hold. Selecting stepsize $\alpha := \frac{\alpha_0}{K^\sigma}$ for the reward update step* (8) *where $\alpha_0 > 0$ and $\sigma \in (0,1)$ are some fixed constants, and K is the total number of iterations to be run by the algorithm. Then the following result holds:*

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\log\pi_{k+1} - \log\pi_{\theta_k}\right\|_\infty\right] = \mathcal{O}(K^{-1}) + \mathcal{O}(K^{-\sigma}) \tag{11a}$$

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla L(\theta_k)\|^2\right] = \mathcal{O}(K^{-\sigma}) + \mathcal{O}(K^{-1+\sigma}) + \mathcal{O}(K^{-1}) \tag{11b}$$

where we denote $\|\log\pi_{k+1} - \log\pi_{\theta_k}\|_\infty := \max_{s\in\mathcal{S},a\in\mathcal{A}}\left|\log\pi_{k+1}(a|s) - \log\pi_{\theta_k}(a|s)\right|$. In particular, setting $\sigma = 1/2$, then both quantities in (11a) *and* (11b) *converge with the rate $\mathcal{O}(K^{-1/2})$.*

In Theorem 2, we present the finite-time guarantee for the convergence of the Alg.1. Moreover, as a special case, when the reward is parameterized as a linear function, we know that (ML-IRL) is concave and Theorem 2 provides a stronger guarantee which identify the global optimal reward estimator in finite time.

We provide a proof sketch below to present the key steps. The detailed proof is in Appendix H.

Proof sketch. We outline our main steps in analyzing (11a) and (11b) respectively.

In order to show the convergence of policy estimates in (11a), there are several key steps. First, we note that both policies π_{k+1} and π_{θ_k} are in the softmax parameterization, where $\pi_{k+1}(\cdot|s) \propto \exp\left(Q^{\rm soft}_{r_{\theta_k},\pi_k}(s,\cdot)\right)$ and $\pi_{\theta_k}(\cdot|s) \propto \exp\left(Q^{\rm soft}_{r_{\theta_k},\pi_{\theta_k}}(s,\cdot)\right)$. Then, we can show a Lipschitz continuity property between the policy and the soft Q-function:

$$\|\log\pi_{k+1} - \log\pi_{\theta_k}\|_\infty \le 2\|Q^{\rm soft}_{r_{\theta_k},\pi_k} - Q^{\rm soft}_{r_{\theta_k},\pi_{\theta_k}}\|_\infty,$$

where the infinity norm $\|\cdot\|_\infty$ is defined over the state-action space $\mathcal{S}\times\mathcal{A}$. Moreover, by analyzing the contraction property of the soft policy iteration (5), we bound $\|Q^{\rm soft}_{r_{\theta_k},\pi_k} - Q^{\rm soft}_{r_{\theta_k},\pi_{\theta_k}}\|_\infty$ as:

$$\|Q^{\rm soft}_{r_{\theta_k},\pi_k} - Q^{\rm soft}_{r_{\theta_k},\pi_{\theta_k}}\|_\infty \le \gamma\|Q^{\rm soft}_{r_{\theta_{k-1}},\pi_{k-1}} - Q^{\rm soft}_{r_{\theta_{k-1}},\pi_{\theta_{k-1}}}\|_\infty + 2L_q\|\theta_k - \theta_{k-1}\|.$$

To ensure that the error term $\|\theta_k - \theta_{k-1}\|$ is small, we select the stepsize of reward parameters as $\alpha := \frac{\alpha_0}{K^\sigma}$, where K is the total number of iterations and $\sigma > 0$. Then, by combining previous two steps, we could further show the convergence rate of the policy estimates in (11a).

To prove the convergence of the reward parameters in (11b), we first leverage the Lipschitz smooth property of $L(\theta)$ in (10b). However, one technical challenge in the convergence analysis is how to handle the bias between the gradient estimator g_k defined in (7) and the exact gradient $\nabla L(\theta_k)$. When we construct the gradient estimator g_k in (7), we need to sample trajectories from the current policy π_{k+1} and the expert dataset $\mathcal{D}$. However, according to the expression of $\nabla L(\theta_k)$ in (6), the trajectories are sampled from the optimal policy π_{θ_k} and the dataset $\mathcal{D}$. Hence, there is a distribution mismatch between π_{k+1} and π_{θ_k}. Our key idea is to leverage (11a) to handle this distribution mismatch error, and thus show that the bias between g_k and $\nabla L(\theta_k)$ could be controlled. □

To the best of our knowledge, Theorem 2 is the first non-asymptotic convergence result for IRL with nonlinear reward parameterization.

6 A Discussion over State-Only Reward

In this section we consider the IRL problems modeled by using rewards that are only a function of the state. A lower dimensional representation of the agent's preferences (i.e. in terms only of states as opposed to states *and* actions) is more likely to facilitate counterfactual analysis such as predicting the optimal policy under different environment dynamics and/or learning new tasks. This is because the estimation of preferences which are only defined in terms of states is less sensitive to the specific environment dynamics in the expert's demonstration dataset. Moreover, in application such as healthcare [39] and autonomous driving [40], where simply imitating the expert policy can potentially result in poor performance, since the learner and the expert may have different transition dynamics. Similar points have also been argued in recent works [14, 41–43].

Next, let us briefly discuss how we can understand (ML-IRL) and Alg.1, when the reward is parameterized as a state-only function. First, it turns out that there is an equivalent formulation of (ML-IRL), when the expert trajectories only contain the visited states.

Lemma 3. *Suppose the expert trajectories τ is sampled from a policy π^{E}, and the reward is parameterized as a state-only function $r(s;\theta)$. Then ML-IRL is equivalent to the following:*

$$\min_{\theta} \ \mathbb{E}_{s_0 \sim \eta(\cdot)}\big[V^{soft}_{r_\theta,\pi_\theta}(s_0)\big] - \mathbb{E}_{s_0 \sim \eta(\cdot)}\big[V^{soft}_{r_\theta,\pi^E}(s_0)\big] \tag{12a}$$

$$s.t. \ \ \pi_\theta := \arg\max_{\pi} \ \mathbb{E}_{\pi}\bigg[\sum_{t=0}^{\infty} \gamma^t \bigg(r(s_t;\theta) + \mathcal{H}(\pi(\cdot|s_t))\bigg)\bigg]. \tag{12b}$$

Please see Appendix F for the detailed derivation. Intuitively, the above lemma says that, when dealing with the state-only IRL, (ML-IRL) minimizes the gap between the soft value functions of the optimal policy π_θ and that of the expert policy π^{E}. Moreover, Alg.1 can also be easily implemented with the state-only reward. In fact, the entire algorithm essentially stays the same, and the only change is that $r(s,a;\theta)$ will be replaced by $r(s;\theta)$. In this way, by only using the visited states in the trajectories, one can still compute the stochastic gradient estimator in (7). Therefore, even under the state-only IRL setting where the expert dataset only contains visited states, our formulation and the proposed algorithm still work if we parameterize the reward as a state-only function.

7 Numerical Results

In this section, we test the performance of our algorithm on a diverse collection of RL tasks and environments. In each experiment set, we train algorithms until convergence and average the scores of the trajectories over multiple random seeds. The hyperparameter settings and simulation details are provided in Appendix B.

MuJoCo Tasks For Inverse Reinforcement Learning. In this experiment set, we test the performance of our algorithm on imitating the expert behavior. We consider several high-dimensional robotics control tasks in MuJoCo [44]. Two class of existing algorithms are considered as the comparison baselines: 1) imitation learning algorithms that only learn the policy to imitate the expert, including Behavior Cloning (BC) [45] and Generative Adversarial Imitation Learning (GAIL) [10]; 2) IRL algorithms which learn a reward function *and* a policy simultaneously, including Adversarial Inverse Reinforcement Learning (AIRL) [11], f-IRL [14] and IQ-Learn [2]. To ensure fair comparison, all imitation learning / IRL algorithms use soft Actor-Critic [22] as the base RL algorithm. For the expert dataset, we use the data provided in the official implementation[2] of f-IRL.

In this experiment, we implement two versions of our proposed algorithm: ML-IRL(State-Action) where the reward is parameterized as a function of state and action; ML-IRL(State-Only) which utilizes the state-only reward function. In Table 1, we present the simulation results under a limited data regime where the expert dataset only contains a single expert trajectory. The scores (cumulative rewards) reported in the table is averaged over 6 random seeds. In each random seed, we train algorithm from initialization and collect 20 trajectories to average their cumulative rewards after the algorithms converge. The results reported in Table 1 show that our proposed algorithms outperform the baselines. The numerical results with confidence intervals are in Table 3 (See Appendix).

We observe that BC fails to imitate the expert's behavior. It is due to the fact that BC is based on supervised learning and thus could not learn a good policy under such a limited data regime. Moreover, we notice the training of IQ-Learn is unstable, which may be due to its inaccurate approximation to the soft Q-function. Therefore, in the MuJoCo tasks where IQ-Learn does not perform well, so that we cannot match the results presented in the original paper [2], we directly report results from there (and mark them by $*$ in Table 1). The results of AIRL are not presented in Table 1 since it performs poorly even after spending significant efforts in parameter tuning (similar observations have been made in in [46, 14]).

Transfer Learning Across Changing Dynamics. We further evaluate IRL algorithms on the transfer learning setting. We follows the environment setup in [11], where two environments with different dynamics are considered: `Custom-Ant` vs `Disabled-Ant`. We compare ML-IRL(State-Only) with several existing IRL methods: 1) AIRL [11], 2) f-IRL [14]; 3) IQ-Learn [2].

[2] `https://github.com/twni2016/f-IRL`

Task	BC	GAIL	IQ-Learn	f-IRL	ML-IRL (State-Only)	ML-IRL (State-Action)	Expert
Hopper	20.49	2815.59	2981.01	3074.55	3089.79	**3121.68**	3592.63
Half-Cheetah	-1.87	3301.52	4175.88	4375.88	**4472.85**	4086.92	5098.3
Walker	-14.01	1112.79	3961.42	4464.20	4380.17	**4504.88**	5344.21
Ant	760.46	1154.27	4362.90*	4571.71	4675.34	**4984.34**	5926.18
Humanoid	78.48	3016.40	5227.10*	5243.90	**5390.31**	5240.57	5351.08

Table 1: **MuJoCo Results.** The performance of benchmark algorithms under a single expert trajectory.

We consider two transfer learning settings: 1) data transfer; 2) reward transfer. For both settings, the expert dataset / trajectories are generated in `Custom-Ant`. In the data transfer setting, we train IRL agents in `Disabled-Ant` by using the expert trajectories, which are generated in `Custom-Ant`. In the reward transfer setting, we first use IRL algorithms to infer the reward functions in `Custom-Ant`, and then transfer these recovered reward functions to `Disabled-Ant` for further evaluation. In both settings, we also train SAC with the ground-truth reward in `Disabled-Ant` and report the scores.

The numerical results are reoprted in Table 2. the proposed ML-IRL(State-Only) achieves superior performance compared with the existing IRL benchmarks in both settings. We notice that IQ-Learn fails in both settings since it indirectly recovers the reward function from a soft Q-function approximator, which could be inaccurate and is highly dependent upon the environment dynamics. Therefore, the reward function recovered by IQ-Learn can not be disentangled from the expert actions and environment dynamics, which leads to its failures in the transfer learning tasks.

Setting	IQ-Learn	AIRL	f-IRL	ML-IRL(State-Only)	Groud-Truth
Data Transfer	-11.78	-5.39	188.85	**221.51**	320.15
Reward Transfer	-1.04	130.3	156.45	**187.69**	320.15

Table 2: **Transfer Learning.** The performance of benchmark algorithms under a single expert trajectory. The scores in the table are obtained similarly as in Table 1.

8 Conclusion

In this paper, we present a maximum likelihood IRL formulation and propose a provably efficient algorithm with a single-loop structure. To our knowledge, we provide the first non-asymptotic analysis for IRL algorithm under nonlinear reward parameterization. As a by-product, when we parameterize the reward as a state-only function, our algorithm could work in state-only IRL setting and enable reward transfer to new environments with different dynamics. Our algorithm outperforms existing IRL methods on high-dimensional robotics control tasks and corresponding transfer learning settings. A limitation of our method is the requirement for online training, so one future direction of this work is to further extend our algorithm and the theoretical analysis to the offline IRL setting.

Potential Negative Social Impacts

Since IRL methods aim to recover the reward function and the associated optimal policy from the observed expert dataset, potential negative social impacts may occur if there are bad demonstrations included in the expert dataset. Thus, for sensitive applications such as autonomous driving and clinical decision support, additional care should be taken to avoid negative biases from the expert demonstrations and ensure safe adaptation.

Acknowledgments

We thank the anonymous reviewers for their valuable comments. M. Hong and S. Zeng are partially supported by NSF grants CIF-1910385, CMMI-1727757, and AFOSR grant 19RT0424. A. Garcia would like to acknowledge partial support from grant FA9550-19-1-00347 by AFOSR.

References

[1] S. Reddy, A. D. Dragan, and S. Levine, "SQIL: Imitation learning via reinforcement learning with sparse rewards," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=S1xKd24twB

[2] D. Garg, S. Chakraborty, C. Cundy, J. Song, and S. Ermon, "Iq-learn: Inverse soft-q learning for imitation," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[3] T. Osa, J. Pajarinen, G. Neumann, J. A. Bagnell, P. Abbeel, and J. Peters, *An Algorithmic Perspective on Imitation Learning*, ser. Foundations and Trends in Robotics, 2018, vol. 7.

[4] K. Kim, S. Garg, K. Shiragur, and S. Ermon, "Reward identification in inverse reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5496–5505.

[5] B. D. Ziebart, A. L. Maas, J. A. Bagnell, A. K. Dey *et al.*, "Maximum entropy inverse reinforcement learning." in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 8. Chicago, IL, USA, 2008, pp. 1433–1438.

[6] B. D. Ziebart, J. A. Bagnell, and A. K. Dey, "Modeling interaction via the principle of maximum causal entropy," in *International Conference on Machine Learning*, 2010.

[7] M. Wulfmeier, P. Ondruska, and I. Posner, "Maximum entropy deep inverse reinforcement learning," *arXiv preprint arXiv:1507.04888*, 2015.

[8] C. Finn, P. Christiano, P. Abbeel, and S. Levine, "A connection between generative adversarial networks, inverse reinforcement learning, and energy-based models," *arXiv preprint arXiv:1611.03852*, 2016.

[9] C. Finn, S. Levine, and P. Abbeel, "Guided cost learning: Deep inverse optimal control via policy optimization," in *International Conference on Machine Learning*. PMLR, 2016, pp. 49–58.

[10] J. Ho and S. Ermon, "Generative adversarial imitation learning," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[11] J. Fu, K. Luo, and S. Levine, "Learning robust rewards with adversarial inverse reinforcement learning," *arXiv preprint arXiv:1710.11248*, 2017.

[12] M. Chen, Y. Wang, T. Liu, Z. Yang, X. Li, Z. Wang, and T. Zhao, "On computation and generalization of generative adversarial imitation learning," *arXiv preprint arXiv:2001.02792*, 2020.

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[14] T. Ni, H. Sikchi, Y. Wang, T. Gupta, L. Lee, and B. Eysenbach, "f-irl: Inverse reinforcement learning via state marginal matching," *arXiv preprint arXiv:2011.04709*, 2020.

[15] K. Kurach, M. Lucic, X. Zhai, M. Michalski, and S. Gelly, "The gan landscape: Losses, architectures, regularization, and normalization," 2018.

[16] I. Kostrikov, K. K. Agrawal, D. Dwibedi, S. Levine, and J. Tompson, "Discriminator-actor-critic: Addressing sample inefficiency and reward bias in adversarial imitation learning," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Hk4fpoA5Km

[17] E. T. Jaynes, "Information theory and statistical mechanics," *Physical review*, vol. 106, no. 4, p. 620, 1957.

[18] B. D. Ziebart, J. A. Bagnell, and A. K. Dey, "The principle of maximum causal entropy for estimating interacting processes," *IEEE Transactions on Information Theory*, vol. 59, no. 4, pp. 1966–1980, 2013.

[19] M. Bloem and N. Bambos, "Infinite time horizon maximum causal entropy inverse reinforcement learning," in *53rd IEEE conference on decision and control*. IEEE, 2014, pp. 4911–4916.

[20] Z. Zhou, M. Bloem, and N. Bambos, "Infinite time horizon maximum causal entropy inverse reinforcement learning," *IEEE Transactions on Automatic Control*, vol. 63, no. 9, pp. 2787–2802, 2017.

[21] T. Haarnoja, H. Tang, P. Abbeel, and S. Levine, "Reinforcement learning with deep energy-based policies," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1352–1361.

[22] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International Conference on Machine Learning*. PMLR, 2018, pp. 1861–1870.

[23] M. Hong, H.-T. Wai, Z. Wang, and Z. Yang, "A two-timescale framework for bilevel optimization: Complexity analysis and application to actor-critic," *arXiv preprint arXiv:2007.05170*, 2020.

[24] K. Ji, J. Yang, and Y. Liang, "Bilevel optimization: Convergence analysis and enhanced design," in *International Conference on Machine Learning*. PMLR, 2021, pp. 4882–4892.

[25] P. Khanduri, S. Zeng, M. Hong, H.-T. Wai, Z. Wang, and Z. Yang, "A near-optimal algorithm for stochastic bilevel optimization via double-momentum," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[26] V. Jain, P. Doshi, and B. Banerjee, "Model-free irl using maximum likelihood estimation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 3951–3958.

[27] N. Sanghvi, S. Usami, M. Sharma, J. Groeger, and K. Kitani, "Inverse reinforcement learning with explicit policy estimates," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 11, 2021, pp. 9472–9480.

[28] S. Cen, C. Cheng, Y. Chen, Y. Wei, and Y. Chi, "Fast global convergence of natural policy gradient methods with entropy regularization," *Operations Research*, 2021.

[29] S. Cayci, N. He, and R. Srikant, "Linear convergence of entropy-regularized natural policy gradient with linear function approximation," *arXiv preprint arXiv:2106.04096*, 2021.

[30] O. Nachum, M. Norouzi, K. Xu, and D. Schuurmans, "Bridging the gap between value and policy based reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[31] V. Konda and J. Tsitsiklis, "Actor-critic algorithms," *Advances in Neural Information Processing Systems*, vol. 12, 1999.

[32] Y. F. Wu, W. Zhang, P. Xu, and Q. Gu, "A finite-time analysis of two time-scale actor-critic methods," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 617–17 628, 2020.

[33] V. S. Borkar, "Stochastic approximation with two time scales," *Systems & Control Letters*, vol. 29, no. 5, pp. 291–294, 1997.

[34] J. Bhandari, D. Russo, and R. Singal, "A finite time analysis of temporal difference learning with linear function approximation," in *Conference on learning theory*. PMLR, 2018, pp. 1691–1692.

[35] S. Zou, T. Xu, and Y. Liang, "Finite-sample analysis for sarsa with linear function approximation," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[36] C. Jin, P. Netrapalli, and M. Jordan, "What is local optimality in nonconvex-nonconcave minimax optimization?" in *International Conference on Machine Learning*. PMLR, 2020, pp. 4880–4889.

[37] Z. Guan, T. Xu, and Y. Liang, "When will generative adversarial imitation learning algorithms attain global convergence," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 1117–1125.

[38] T. Chen, Y. Sun, and W. Yin, "Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems," *Advances in Neural Information Processing Systems*, vol. 34, pp. 25 294–25 307, 2021.

[39] C. Yu, J. Liu, S. Nemati, and G. Yin, "Reinforcement learning in healthcare: A survey," *ACM Computing Surveys (CSUR)*, vol. 55, no. 1, pp. 1–36, 2021.

[40] B. R. Kiran, I. Sobh, V. Talpaert, P. Mannion, A. A. Al Sallab, S. Yogamani, and P. Pérez, "Deep reinforcement learning for autonomous driving: A survey," *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[41] T. Gangwani and J. Peng, "State-only imitation with transition dynamics mismatch," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=HJgLLyrYwB

[42] L. Viano, Y.-T. Huang, P. Kamalaruban, A. Weller, and V. Cevher, "Robust inverse reinforcement learning under transition dynamics mismatch," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[43] F. Torabi, G. Warnell, and P. Stone, "Generative adversarial imitation from observation," *arXiv preprint arXiv:1807.06158*, 2018.

[44] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *2012 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 2012, pp. 5026–5033.

[45] D. A. Pomerleau, "ALVINN: An autonomous land vehicle in a neural network," *Advances in Neural Information Processing Systems*, vol. 1, 1988.

[46] F. Liu, Z. Ling, T. Mu, and H. Su, "State alignment-based imitation learning," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=rylrdxHFDr

[47] A. Kamoutsi, G. Banjac, and J. Lygeros, "Efficient performance bounds for primal-dual reinforcement learning from demonstrations," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5257–5268.

[48] H. Cao, S. Cohen, and L. Szpruch, "Identifiability in inverse reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[49] R. Wang, C. Ciliberto, P. V. Amadori, and Y. Demiris, "Random expert distillation: Imitation learning via expert policy support estimation," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6536–6544.

[50] D. S. Brown, W. Goo, and S. Niekum, "Better-than-demonstrator imitation learning via automatically-ranked demonstrations," in *Conference on robot learning*. PMLR, 2020, pp. 330–359.

[51] P. Barde, J. Roy, W. Jeon, J. Pineau, C. Pal, and D. Nowrouzezahrai, "Adversarial soft advantage fitting: Imitation learning without policy optimization," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 334–12 344, 2020.

[52] H. Kretzschmar, M. Spies, C. Sprunk, and W. Burgard, "Socially compliant mobile robot navigation via inverse reinforcement learning," *The International Journal of Robotics Research*, vol. 35, no. 11, pp. 1289–1307, 2016.

[53] C. Xia and A. El Kamel, "Neural inverse reinforcement learning in autonomous navigation," *Robotics and Autonomous Systems*, vol. 84, pp. 1–14, 2016.

[54] J. Rust and C. Phelan, "How social security and medicare affect retirement behavior in a world of incomplete markets," *Econometrica: Journal of the Econometric Society*, pp. 781–831, 1997.

[55] Z. Eckstein and K. I. Wolpin, “Why youths drop out of high school: The impact of preferences, opportunities, and abilities,” *Econometrica*, vol. 67, no. 6, pp. 1295–1339, 1999.

[56] E. Duflo, R. Hanna, and S. P. Ryan, “Incentives work: Getting teachers to come to school,” *American Economic Review*, vol. 102, no. 4, pp. 1241–78, 2012.

[57] H. Fang and Y. Wang, “Estimating dynamic discrete choice models with hyperbolic discounting, with an application to mammography decisions,” *International Economic Review*, vol. 56, no. 2, pp. 565–596, 2015.

[58] L. Caliendo, M. Dvorkin, and F. Parro, “Trade and labor market dynamics: General equilibrium analysis of the china trade shock,” *Econometrica*, vol. 87, no. 3, pp. 741–835, 2019.

[59] C. Cirillo, R. Xu, and F. Bastin, “A dynamic formulation for car ownership modeling,” *Transportation Science*, vol. 50, no. 1, pp. 322–335, 2016.

[60] T. Xu, Z. Wang, and Y. Liang, “Improving sample complexity bounds for (natural) actor-critic algorithms,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 4358–4369, 2020.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
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)? [No]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (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]
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? [Yes]
(b) Did you mention the license of the assets? [Yes]
(c) Did you include any new assets either in the supplemental material or as a URL? [No]
(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]

Appendix

A Related Works

Under the maximum entropy framework, IRL algorithms [7, 9] are proposed to learn the nonlinear structure of reward function. In general, these works [5, 7, 9] recover the reward function through minimizing forward KL divergence in trajectory space. There is one branch of IRL methods which train a generative adversial netowrk (GAN) [13] to learn the reward function through designing a specific structure in the discriminator network. In [8], the authors reveal a connection between GAN and guided cost learning (GCL) [9]. Observing the Lagrangian function of the MaxEnt-IRL [5, 18, 20] is a min-max problem with convex-nonconcave structure, [12, 37] first swap the optimization order of the reward parameter and the policy parameter and further regularize the reward parameter for analyzing the convergence of such constructed nonconcave strongly-convex optimization problem. In [47], in order to provide a non-asymptotic analysis of the IRL problem, the authors introduce a bilinear saddlepoint framework through using Lagrangian duality.

Under the MaxEnt-IRL framework, there is a line of works focusing on disentangling the reward function from the environment dynamics, so that the recovered reward functions could be transferred across environments with different dynamics. In [11], the authors propose an algorithm called adversarial inverse reinforcement learning (AIRL). Constructing the estimated reward as a function which depends on the current state and next state, AIRL enables the agent to learn policies in a new environment through leveraging the estimated reward function recovered from the training environment. In [4], the authors prove necessary and sufficient conditions for reward identifiability in deterministic MDP models with the maximum entropy reinforcement learning objective. In [48], the authors present a theoretical analysis to show the necessary and sufficient condition to identify an action-independent time-homogeneous reward function under MaxEnt-IRL.

A recent line of works consider a more challenging setting, where the learner has no access to the expert environment, and there is a transition dynamics mismatch between the expert and the learner. In [46], the authors propose a state alignment based imitation learning method so that the imitator could follow the state sequences in expert demonstrations as much as possible. Arguing that the expert actions are not efficient demonstrations under transition dynamics mismatch, [41] further develops a state-only imitation learning method. In [42], the authors revisit the Maximum Causal Entropy IRL when there is a transition dynamics mismatch between the expert and the learner. A theoretical analysis is further provided in [42] to show the upper bound on the learner's performance degradation, which is measured in terms of the ℓ_1-distance between the transition dynamics of the expert and the learner.

We would like to further introduce several interesting works which seek to make imitating expert policy more tractable. In [49], the authors utilize random network distillation and propose a new general framework of imitation learning via expert policy support estimation. In [50], a ranking-based imitation learning method is proposed and the authors show that such imitation learning method could outperform the demonstrator. Witnessing the instability of adversarial training, [51] proposes an imitation learning method without performing any policy optimization steps.

In the end, we introduce the wide applications of inverse reinforcement learning. The problem of inverse reinforcement learning IRL has been widely in studied by the robotics and artificial intelligence research communities [52, 53]. It has also been applied (under the label of dynamic discrete choice estimation) in a wide variety of application domains including modeling of employee retirement decisions [54], occupational choices and career decisions of young professionals [55], incentives to get teachers to work [56], adult women's mammography decisions [57], trade and labor markets [58], car ownership [59]. The techniques developed in the present paper will enable new applications to settings with high dimensional state space.

B Experiment Details

B.1 MuJoCo Tasks For Inverse Reinforcement Learning.

In all experiments, we test the performance of benchmark algorithms on `Hopper, Half-Cheetah, Walker, Ant, Humanoid` environments from OpenAI Gym. To ensure fair comparison, we use

Algorithm 2 *Practical Implementation of ML-IRL*

Input: Initialize reward parameter θ_0 and policy π_0. Set the reward parameter's stepsize as α.
Data Preparation: Collect a dataset $\mathcal{D}$ which contains multiple expert trajectories
for $k = 0, 1, \ldots, K-1$ **do**
  Policy Update: $\pi_{k+1} \leftarrow$ several SAC steps under reward function $r(\cdot, \cdot; \theta_k)$ and policy π_k.
  Data Sampling I: Sampling expert trajectory $\tau_k^{\mathrm{E}} := \{s_t, a_t\}_{t \geq 0}$ from the dataset $\mathcal{D}$
  Data Sampling II: Sampling agent trajectory $\tau_k^{\mathrm{A}} := \{s_t, a_t\}_{t \geq 0}$ from the policy π_{k+1}
  Estimating Gradient: $g_k := h(\theta_k; \tau_k^{\mathrm{E}}) - h(\theta_k; \tau_k^{\mathrm{A}})$ where $h(\theta; \tau) := \sum_{t \geq 0} \gamma^t \nabla_\theta r(s_t, a_t; \theta)$
  Reward Parameter Update: $\theta_{k+1} := \theta_k + \alpha g_k$
end for

an open-source implementation[3] of SAC as the base RL algorithm for all imitation learning / IRL methods. Moreover, Adam is used as the optimizer in SAC.

In SAC, both policy network and Q-network are $(64, 64)$ MLPs with ReLU activation function, and we set their stepsizes as 3×10^{-3}. Moreover, in our proposed algorithms, we parameterize the reward function by a $(64, 64)$ MLPs with ReLU activation function. For the reward network, we use Adam as the optimizer and the stepsize is set to be 1×10^{-4}.

We present the practical implementation procedure of our proposed algorithm in Table.2. At each iteration, we first warm-start both policy network and Q-network in SAC by using the trained neural networks from the previous iteration. Then, we run 10 episodes in the corresponding MuJoCo environment to train the policy network and Q-network in SAC. After that, we sample 5 agent trajectories and expert trajectories to construct the reward gradient estimator, and then update the reward network by a gradient update.

For the imitation learning / IRL benchmark algorithms, we use their open-source implementations in our experiments. The official implementations of f-IRL is provided in `https://github.com/twni2016/f-IRL`. The offical code base for IQ-Learn is provided in `https://github.com/Div99/IQ-Learn`. For the remaining benchmarks including BC, GAIL and AIRL, we refer to a open-source implementation: `https://github.com/KamyarGh/rl_swiss`.

B.2 Transfer Learning Across Changing Dynamics.

In this experiment, we follow the setup in [11]. A standard ant (`Custom-Ant`) and an ant with two disabled legs (`Disabled-Ant`) are simulated in MuJoCo. For all benchmark algorithms tested in this experiment, we follow same network structure and hyperparameter settings described in Section B.1.

[3] `https://github.com/openai/spinningup`

Here, we provide a supplementary experiment result to show the performance of benchmark algorithms under different number of expert trajectory. The performance of AIRL and IQ-Learn is not presented in Table 3 since we found their training is unstable (as we have mentioned in Sec.7). The scores in Hopper are recorded after 1×10^6 environment steps and the scores in other environments are recorded after 2×10^6 environment steps. The scores are reported after 6 independent Monte Carlo (MC) trials for each algorithm.

Task	Hopper		
# Expert Trajectory	1	5	10
Expert Performance	3592.63	3530.63 ± 2.73	3531.72 ± 6.41
BC	20.49 ± 3.24	104.50 ± 59.12	378.42 ± 56.08
GAIL	2815.59 ± 203.80	2840.71 ± 166.36	2941.82 ± 128.34
f-IRL	3074.55 ± 237.03	3118.49 ± 83.25	$\mathbf{3127.05} \pm 103.60$
ML-IRL(State-Only)	3089.79 ± 18.39	3116.55 ± 49.54	3039.91 ± 50.61
ML-IRL(State-Action)	$\mathbf{3121.68} \pm 286.58$	$\mathbf{3200.33} \pm 114.28$	2943.42 ± 271.20

Task	Half-Cheetah		
# Expert Trajectory	1	5	10
Expert Performance	5098.30	5072.53 ± 145.12	5043.02 ± 104.32
BC	-1.87 ± 0.24	145.72 ± 78.45	342.64 ± 154.97
GAIL	3301.52 ± 243.79	3465 ± 178.44	3400.05 ± 115.92
f-IRL	4375.88 ± 230.41	4438.96 ± 365.20	4427.91 ± 426.64
ML-IRL(State-Only)	$\mathbf{4472.85} \pm 166.61$	$\mathbf{4669.85} \pm 483.51$	$\mathbf{4653.11} \pm 401.04$
ML-IRL(State-Action)	4086.92 ± 195.30	4338.35 ± 107.07	4416.79 ± 61.35

Task	Walker		
# Expert Trajectory	1	5	10
Expert Performance	5344.21	5471.58 ± 13.93	5471.70 ± 10.59
BC	-14.01 ± 0.24	285.43 ± 3.24	450.71 ± 72.08
GAIL	1112.79 ± 143.81	3523.42 ± 487.70	3789.40 ± 61.62
f-IRL	4464.20 ± 332.49	$\mathbf{4507.58} \pm 254.13$	4164.77 ± 648.66
ML-IRL(State-Only)	4380.17 ± 393.90	4402.41 ± 491.72	$\mathbf{4541.04} \pm 272.54$
ML-IRL(State-Action)	$\mathbf{4504.88} \pm 120.82$	4027.75 ± 478.39	4114.00 ± 732.54

Task	Ant		
# Expert Trajectory	1	5	10
Expert Performance	5926.18	5856.84 ± 79.48	5901.73 ± 122.40
BC	760.46 ± 0.59	942.58 ± 14.87	1032.70 ± 73.94
GAIL	1154.27 ± 422.23	3042.55 ± 403.72	4233.51 ± 280.51
f-IRL	4571.71 ± 169.97	4681.54 ± 221.83	4574.20 ± 303.49
ML-IRL(State-Only)	4675.34 ± 26.03	4632.38 ± 77.90	4498.31 ± 212.81
ML-IRL(State-Action)	$\mathbf{4984.34} \pm 177.97$	$\mathbf{5190.48} \pm 159.65$	$\mathbf{5011.67} \pm 252.75$

Task	Humanoid		
# Expert Trajectory	1	5	10
Expert Performance	5351.08	5339.12 ± 22.21	5343.76 ± 24.19
BC	78.48 ± 0.34	532.61 ± 48.74	582.78 ± 18.82
GAIL	3016.40 ± 714.26	3030.25 ± 594.38	3204.10 ± 358.88
f-IRL	5243.90 ± 163.64	5338.20 ± 122.219	$\mathbf{5472.72} \pm 117.21$
ML-IRL(State-Only)	$\mathbf{5390.31} \pm 161.98$	$\mathbf{5431.23} \pm 244.17$	5257.24 ± 323.61
ML-IRL(State-Action)	5240.57 ± 72.18	5251.86 ± 218.98	5342.98 ± 608.32

Table 3: **MuJoCo Results.** The performance versus different number of expert trajectory.

C Auxiliary Lemmas

Throughout this section, we assume Assumptions 1 - 2 hold true.

Lemma 4. *([60, Lemma 3]) Consider the initialization distribution $\eta(\cdot)$ and transition kernel $\mathcal{P}(\cdot|s,a)$. Under $\eta(\cdot)$ and $\mathcal{P}(\cdot|s,a)$, denote $d_w(\cdot,\cdot)$ as the state-action visitation distribution of MDP with the Boltzman policy parameterized by parameter w. Suppose Assumption 1 holds, for all policy parameter w and w', we have*

$$\|d_w(\cdot,\cdot) - d_{w'}(\cdot,\cdot)\|_{TV} \leq C_d\|w - w'\| \tag{13}$$

where C_d is a positive constant.

Lemma 5. *([21, Theorem 4]) Under a reward function $r(\cdot,\cdot)$, given a policy π, we define a new policy $\tilde{\pi}$ as*

$$\tilde{\pi}(\cdot|s) \propto \exp\left(Q^{soft}_{r,\pi}(s,\cdot)\right), \ \forall\, s \in \mathcal{S}.$$

For any $s \in \mathcal{S}, a \in \mathcal{A}$, it holds that $Q^{soft}_{r,\tilde{\pi}}(s,a) \geq Q^{soft}_{r,\pi}(s,a)$.

Next, in order to facilitate analysis for entropy-regularized MDPs, we introduce a "soft" Bellman optimality operator $\mathcal{T} : \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|} \to \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ as follows:

$$\mathcal{T}(Q)(s,a) := r(s,a) + \gamma\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)}\left[\max_{\pi(\cdot|s)}\ \mathbb{E}_{a'\sim\pi(\cdot|s')}\left[Q(s',a') - \log\pi(a'|s')\right]\right]. \tag{14}$$

In the following lemma, the properties of entropy-regularized MDPs are characterized.

Lemma 6. *([28, Lemma 2]) The operator $\mathcal{T}$ as defined in* (14) *satisfies the properties below:*

- *$\mathcal{T}$ has the following closed-form expression:*

$$\mathcal{T}(Q)(s,a) = r(s,a) + \gamma\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)}\left[\log\left(\sum_{a'}\exp\left(Q(s',a')\right)\right)\right]. \tag{15}$$

- *$\mathcal{T}$ is a γ-contraction in the ℓ_∞ norm, namely, for any $Q_1, Q_2 \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, it holds that*

$$\|\mathcal{T}(Q_1) - \mathcal{T}(Q_2)\|_\infty \leq \gamma\|Q_1 - Q_2\|_\infty. \tag{16}$$

- *Under a given reward function $r(\cdot,\cdot)$, the corresponding optimal soft Q-function Q^{soft}_{r,π^*} is a unique fixed point of the operator $\mathcal{T}$, namely,*

$$\mathcal{T}(Q^{soft}_{r,\pi^*}) = Q^{soft}_{r,\pi^*} \tag{17}$$

Proof. This Lemma is proved in [28, Lemma]. We refine its analysis as below.

We first show that

$$\mathbb{E}_{a\sim\pi(\cdot|s)}\Big[Q(s,a) - \log\pi(a|s)\Big] = \sum_a \pi(a|s)\log\left(\frac{\exp(Q(s,a))}{\pi(a|s)}\right) \overset{(i)}{\leq} \log\left(\sum_a \exp\left(Q(s,a)\right)\right) \tag{18}$$

where (i) is from Jensen's inequality. Moreover, the equality between both sides of (i) holds when the policy π has the expression $\pi(\cdot|s) \propto \exp(Q(s,\cdot))$. Therefore, through applying the inequality (18) to (14), it obtains that

$$\mathcal{T}(Q)(s,a) = r(s,a) + \gamma\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)}\left[\log\left(\sum_{a'}\exp\left(Q(s',a')\right)\right)\right], \tag{19}$$

which proves the equality (15).

We define $\|Q_1 - Q_2\|_\infty := \max_{s\in\mathcal{S},a\in\mathcal{A}} |Q_1(s,a) - Q_2(s,a)|$ and $\epsilon = \|Q_1 - Q_2\|_\infty$. Then for any $s \in \mathcal{S}$ and $a \in \mathcal{A}$, it follows that

$$\begin{aligned}\log\left(\sum_a \exp\big(Q_1(s,a)\big)\right) &\leq \log\left(\sum_a \exp\big(Q_2(s,a)+\epsilon\big)\right)\\ &= \log\left(\exp(\epsilon)\sum_a \exp\big(Q_2(s,a)\big)\right)\\ &= \epsilon + \log\left(\sum_a \exp\big(Q_2(s,a)\big)\right)\end{aligned}$$

Similarly, it is easy to obtain that $\log\big(\sum_a \exp\big(Q_1(s,a)\big)\big) \geq -\epsilon + \log\big(\sum_a \exp\big(Q_2(s,a)\big)\big)$. Hence, it leads to the contraction property that

$$\|\mathcal{T}(Q_1) - \mathcal{T}(Q_2)\|_\infty \leq \gamma\epsilon = \gamma\|Q_1 - Q_2\|_\infty \tag{20}$$

which proves the contraction property (16).

Moreover, we have

$$\mathcal{T}(Q^{\text{soft}}_{r,\pi^*})(s,a) \overset{(i)}{=} r(s,a) + \gamma\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)}\left[\log\left(\sum_{a'}\exp\big(Q^{\text{soft}}_{r,\pi^*}(s',a')\big)\right)\right] \overset{(ii)}{=} Q^{\text{soft}}_{r,\pi^*}(s,a) \tag{21}$$

where (i) follows the equality (19). Based on the defition of the soft Q-function $Q^{\text{soft}}_{r,\pi^*}$, we have

$$Q^{\text{soft}}_{r,\pi^*}(s,a) = r(s,a) + \gamma\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)}\big[\mathbb{E}_{a'\sim\pi^*(\cdot|s')}[-\log\pi^*(a'|s') + Q^{\text{soft}}_{r,\pi^*}(s',a')]\big]. \tag{22}$$

We prove the equality (ii) in (21) through combining (22) and the fact that the optimal soft policy has the closed form $\pi^*(\cdot|s) \propto \exp\big(Q^{\text{soft}}_{r,\pi^*}(s,\cdot)\big)$. Suppose two different fixed points of the soft Bellman operator exist, then it contradicts with the contraction property in (20).

Hence, we proved the uniqueness of the optimal soft Q-function $Q^{\text{soft}}_{r,\pi^*}$. Moreover, the optimal soft Q-function $Q^{\text{soft}}_{r,\pi^*}$ is a fixed point to the soft Bellman operator $\mathcal{T}$ in (17). $\square$

Lemma 7. *Suppose Assumption 2 holds. Under an arbitrary policy π, for any $s \in \mathcal{S}$, $a \in \mathcal{A}$ and any reward parameters θ_1 and θ_2, the following inequality holds:*

$$|Q^{\text{soft}}_{r_{\theta_1},\pi}(s,a) - Q^{\text{soft}}_{r_{\theta_2},\pi}(s,a)| \leq L_q\|\theta_1 - \theta_2\|,$$

where $L_q := \frac{L_r}{1-\gamma}$ and L_r is the positive constant in Assumption 2.

Proof. Based on the definition of soft-Q function, we have

$$Q^{\text{soft}}_{r,\pi}(s,a) := r(s,a) + \mathbb{E}_\pi\left[\sum_{t=1}^{\infty}\gamma^t\Big(r(s_t,a_t) + \mathcal{H}(\pi(\cdot|s_t))\Big)\Big|(s_0,a_0) = (s,a)\right].$$

Then it holds that

$$\begin{aligned}&|Q^{\text{soft}}_{r_{\theta_1},\pi}(s,a) - Q^{\text{soft}}_{r_{\theta_2},\pi}(s,a)|\\ &= \left|\mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t\Big(r(s_t,a_t;\theta_1) - r(s_t,a_t;\theta_2)\Big)\right]\right|\\ &\overset{(i)}{\leq} \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t\Big|r(s_t,a_t;\theta_1) - r(s_t,a_t;\theta_2)\Big|\right]\\ &\overset{(ii)}{\leq} \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t\Big\|\max_\theta \nabla_\theta r(s_t,a_t;\theta)\Big\|\cdot\Big\|\theta_1 - \theta_2\Big\|\right]\\ &\overset{(iii)}{\leq} \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t L_r\|\theta_1 - \theta_2\|\right]\\ &= \frac{L_r}{1-\gamma}\|\theta_1 - \theta_2\|\end{aligned} \tag{23}$$

where (i) follows Jensen's inequality; (ii) follows the mean value theorem; (iii) follows inequality (9a) in Assumption 2.

□

D Proof of Lemma 1

Proof. First, we are able to express the objective function $L(\theta)$ in (ML-IRL) as below:

$$L(\theta) := \mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\Big[\sum_{t=0}^{\infty}\gamma^t \log \pi_\theta(a_t|s_t)\Big] \overset{(i)}{=} \mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\Big[\sum_{t=0}^{\infty}\gamma^t \log\Big(\frac{\exp\big(Q^{\text{soft}}_{r_\theta,\pi_\theta}(s_t,a_t)\big)}{\sum_a \exp\big(Q^{\text{soft}}_{r_\theta,\pi_\theta}(s_t,a)\big)}\Big)\Big]$$

where (i) is due to the fact that the optimal policy has the closed form $\pi_\theta(\cdot|s) \propto \exp\big(Q^{\text{soft}}_{r_\theta,\pi_\theta}(s,\cdot)\big)$. Therefore, we could express the objective function in this form:

$$\begin{aligned}
L(\theta) &:= \mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\Big[\sum_{t=0}^{\infty}\gamma^t\Big(Q^{\text{soft}}_{r_\theta,\pi_\theta}(s_t,a_t) - \log\big(\sum_a \exp\big(Q^{\text{soft}}_{r_\theta,\pi_\theta}(s_t,a)\big)\big)\Big)\Big] \\
&\overset{(i)}{=} \mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\Big[\sum_{t=0}^{\infty}\gamma^t\Big(Q^{\text{soft}}_{r_\theta,\pi_\theta}(s_t,a_t) - V^{\text{soft}}_{r_\theta,\pi_\theta}(s_t)\Big)\Big] \\
&= \mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\Big[\sum_{t=0}^{\infty}\gamma^t\Big(r(s_t,a_t;\theta) + \gamma\mathbb{E}_{s_{t+1}\sim\mathcal{P}(\cdot|s_t,a_t)}\big[V^{\text{soft}}_{r_\theta,\pi_\theta}(s_{t+1})\big] - V^{\text{soft}}_{r_\theta,\pi_\theta}(s_t)\Big)\Big] \\
&= \mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\Big[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t;\theta)\Big] + \mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\Big[\sum_{t=1}^{\infty}\gamma^t V^{\text{soft}}_{r_\theta,\pi_\theta}(s_t)\Big] - \mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\Big[\sum_{t=0}^{\infty}\gamma^t V^{\text{soft}}_{r_\theta,\pi_\theta}(s_t)\Big] \\
&= \mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\Big[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t;\theta)\Big] - \mathbb{E}_{s_0\sim\eta(\cdot)}\Big[V^{\text{soft}}_{r_\theta,\pi_\theta}(s_0)\Big] \qquad (24) \\
&\overset{(ii)}{=} \mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\Big[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t;\theta)\Big] - \mathbb{E}_{s_0\sim\eta(\cdot)}\Big[\log\Big(\sum_a \exp\big(Q^{\text{soft}}_{r_\theta,\pi_\theta}(s_0,a)\big)\Big)\Big] \qquad (25)
\end{aligned}$$

where (i) and (ii) follows the fact that the the optimal soft value function could be expressed as $V^{\text{soft}}_{r_\theta,\pi_\theta}(s) = \log\big(\sum_a \exp\big(Q^{\text{soft}}_{r_\theta,\pi_\theta}(s,a)\big)\big)$.

Based on (25), we calculate the exact gradient of the objective function $L(\theta)$ as below:

$$\begin{aligned}
\nabla L(\theta) &:= \mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\Big[\sum_{t=0}^{\infty}\gamma^t \nabla_\theta r(s_t,a_t;\theta)\Big] - \mathbb{E}_{s_0\sim\eta(\cdot)}\Big[\nabla_\theta \log\Big(\sum_a \exp\big(Q^{\text{soft}}_{r_\theta,\pi_\theta}(s_0,a)\big)\Big)\Big] \\
&= \mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\Big[\sum_{t=0}^{\infty}\gamma^t \nabla_\theta r(s_t,a_t;\theta)\Big] - \mathbb{E}_{s_0\sim\eta(\cdot)}\Big[\sum_a\Big(\frac{\exp\big(Q^{\text{soft}}_{r_\theta,\pi_\theta}(s_0,a)\big)}{\sum_{\tilde{a}}\exp\big(Q^{\text{soft}}_{r_\theta,\pi_\theta}(s_0,\tilde{a})\big)}\nabla_\theta Q^{\text{soft}}_{r_\theta,\pi_\theta}(s_0,a)\Big)\Big] \\
&= \mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\Big[\sum_{k=0}^{\infty}\gamma^t \nabla_\theta r(s_t,a_t;\theta)\Big] - \mathbb{E}_{s_0\sim\eta(\cdot)}\Big[\sum_a \pi_\theta(a|s_0)\nabla_\theta Q^{\text{soft}}_{r_\theta,\pi_\theta}(s_0,a)\Big] \qquad (26)
\end{aligned}$$

Then we need to calculate the gradient $\nabla_\theta Q^{\text{soft}}_{r_\theta,\pi_\theta}(s_0,a_0)$ as follows.

$$\begin{aligned}
&\nabla_\theta Q^{\text{soft}}_{r_\theta,\pi_\theta}(s_0,a_0)\\
&\overset{(i)}{=} \nabla_\theta\bigg(r(s_0,a_0;\theta)+\gamma\mathbb{E}_{s_1\sim\mathcal{P}(\cdot|s_0,a_0)}\Big[V^{\text{soft}}_{r_\theta,\pi_\theta}(s_1)\Big]\bigg)\\
&\overset{(ii)}{=} \nabla_\theta r(s_0,a_0;\theta)+\gamma\mathbb{E}_{s_1\sim\mathcal{P}(\cdot|s_0,a_0)}\bigg[\nabla_\theta\log\bigg(\sum_a \exp\big(Q^{\text{soft}}_{r_\theta,\pi_\theta}(s_0,a)\big)\bigg)\bigg]\\
&= \nabla_\theta r(s_0,a_0;\theta)+\gamma\mathbb{E}_{s_1\sim\mathcal{P}(\cdot|s_0,a_0)}\bigg[\sum_a \frac{\exp(Q^{\text{soft}}_{r_\theta,\pi_\theta}(s_1,a))}{\sum_{\tilde a}\exp(Q^{\text{soft}}_{r_\theta,\pi_\theta}(s_1,\tilde a))}\nabla_\theta Q^{\text{soft}}_{r_\theta,\pi_\theta}(s_1,a)\bigg]\\
&\overset{(iii)}{=} \nabla_\theta r(s_0,a_0;\theta)+\gamma\mathbb{E}_{s_1\sim\mathcal{P}(\cdot|s_0,a_0)}\bigg[\sum_a \pi_\theta(a|s_1)\nabla_\theta Q^{\text{soft}}_{r_\theta,\pi_\theta}(s_1,a)\bigg]\\
&\overset{(iv)}{=} \nabla_\theta r(s_0,a_0;\theta)+\gamma\mathbb{E}_{s_1\sim\mathcal{P}(\cdot|s_0,a_0),a_1\sim\pi_\theta(\cdot|s_1)}\bigg[\nabla_\theta\bigg(r(s_1,a_1;\theta)+\gamma\mathbb{E}_{s_2\sim\mathcal{P}(\cdot|s_1,a_1)}\Big[V^{\text{soft}}_{r_\theta,\pi_\theta}(s_2)\Big]\bigg)\bigg]\\
&\overset{(v)}{=} \mathbb{E}_{\tau\sim\pi_\theta}\bigg[\sum_{t\geq 0}\nabla_\theta r(s_t,a_t;\theta)\mid s_0,a_0\bigg]
\end{aligned}\tag{27}$$

where (i) and (iv) follows the definition of the soft Q-function; (ii) follows the fact that $V^{\text{soft}}_{r_\theta,\pi_\theta}(s)=\log(\sum_a\exp(Q^{\text{soft}}_{r_\theta,\pi_\theta}(s,a)))$; (iii) follows the fact that $\pi_\theta(a|s)\propto\exp(Q^{\text{soft}}_{r_\theta,\pi_\theta}(s,a))$; (v) is shown by recursively applying (i) - (iv).

Finally, plugging equation (27) into (26), the gradient of the maximum likelihood objective is:

$$\nabla L(\theta)=\mathbb{E}_{\tau\sim\pi^{\text{E}}}\bigg[\sum_{t\geq 0}\gamma^t\nabla_\theta r(s_t,a_t;\theta)\bigg]-\mathbb{E}_{\tau\sim\pi_\theta}\bigg[\sum_{t\geq 0}\gamma^t\nabla_\theta r(s_t,a_t;\theta)\bigg].\tag{28}$$

□

E Proof of Lemma 2

To proof Lemma 2, we proof the equality (10a) and the equality (10b) respectively. The constants L_q and L_c in Lemma 2 has the expression:

$$L_q:=\frac{L_r}{1-\gamma},\quad L_c:=\frac{2L_qL_rC_d\sqrt{|\mathcal{S}|\cdot|\mathcal{A}|}}{1-\gamma}+\frac{2L_g}{1-\gamma}.$$

E.1 Proof of Inequality (10a)

In this subsection, we prove the inequality (10a) in Lemma 2.

Proof. We show that $Q^{\text{soft}}_{r_\theta,\pi_\theta}$ has bounded gradient with respect to any reward parameter θ, then the inequality (10a) holds due to the mean value theorem. According to the equality (27), we have shown the explicit expression of $\nabla_\theta Q^{\text{soft}}_{r_\theta,\pi_\theta}(s,a)$ for any $s\in\mathcal{S}$ and $a\in\mathcal{A}$. Using this expression, we have the following series of relations:

$$\begin{aligned}
\|\nabla_\theta Q^{\text{soft}}_{r_\theta,\pi_\theta}(s,a)\| &\overset{(i)}{=} \bigg\|\mathbb{E}_{\tau\sim\pi_\theta}\bigg[\sum_{t\geq 0}\gamma^t\nabla_\theta r(s_t,a_t;\theta)\,\bigg|(s_0,a_0)=(s,a)\bigg]\bigg\|\\
&\overset{(ii)}{\leq} \mathbb{E}_{\tau\sim\pi_\theta}\bigg[\sum_{t\geq 0}\gamma^t\Big\|\nabla_\theta r(s_t,a_t;\theta)\Big\|\,\bigg|(s_0,a_0)=(s,a)\bigg]\\
&\overset{(iii)}{\leq} \mathbb{E}_{\tau\sim\pi_\theta}\bigg[\sum_{t\geq 0}\gamma^t L_r\,\bigg|(s_0,a_0)=(s,a)\bigg]\\
&= \frac{L_r}{1-\gamma}
\end{aligned}\tag{29}$$

where (i) is from the equality (27) in the proof of Lemma 1, (ii) follows Jensen's inequality and (iii) follows the inequality (9a) in Assumption 2. To complete this proof, we use the mean value theorem to show that

$$|Q^{\text{soft}}_{r_{\theta_1},\pi_{\theta_1}}(s,a) - Q^{\text{soft}}_{r_{\theta_2},\pi_{\theta_2}}(s,a)| \leq \| \max_{\theta} \nabla_\theta Q^{\text{soft}}_{r_\theta,\pi_\theta}(s,a)\| \cdot \|\theta_1 - \theta_2\| \leq L_q \|\theta_1 - \theta_2\|$$

where the last inequality follows (29) and we denote $L_q := \frac{L_r}{1-\gamma}$. Therefore, we have proved the Lipschitz continuous inequality in (10a).

$\square$

E.2 Proof of Inequality (10b)

In this section, we prove the inequality (10b) in Lemma 2.

Proof. According to Lemma 1, the gradient $\nabla L(\theta)$ is expressed as:

$$\nabla L(\theta) = \mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\Big[\sum_{t\geq 0}\gamma^t \nabla_\theta r(s_t,a_t;\theta)\Big] - \mathbb{E}_{\tau\sim\pi_\theta}\Big[\sum_{t\geq 0}\gamma^t \nabla_\theta r(s_t,a_t;\theta)\Big]. \tag{30}$$

Using the above relation, we have

$$\begin{aligned}
&\|\nabla L(\theta_1) - \nabla L(\theta_2)\| \\
&\overset{(i)}{=} \bigg\| \bigg(\mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\Big[\sum_{t\geq 0}\gamma^t \nabla_\theta r(s_t,a_t;\theta_1)\Big] - \mathbb{E}_{\tau\sim\pi_{\theta_1}}\Big[\sum_{t\geq 0}\gamma^t \nabla_\theta r(s_t,a_t;\theta_1)\Big]\bigg) - \\
&\qquad \bigg(\mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\Big[\sum_{t\geq 0}\gamma^t \nabla_\theta r(s_t,a_t;\theta_2)\Big] - \mathbb{E}_{\tau\sim\pi_{\theta_2}}\Big[\sum_{t\geq 0}\gamma^t \nabla_\theta r(s_t,a_t;\theta_2)\Big]\bigg) \bigg\| \\
&\leq \underbrace{\bigg\| \mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\Big[\sum_{t\geq 0}\gamma^t \nabla_\theta r(s_t,a_t;\theta_1)\Big] - \mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\Big[\sum_{t\geq 0}\gamma^t \nabla_\theta r(s_t,a_t;\theta_2)\Big] \bigg\|}_{:=\text{term A}} + \\
&\qquad \underbrace{\bigg\| \mathbb{E}_{\tau\sim\pi_{\theta_1}}\Big[\sum_{t\geq 0}\gamma^t \nabla_\theta r(s_t,a_t;\theta_1)\Big] - \mathbb{E}_{\tau\sim\pi_{\theta_2}}\Big[\sum_{t\geq 0}\gamma^t \nabla_\theta r(s_t,a_t;\theta_2)\Big] \bigg\|}_{:=\text{term B}}
\end{aligned} \tag{31}$$

where (i) follows the exact gradient expression in equation (30). Then we separately analyze term A and term B in (31).

For term A, it follows that

$$\begin{aligned}
&\bigg\| \mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\Big[\sum_{t\geq 0}\gamma^t \nabla_\theta r(s_t,a_t;\theta_1)\Big] - \mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\Big[\sum_{t\geq 0}\gamma^t \nabla_\theta r(s_t,a_t;\theta_2)\Big] \bigg\| \\
&\overset{(i)}{\leq} \mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\Big[\sum_{t\geq 0}\gamma^t \big\|\nabla_\theta r(s_t,a_t;\theta_1) - \nabla_\theta r(s_t,a_t;\theta_2)\big\|\Big] \\
&\overset{(ii)}{\leq} \mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\Big[\sum_{t\geq 0}\gamma^t L_g \|\theta_1 - \theta_2\|\Big] \\
&= \frac{L_g}{1-\gamma}\|\theta_1 - \theta_2\|
\end{aligned} \tag{32}$$

where (i) follows Jensen's inequality and (ii) is from (9b) in Assumption 2.

For the term B, it holds that

$$
\begin{aligned}
&\left\|\mathbb{E}_{\tau\sim\pi_{\theta_1}}\left[\sum_{t\geq 0}\gamma^t\nabla_\theta r(s_t,a_t;\theta_1)\right]-\mathbb{E}_{\tau\sim\pi_{\theta_2}}\left[\sum_{t\geq 0}\gamma^t\nabla_\theta r(s_t,a_t;\theta_2)\right]\right\| \\
&\overset{(i)}{\leq}\left\|\mathbb{E}_{\tau\sim\pi_{\theta_1}}\left[\sum_{t\geq 0}\gamma^t\nabla_\theta r(s_t,a_t;\theta_1)\right]-\mathbb{E}_{\tau\sim\pi_{\theta_2}}\left[\sum_{t\geq 0}\gamma^t\nabla_\theta r(s_t,a_t;\theta_1)\right]\right\| \\
&\quad+\left\|\mathbb{E}_{\tau\sim\pi_{\theta_2}}\left[\sum_{t\geq 0}\gamma^t\nabla_\theta r(s_t,a_t;\theta_1)\right]-\mathbb{E}_{\tau\sim\pi_{\theta_2}}\left[\sum_{t\geq 0}\gamma^t\nabla_\theta r(s_t,a_t;\theta_2)\right]\right\| \\
&\overset{(ii)}{\leq}\frac{1}{1-\gamma}\left\|\mathbb{E}_{(s,a)\sim d(\cdot,\cdot;\pi_{\theta_1})}\left[\nabla_\theta r(s_t,a_t;\theta_1)\right]-\mathbb{E}_{(s,a)\sim d(\cdot,\cdot;\pi_{\theta_2})}\left[\nabla_\theta r(s_t,a_t;\theta_1)\right]\right\| \\
&\quad+\mathbb{E}_{\tau\sim\pi_{\theta_2}}\left[\sum_{t\geq 0}\gamma^t\left\|\nabla_\theta r(s_t,a_t;\theta_1)-\nabla_\theta r(s_t,a_t;\theta_2)\right\|\right] \\
&\overset{(iii)}{\leq}\frac{1}{1-\gamma}\left\|\sum_{s\in\mathcal{S},a\in\mathcal{A}}\nabla_\theta r(s_t,a_t;\theta_1)\Big(d(s,a;\pi_{\theta_1})-d(s,a;\pi_{\theta_2})\Big)\right\|+\mathbb{E}_{\tau\sim\pi_{\theta_2}}\left[\sum_{k\geq 0}\gamma^k L_g\|\theta_1-\theta_2\|\right] \\
&\overset{(iv)}{\leq}\frac{2L_r}{1-\gamma}\|d(\cdot,\cdot;\pi_{\theta_1})-d(\cdot,\cdot;\pi_{\theta_1})\|_{TV}+\frac{L_g}{1-\gamma}\|\theta_1-\theta_2\| \qquad (33)
\end{aligned}
$$

where (i) follows the triangle inequality, (ii) is from Jensen's inequality and the definition of the discounted state-action visitation measure $d(s,a;\pi):=(1-\gamma)\pi(a|s)\sum_{t\geq 0}\gamma^t\mathcal{P}^\pi(s_t=s|s_0\sim\eta)$; (iii) is from (9b) in Assumption 2;(iv) is from (9a) and the definition of the total variation norm.

Plugging the inequalities (32), (33) to (31), it holds that

$$
\begin{aligned}
&\|\nabla L(\theta_1)-\nabla L(\theta_2)\| \\
&\leq\frac{2L_r}{1-\gamma}\|d(\cdot,\cdot;\pi_{\theta_1})-d(\cdot,\cdot;\pi_{\theta_2})\|_{TV}+\frac{2L_g}{1-\gamma}\|\theta_1-\theta_2\| \\
&\overset{(i)}{\leq}\frac{2L_rC_d}{1-\gamma}\|Q^{\text{soft}}_{r_{\theta_1},\pi_{\theta_1}}-Q^{\text{soft}}_{r_{\theta_2},\pi_{\theta_2}}\|+\frac{2L_g}{1-\gamma}\|\theta_1-\theta_2\| \\
&\overset{(ii)}{\leq}\frac{2L_rC_d\sqrt{|\mathcal{S}|\cdot|\mathcal{A}|}}{1-\gamma}\|Q^{\text{soft}}_{r_{\theta_1},\pi_{\theta_1}}-Q^{\text{soft}}_{r_{\theta_2},\pi_{\theta_2}}\|_\infty+\frac{2L_g}{1-\gamma}\|\theta_1-\theta_2\| \\
&\overset{(iii)}{\leq}\left(\frac{2L_qL_rC_d\sqrt{|\mathcal{S}|\cdot|\mathcal{A}|}}{1-\gamma}+\frac{2L_g}{1-\gamma}\right)\|\theta_1-\theta_2\|. \qquad (34)
\end{aligned}
$$

Given the fact that π_θ is a Boltzmann policy parameterized by $Q^{\text{soft}}_{r_\theta,\pi_\theta}$ where $\pi_\theta(a|s)\propto\exp(Q^{\text{soft}}_{r_\theta,\pi_\theta}(s,a))$, we show the inequality (i) from the inequality (13) in Lemma 4. Moreover, the inequality (ii) follows the equivalence relation between Frobenius norm and infinity norm and (iii) is from the inequality (10a) in Lemma 2.

Define the constant $L_c:=\frac{2L_qL_rC_d\sqrt{|\mathcal{S}|\cdot|\mathcal{A}|}}{1-\gamma}+\frac{2L_g}{1-\gamma}$, we have the following inequality:

$$\|\nabla L(\theta_1)-\nabla L(\theta_2)\|\leq L_c\|\theta_1-\theta_2\|.$$

Therefore, we complete the proof of the inequality (10b) in Lemma 2. □

F Proof of Lemma 3

Proof. Suppose the expert trajectories τ in ML-IRL is sampled from an expert policy π^{E}. Moreover, we parameterize the state-only reward as $r(s;\theta)$. Then the objective function $L(\theta)$ in ML-IRL could be rewritten as follows.

$$
\begin{aligned}
L(\theta) &:= \mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\big[\sum_{t\geq 0}\gamma^t \log \pi_\theta(a_t|s_t)\big] \\
&\overset{(i)}{=} \mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\bigg[\sum_{t=0}^{\infty}\gamma^t r(s_t;\theta)\bigg] - \mathbb{E}_{s_0\sim\eta(\cdot)}\bigg[V^{\text{soft}}_{r_\theta,\pi_\theta}(s_0)\bigg] \\
&\overset{(ii)}{=} \mathbb{E}_{s_0\sim\eta(\cdot)}\bigg[V^{\text{soft}}_{r_\theta,\pi^{\mathrm{E}}}(s_0)\bigg] - \mathbb{E}_{s_0\sim\eta(\cdot)}\bigg[V^{\text{soft}}_{r_\theta,\pi_\theta}(s_0)\bigg] - H(\pi^{\mathrm{E}})
\end{aligned} \tag{35}
$$

where (i) follows (24) and the fact that the reward is a state-only function $r(s;\theta)$; (ii) follows the definitions of the soft value function.

Ignoring the constant term $H(\pi^{\mathrm{E}})$ in (35), the maximum likelihood formulation (ML-IRL) is equivalent to the following bi-level problem:

$$
\begin{aligned}
&\min_\theta \;\; \mathbb{E}_{s_0\sim\eta(\cdot)}\big[V^{\text{soft}}_{r_\theta,\pi_\theta}(s_0)\big] - \mathbb{E}_{s_0\sim\eta(\cdot)}\big[V^{\text{soft}}_{r_\theta,\pi^E}(s_0)\big] \\
&s.t. \;\; \pi_\theta := \arg\max_\pi \; \mathbb{E}_\pi\bigg[\sum_{t=0}^{\infty}\gamma^t\bigg(r(s_t;\theta) + \mathcal{H}(\pi(\cdot|s_t))\bigg)\bigg].
\end{aligned}
$$

Therefore, we complete the proof of Lemma 3. As an alternative interpretation to (ML-IRL), the formulation above aims to minimize the gap between the soft value function of π_θ and π^{E} under the state-only IRL setting. □

G Proof of Theorem 1

Proof. Calculate the Lagrangian of MaxEnt-IRL, we obtain that

$$H(\pi) + \left\langle \theta, \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t\phi(s_t,a_t)\right] - \mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\left[\sum_{t=0}^{\infty}\gamma^t\phi(s_t,a_t)\right]\right\rangle + \sum_{s\in\mathcal{S},t\geq 0} C_{s_t=s}\Big(1-\sum_{a\in\mathcal{A}}\pi(a|s_t)\Big)$$

$$= \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}-\gamma^t\log\pi(a_t|s_t=s)\right] + \left\langle \theta, \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t\phi(s_t,a_t)\right] - \mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\left[\sum_{t=0}^{\infty}\gamma^t\phi(s_t,a_t)\right]\right\rangle$$
$$+ \sum_{s\in\mathcal{S},t\geq 0} C_{s_t=s}\Big(1-\sum_{a\in\mathcal{A}}\pi(a|s_t=s)\Big) \tag{36}$$

where θ is the dual variable to ensure the feature matching equality, and $C_{s_t=s}$ is the dual variable to ensure that π is a well-defined policy satisfying $\sum_{a\in\mathcal{A}}\pi(a|s_t=s)=1$.

Then we could calculate the gradient of (36) w.r.t. $\pi(a|s_t=s)$, and set it to 0. Then it holds that

$$0 = \mathcal{P}^{\pi}(s_t=s)\bigg(-\gamma^t\big(\log\pi(a|s_t=s)+1\big) + \mathbb{E}_{\pi}\left[\sum_{\kappa=t}^{\infty}-\gamma^{\kappa+1}\log\pi(a_{\kappa+1}|s_{\kappa+1}) \mid s_t=s,a_t=a\right]$$
$$+ \theta^T\mathbb{E}_{\pi}\left[\sum_{\kappa=t}^{\infty}\gamma^{\kappa}\phi(s_\kappa,a_\kappa)|s_t=s,a_t=a\right]\bigg) - C_{s_t=s}. \tag{37}$$

Dividing $\gamma^t\mathcal{P}^{\pi}(s_t=s)$ on both sides of (37) and further moving $\log\pi(a|s_t=s)$ to the left side, then we have the equality as below:

$$\log\pi(a|s_t=s) = \Big(-\frac{C_{s_t=s}}{\gamma^t\mathcal{P}^{\pi}(s_t=s)}-1\Big) + \mathbb{E}_{\pi}\left[\sum_{\kappa=t+1}^{\infty}-\gamma^{\kappa-t}\log\pi(a_\kappa|s_\kappa) \mid s_t=s,a_t=a\right]$$
$$+ \theta^T\mathbb{E}_{\pi}\left[\sum_{\kappa=t}^{\infty}\gamma^{\kappa-t}\phi(s_\kappa,a_\kappa)|s_t=s,a_t=a\right] \tag{38}$$

Given that $-\frac{C_{s_t=s}}{\gamma^t\mathcal{P}^{\pi}(s_t=s)}-1$ is independent of action a, we could express the closed form of $\pi(a|s_t=s)$ as below:

$$\pi(a|s_t=s) \propto \exp\left(\mathbb{E}_{\pi}\left[\sum_{\kappa=t+1}^{\infty}-\gamma^{\kappa-t}\log\pi(a_\kappa|s_\kappa) \mid s_t=s,a_t=a\right] + \theta^T\mathbb{E}_{\pi}\left[\sum_{\kappa=t}^{\infty}\gamma^{\kappa-t}\phi(s_\kappa,a_\kappa)|s_t=s,a_t=a\right]\right).$$

According to the closed form of the policy above, it shows that $\pi(a|s_t=s)$ is a stationary policy being independent of the time index t. Therefore, it holds that $\pi(a|s_t=s)=\pi(a|s)$ for any $t\geq 0$.

Denoting a linearly parameterized reward as $r(s,a;\theta) := \theta^T\phi(s,a)$, it holds that

$$\pi(a|s) \propto \exp\left(\theta^T\mathbb{E}_{\pi}\left[\sum_{\kappa=0}^{\infty}\gamma^{\kappa}\phi(s_\kappa,a_\kappa) \mid s_0=s,a_0=a\right] + \mathbb{E}_{\pi}\left[\sum_{\kappa=0}^{\infty}-\gamma^{\kappa+1}\log\pi(a_{\kappa+1}|s_{\kappa+1}) \mid s_0=s,a_0=a\right]\right)$$
$$= \exp\left(\mathbb{E}_{\pi}\left[\sum_{\kappa=0}^{\infty}\gamma^{\kappa}r(s_\kappa,a_\kappa;\theta) \mid s_0=s,a_0=a\right] + \mathbb{E}_{\pi}\left[\sum_{\kappa=0}^{\infty}-\gamma^{\kappa+1}\log\pi(a_{\kappa+1}|s_{\kappa+1}) \mid s_0=s,a_0=a\right]\right) \tag{39}$$

Here, the optimal $\pi(a|s)$ is a function of the dual variables (reward parameters) θ. In the maximum entropy reinforcement learning [21], under a reward function $r(\cdot,\cdot)$ and policy π, the soft value function and soft Q-function are defined as below:

$$V^{\text{soft}}_{r,\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty}\gamma^t\Big(r(s_t,a_t)+\mathcal{H}(\pi(\cdot|s_t))\Big)\Big|s_0=s\right] \tag{40a}$$

$$Q^{\text{soft}}_{r,\pi}(s,a) = r(s,a) + \gamma\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)}\left[V^{\text{soft}}_{r,\pi}(s)\right] \tag{40b}$$

Based on the definitions in (40a) - (40b), we could further express the closed form of the policy in (39) as below:

$$\pi(a|s) = \frac{\exp\big(Q^{\text{soft}}_{r_\theta,\pi}(s,a)\big)}{\sum_{a\in\mathcal{A}}\exp\big(Q^{\text{soft}}_{r_\theta,\pi}(s,a)\big)}. \tag{41}$$

According to [21], under a reward function $r(\cdot,\cdot)$, the optimal soft policy π satisfies $\pi(\cdot|s) \propto \exp(Q^{soft}_{r,\pi}(s,\cdot))$. Hence, we have shown that the policy in (41) is the optimal policy under the reward function $r(\cdot,\cdot;\theta)$. After denoting the optimal policy under $r(\cdot,\cdot;\theta)$ as π_θ, we have the following relation:

$$\pi_\theta(a|s) = \frac{\exp\big(Q^{\text{soft}}_{r_\theta,\pi_\theta}(s,a)\big)}{\sum_{a\in\mathcal{A}}\exp\big(Q^{\text{soft}}_{r_\theta,\pi_\theta}(s,a)\big)} \overset{(a)}{=} \frac{\exp\big(Q^{\text{soft}}_{r_\theta,\pi_\theta}(s,a)\big)}{\exp\big(V^{\text{soft}}_{r_\theta,\pi_\theta}(s)\big)} = \exp\Bigg(Q^{\text{soft}}_{r_\theta,\pi_\theta}(s,a) - V^{\text{soft}}_{r_\theta,\pi_\theta}(s)\Bigg) \tag{42}$$

where (a) is due to the equality shown as below:

$$\begin{aligned}
V^{\text{soft}}_{r_\theta,\pi_\theta}(s) &= \mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\Big[-\log\big(\pi_\theta(a|s)\big) + Q^{\text{soft}}_{r_\theta,\pi_\theta}(s,a)\Big] \\
&= \mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\Bigg[-\log\Bigg(\frac{\exp\big(Q^{\text{soft}}_{r_\theta,\pi_\theta}(s,a)\big)}{\sum_{a\in\mathcal{A}}\exp\big(Q^{\text{soft}}_{r_\theta,\pi_\theta}(s,a)\big)}\Bigg) + Q^{\text{soft}}_{r_\theta,\pi_\theta}(s,a)\Bigg] \\
&= \log\big(\sum_{a\in\mathcal{A}}\exp\big(Q^{\text{soft}}_{r_\theta,\pi_\theta}(s,a)\big)\big).
\end{aligned}$$

Rewriting the equality (37), we are able to show the expression of $C_{s_t=s}$ as below:

$$\begin{aligned}
C_{s_t=s} = \mathcal{P}^\pi(s_t=s)\Bigg(&-\gamma^t\big(\log\pi_\theta(a|s_t=s)+1\big) + \mathbb{E}_{\pi_\theta}\Bigg[\sum_{\kappa=t}^{\infty}-\gamma^{\kappa+1}\log\pi_\theta(a_{\kappa+1}|s_{\kappa+1}) \mid s_t=s, a_t=a\Bigg] \\
&+ \theta^T\mathbb{E}_{\pi_\theta}\Bigg[\sum_{\kappa=t}^{\infty}\gamma^\kappa\phi(s_\kappa,a_\kappa)|s_t=s,a_t=a\Bigg]\Bigg) \\
= \gamma^t\mathcal{P}^{\pi_\theta}(s_t=s)\Bigg(&-1-\log\pi_\theta(a|s) + \mathbb{E}_{\pi_\theta}\Bigg[\sum_{\kappa=0}^{\infty}-\gamma^{\kappa+1}\log\pi_\theta(a_{\kappa+1}|s_{\kappa+1}) \mid s_0=s, a_0=a\Bigg] \\
&+ \mathbb{E}_{\pi_\theta}\Bigg[\sum_{\kappa=0}^{\infty}\gamma^\kappa r(s_\kappa,a_\kappa;\theta)|s_0=s,a_0=a\Bigg]\Bigg) \\
\overset{(a)}{=} \gamma^t\mathcal{P}^\pi(s_t=s)\Bigg(&-1-\log\pi_\theta(a|s) + Q^{\text{soft}}_{r_\theta,\pi_\theta}(s,a)\Bigg) \\
\overset{(b)}{=} \gamma^t\mathcal{P}^\pi(s_t=s)\Bigg(&V^{\text{soft}}_{r_\theta,\pi_\theta}(s) - 1\Bigg)
\end{aligned} \tag{43}$$

where (a) follows the definition of the soft Q-function in (40b), and (b) follows (42). According to (43), we are able to show the exact expression of $C_{s_t=s}$.

Plugging π_θ and $C_{s_t=s}$ into (36), we have

$$
\begin{aligned}
&\mathbb{E}_{\tau\sim\pi_\theta}\Big[\sum_{t=0}^{\infty} -\gamma^t \log \pi_\theta(a_t|s_t)\Big] + \Big\langle \theta, \mathbb{E}_{\tau\sim\pi_\theta}\Big[\sum_{t=0}^{\infty}\gamma^t\phi(s_t,a_t)\Big] - \mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\Big[\sum_{t=0}^{\infty}\gamma^t\phi(s_t,a_t)\Big]\Big\rangle \\
&\quad + \sum_{s\in\mathcal{S},t\geq 0} C_{s_t=s}\Big(1-\sum_{a\in\mathcal{A}}\pi_\theta(a|s_t=s)\Big) \\
&\overset{(a)}{=} \mathbb{E}_{\tau\sim\pi_\theta}\Big[\sum_{t=0}^{\infty} -\gamma^t \log \pi_\theta(a_t|s_t)\Big] + \Big\langle \theta, \mathbb{E}_{\tau\sim\pi_\theta}\Big[\sum_{t=0}^{\infty}\gamma^t\phi(s_t,a_t)\Big] - \mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\Big[\sum_{t=0}^{\infty}\gamma^t\phi(s_t,a_t)\Big]\Big\rangle \\
&\overset{(b)}{=} \mathbb{E}_{\tau\sim\pi_\theta}\Big[\sum_{t=0}^{\infty} -\gamma^t \Big(Q^{\text{soft}}_{r_\theta,\pi_\theta}(s,a) - V^{\text{soft}}_{r_\theta,\pi_\theta}(s)\Big)\Big] + \Big\langle \theta, \mathbb{E}_{\tau\sim\pi_\theta}\Big[\sum_{t=0}^{\infty}\gamma^t\phi(s_t,a_t)\Big] - \mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\Big[\sum_{t=0}^{\infty}\gamma^t\phi(s_t,a_t)\Big]\Big\rangle \\
&\overset{(c)}{=} \mathbb{E}_{\tau\sim\pi_\theta}\Big[\sum_{t=0}^{\infty} -\gamma^t \Big(\theta^T\phi(s_t,a_t) + \gamma V^{soft}_{r_\theta,\pi_\theta}(s_{t+1}) - V^{soft}_{r_\theta,\pi_\theta}(s_t)\Big)\Big] \\
&\quad + \Big\langle \theta, \mathbb{E}_{\tau\sim\pi_\theta}\Big[\sum_{t=0}^{\infty}\gamma^t\phi(s_t,a_t)\Big] - \mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\Big[\sum_{t=0}^{\infty}\gamma^t\phi(s_t,a_t)\Big]\Big\rangle \\
&= \mathbb{E}_{s_0\sim\eta(\cdot)}\Big[V^{\text{soft}}_{r_\theta,\pi_\theta}(s_0)\Big] - \theta^T\mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\Big[\sum_{t=0}^{\infty}\gamma^t\phi(s_t,a_t)\Big] \qquad (44)
\end{aligned}
$$

where (a) is due to the fact that $\sum_{a\in\mathcal{A}}\pi_\theta(a|s_t=s)=1$ for all $a\in\mathcal{A}$ and $s\in\mathcal{S}$; (b) follows (42); (c) is due to the definition of the soft Q-function in (40b). Here, we could further show the problem in (44) is equivalent to ML-IRL as below:

$$
\begin{aligned}
\mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\Big[\sum_{t=0}^{\infty}\gamma^t \ln \pi_\theta(a_t|s_t)\Big] &= \sum_{t=0}^{\infty}\gamma^t\cdot\mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\Big[r(s_t,a_t;\theta) + \gamma V^{\text{soft}}_{r_\theta,\pi_\theta}(s_{t+1}) - V^{\text{soft}}_{r_\theta,\pi_\theta}(s_t)\Big] \\
&= \mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\Big[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t;\theta)\Big] + \sum_{t=0}^{\infty}\gamma^t\cdot\mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\Big[\gamma V^{\text{soft}}_{r_\theta,\pi_\theta}(s_{t+1}) - V^{\text{soft}}_{r_\theta,\pi_\theta}(s_t)\Big] \\
&= \mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\Big[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t;\theta)\Big] - \mathbb{E}_{s_0\sim\eta(\cdot)}\Big[V^{\text{soft}}_{r_\theta,\pi_\theta}(s_0)\Big] \\
&= \theta^T\mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\Big[\sum_{t=0}^{\infty}\gamma^t\phi(s_t,a_t)\Big] - \mathbb{E}_{s_0\sim\eta(\cdot)}\Big[V^{\text{soft}}_{r_\theta,\pi_\theta}(s_0)\Big] \qquad (45)
\end{aligned}
$$

Finally, through combining (44) and (45), we are able to know that the maximum likelihood formulation ML-IRL is the dual form of MaxEnt-IRL.

□

H Proof of Theorem 2

In this section, we prove (11a) and (11b) respectively, to show the convergence of the lower-level problem and the upper-level problem.

H.1 Proof of (11a)

Proof. In this proof, we first show the convergence of the lower-level variable $\{\pi_k\}_{k\geq 0}$. Recall that we approximate the optimal policy π_{θ_k} by π_{k+1} at each iteration k. We first analyze the approximation error between π_{θ_k} and π_{k+1} as follows. For any $s \in \mathcal{S}$ and $a \in \mathcal{A}$, we have the following relation:

$$
\begin{aligned}
&\big|\log\big(\pi_{k+1}(a|s)\big) - \log\big(\pi_{\theta_k}(a|s)\big)\big| \\
&\overset{(i)}{=} \left|\log\left(\frac{\exp\big(Q^{\text{soft}}_{r_{\theta_k},\pi_k}(s,a)\big)}{\sum_{\tilde a}\exp\big(Q^{\text{soft}}_{r_{\theta_k},\pi_k}(s,\tilde a)\big)}\right) - \log\left(\frac{\exp\big(Q^{\text{soft}}_{r_{\theta_k},\pi_{\theta_k}}(s,a)\big)}{\sum_{\tilde a}\exp\big(Q^{\text{soft}}_{r_{\theta_k},\pi_{\theta_k}}(s,\tilde a)\big)}\right)\right| \\
&\overset{(ii)}{\leq} \big|Q^{\text{soft}}_{r_{\theta_k},\pi_k}(s,a) - Q^{\text{soft}}_{r_{\theta_k},\pi_{\theta_k}}(s,a)\big| + \left|\log\Big(\sum_{\tilde a}\exp\big(Q^{\text{soft}}_{r_{\theta_k},\pi_k}(s,\tilde a)\big)\Big) - \log\Big(\sum_{\tilde a}\exp\big(Q^{\text{soft}}_{r_{\theta_k},\pi_{\theta_k}}(s,\tilde a)\big)\Big)\right|
\end{aligned} \tag{46}
$$

where (i) follows (5) and the fact that $\pi_\theta(a|s) \propto \exp(Q^{\text{soft}}_{r_\theta,\pi_\theta}(s,a))$; (ii) follows the triangle inequality. We further analyze the second term in (46).

We first denote the operator $\log(\|\exp(v)\|_1) := \log(\|\sum_{\tilde a\in\mathcal{A}}\exp(v_{\tilde a})\|_1)$, where the vector $v \in \mathbb{R}^{|\mathcal{A}|}$ and $v = [v_1, v_2, \cdots, v_{|\mathcal{A}|}]$. Then for any $v', v'' \in \mathbb{R}^{|\mathcal{A}|}$, we have the following relation:

$$
\begin{aligned}
\big|\log\big(\|\exp(v')\|_1\big) - \log\big(\|\exp(v'')\|_1\big) &\overset{(i)}{=} \big\langle v'-v'', \nabla_v \log\big(\|\exp(v)\|_1\big)|_{v=v^c}\big\rangle \\
&\leq \|v'-v''\|_\infty \cdot \|\nabla_v \log\big(\|\exp(v)\|_1\big)|_{v=v^c}\|_1 \\
&\overset{(ii)}{=} \|v'-v''\|_\infty
\end{aligned} \tag{47}
$$

where (i) follows the mean value theorem and v_c is a convex combination of v' and v''; (ii) follows the following equalities:

$$
[\nabla_v \log\big(\|\exp(v)\|_1\big)]_i = \frac{\exp(v_i)}{\sum_{1\leq a\leq|\mathcal{A}|}\exp(v_a)}, \quad \|\nabla_v \log\big(\|\exp(v)\|_1\big)\|_1 = 1, \quad \forall v \in \mathbb{R}^{|\mathcal{A}|}.
$$

Through plugging (47) into (46), it holds that

$$
\begin{aligned}
&\big|\log\big(\pi_{k+1}(a|s)\big) - \log\big(\pi_{\theta_k}(a|s)\big)\big| \\
&\leq \big|Q^{\text{soft}}_{r_{\theta_k},\pi_k}(s,a) - Q^{\text{soft}}_{r_{\theta_k},\pi_{\theta_k}}(s,a)\big| + \max_{\tilde a\in\mathcal{A}}\big|Q^{\text{soft}}_{r_{\theta_k},\pi_k}(s,\tilde a) - Q^{\text{soft}}_{r_{\theta_k},\pi_{\theta_k}}(s,\tilde a)\big|
\end{aligned} \tag{48}
$$

Taking the infinity norm over $\mathbb{R}^{|\mathcal{S}|\cdot|\mathcal{A}|}$, the following result holds:

$$
\|\log\pi_{k+1} - \log\pi_{\theta_k}\|_\infty \leq 2\|Q^{\text{soft}}_{r_{\theta_k},\pi_k} - Q^{\text{soft}}_{r_{\theta_k},\pi_{\theta_k}}\|_\infty \tag{49}
$$

where $\|\log\pi_{k+1} - \log\pi_{\theta_k}\|_\infty = \max_{s\in\mathcal{S},a\in\mathcal{A}}|\log\pi_{k+1}(a|s) - \log\pi_{\theta_k}(a|s)|$ and $\|Q^{\text{soft}}_{r_{\theta_k},\pi_k} - Q^{\text{soft}}_{r_{\theta_k},\pi_{\theta_k}}\|_\infty = \max_{s\in\mathcal{S},a\in\mathcal{A}}|Q^{\text{soft}}_{r_{\theta_k},\pi_k}(s,a) - Q^{\text{soft}}_{r_{\theta_k},\pi_{\theta_k}}(s,a)|$.

Based on the inequality (49), we analyze $\|Q^{\text{soft}}_{r_{\theta_k},\pi_k} - Q^{\text{soft}}_{r_{\theta_k},\pi_{\theta_k}}\|_\infty$ to show the convergence of the policy estimates. It leads to the following analysis:

$$
\begin{aligned}
&\|Q^{\text{soft}}_{r_{\theta_k},\pi_k} - Q^{\text{soft}}_{r_{\theta_k},\pi_{\theta_k}}\|_\infty \\
&= \|Q^{\text{soft}}_{r_{\theta_k},\pi_k} - Q^{\text{soft}}_{r_{\theta_k},\pi_{\theta_k}} + Q^{\text{soft}}_{r_{\theta_{k-1}},\pi_{\theta_{k-1}}} - Q^{\text{soft}}_{r_{\theta_{k-1}},\pi_{\theta_{k-1}}} + Q^{\text{soft}}_{r_{\theta_{k-1}},\pi_k} - Q^{\text{soft}}_{r_{\theta_{k-1}},\pi_k}\|_\infty \\
&\leq \|Q^{\text{soft}}_{r_{\theta_k},\pi_{\theta_k}} - Q^{\text{soft}}_{r_{\theta_{k-1}},\pi_{\theta_{k-1}}}\|_\infty + \|Q^{\text{soft}}_{r_{\theta_{k-1}},\pi_k} - Q^{\text{soft}}_{r_{\theta_{k-1}},\pi_{\theta_{k-1}}}\|_\infty + \|Q^{\text{soft}}_{r_{\theta_k},\pi_k} - Q^{\text{soft}}_{r_{\theta_{k-1}},\pi_k}\|_\infty \\
&\overset{(i)}{\leq} L_q\|\theta_k - \theta_{k-1}\| + \|Q^{\text{soft}}_{r_{\theta_{k-1}},\pi_k} - Q^{\text{soft}}_{r_{\theta_{k-1}},\pi_{\theta_{k-1}}}\|_\infty + \|Q^{\text{soft}}_{r_{\theta_k},\pi_k} - Q^{\text{soft}}_{r_{\theta_{k-1}},\pi_k}\|_\infty \\
&\overset{(ii)}{\leq} \|Q^{\text{soft}}_{r_{\theta_{k-1}},\pi_k} - Q^{\text{soft}}_{r_{\theta_{k-1}},\pi_{\theta_{k-1}}}\|_\infty + 2L_q\|\theta_k - \theta_{k-1}\|
\end{aligned} \tag{50}
$$

where (i) is from (10a) in Lemma 2; (ii) follows Lemma 7. Based on (50), we further analyze the two terms in (50) as below.

Recall Lemma 6, we have the "soft" Bellman operator expressed as below:

$$\mathcal{T}_\theta(Q)(s,a) = r(s,a;\theta) + \gamma \mathbb{E}_{s' \sim P(\cdot|s',a')}\left[\log\left(\sum_{a'} \exp\left(Q(s',a') \right) \right) \right] \tag{51}$$

According to the soft Bellman operator, it holds that

$$\begin{aligned} Q^{\text{soft}}_{r_{\theta_k},\pi_{k+1}}(s,a) &= r(s,a;\theta_k) + \gamma \mathbb{E}_{s' \sim \mathcal{P}(\cdot|s,a)}[V^{\text{soft}}_{r_{\theta_k},\pi_{k+1}}(s')] \\ &= r(s,a;\theta_k) + \gamma \mathbb{E}_{s' \sim \mathcal{P}(\cdot|s,a), a' \sim \pi_{k+1}(\cdot|s')}[-\log \pi_{k+1}(a'|s') + Q^{\text{soft}}_{r_{\theta_k},\pi_{k+1}}(s',a')] \\ &\overset{(i)}{\geq} r(s,a;\theta_k) + \gamma \mathbb{E}_{s' \sim \mathcal{P}(\cdot|s,a), a' \sim \pi_{k+1}(\cdot|s')}[-\log \pi_{k+1}(a'|s') + Q^{\text{soft}}_{r_{\theta_k},\pi_k}(s',a')] \\ &\overset{(ii)}{=} r(s,a;\theta_k) + \gamma \mathbb{E}_{s' \sim \mathcal{P}(\cdot|s,a)}\left[\log\left(\sum_{a'} \exp\left(Q^{\text{soft}}_{r_{\theta_k},\pi_k}(s',a') \right) \right) \right] \\ &\overset{(iii)}{=} \mathcal{T}_{\theta_k}(Q^{\text{soft}}_{r_{\theta_k},\pi_k})(s,a) \end{aligned} \tag{52}$$

where (i) follows the policy improvement result in Lemma 5, (ii) follows the definition $\pi_{k+1}(a|s) := \frac{\exp\left(Q^{\text{soft}}_{r_{\theta_k},\pi_k}(s,a)\right)}{\sum_{\tilde{a}} \exp\left(Q^{\text{soft}}_{r_{\theta_k},\pi_k}(s,\tilde{a})\right)}$ in (5); (iii) follows the definition of the soft Bellman operator in (51).

For any $s \in \mathcal{S}$ and $a \in \mathcal{A}$, it holds that

$$0 \overset{(i)}{\leq} Q^{\text{soft}}_{r_{\theta_k},\pi_{\theta_k}}(s,a) - Q^{\text{soft}}_{r_{\theta_k},\pi_{k+1}}(s,a) \overset{(ii)}{\leq} Q^{\text{soft}}_{r_{\theta_k},\pi_{\theta_k}}(s,a) - \mathcal{T}_{\theta_k}(Q^{\text{soft}}_{r_{\theta_k},\pi_k})(s,a) \tag{53}$$

where (i) is due to the fact that π_{θ_k} is the optimal policy under reward parameter θ_k; (ii) is from (52).

Hence, it further leads to

$$\begin{aligned} \|Q^{\text{soft}}_{r_{\theta_k},\pi_{\theta_k}} - Q^{\text{soft}}_{r_{\theta_k},\pi_{k+1}}\|_\infty &\overset{(i)}{\leq} \|Q^{\text{soft}}_{r_{\theta_k},\pi_{\theta_k}} - \mathcal{T}_{\theta_k}(Q^{\text{soft}}_{r_{\theta_k},\pi_k})\|_\infty \\ &\overset{(ii)}{=} \|\mathcal{T}_{\theta_k}(Q^{\text{soft}}_{r_{\theta_k},\pi_{\theta_k}}) - \mathcal{T}_{\theta_k}(Q^{\text{soft}}_{r_{\theta_k},\pi_k})\|_\infty \\ &\overset{(iii)}{\leq} \gamma \|Q^{\text{soft}}_{r_{\theta_k},\pi_{\theta_k}} - Q^{\text{soft}}_{r_{\theta_k},\pi_k}\|_\infty \end{aligned} \tag{54}$$

where (i) is from (53); (ii) is from the fixed-point property in (17); (iii) is from the contraction property in (16). Therefore, we have the following result:

$$\begin{aligned} &\|Q^{\text{soft}}_{r_{\theta_k},\pi_k} - Q^{\text{soft}}_{r_{\theta_k},\pi_{\theta_k}}\|_\infty \\ &\overset{(i)}{\leq} \|Q^{\text{soft}}_{r_{\theta_{k-1}},\pi_k} - Q^{\text{soft}}_{r_{\theta_{k-1}},\pi_{\theta_{k-1}}}\|_\infty + 2L_q\|\theta_k - \theta_{k-1}\| \\ &\overset{(ii)}{\leq} \gamma \|Q^{\text{soft}}_{r_{\theta_{k-1}},\pi_{k-1}} - Q^{\text{soft}}_{r_{\theta_{k-1}},\pi_{\theta_{k-1}}}\|_\infty + 2L_q\|\theta_k - \theta_{k-1}\| \end{aligned} \tag{55}$$

where (i) is from (50); (ii) is from (54).

To show the convergence of the soft Q-function based on (55), we further analyze the error between the reward parameters θ_k and θ_{k-1}. Recall in Alg.1, the updates in reward parameters follows (8):

$$\begin{aligned} \theta_k &= \theta_{k-1} + \alpha g_{k-1} \\ &= \theta_{k-1} + \alpha\big(h(\theta_{k-1}, \tau^E_{k-1}) - h(\theta_{k-1}, \tau^A_{k-1})\big) \end{aligned}$$

where we denote $\tau = \{(s_t, a_t)\}_{t=0}^\infty$, $h(\theta, \tau) := \sum_{t \geq 0} \gamma^t \nabla_\theta r(s_t, a_t; \theta)$ and g_{k-1} is the stochastic gradient estimator at iteration $k-1$. Here, τ^E_{k-1} denotes the trajectory sampled from the expert's dataset D at iteration $k-1$ and τ^A_{k-1} denotes the trajectory sampled from the agent's policy π_k at time $k-1$. Then according to the inequality (9a) in Assumption 2, we could show that

$$\|g_{k-1}\| \leq \|h(\theta_{k-1}, \tau^E_{k-1})\| + \|h(\theta_{k-1}, \tau^A_{k-1})\| \leq 2L_r \sum_{t \geq 0} \gamma^t = \frac{2L_r}{1-\gamma} = 2L_q \tag{56}$$

where the last equality follows the fact that we have defined the constant $L_q := \frac{L_r}{1-\gamma}$. Then we could further show that

$$
\begin{aligned}
&\|Q^{\text{soft}}_{r_{\theta_k},\pi_k} - Q^{\text{soft}}_{r_{\theta_k},\pi_{\theta_k}}\|_\infty \\
&\overset{(i)}{\leq} \gamma\|Q^{\text{soft}}_{r_{\theta_{k-1}},\pi_{k-1}} - Q^{\text{soft}}_{r_{\theta_{k-1}},\pi_{\theta_{k-1}}}\|_\infty + 2L_q\|\theta_k - \theta_{k-1}\| \\
&\overset{(ii)}{=} \gamma\|Q^{\text{soft}}_{r_{\theta_{k-1}},\pi_{k-1}} - Q^{\text{soft}}_{r_{\theta_{k-1}},\pi_{\theta_{k-1}}}\|_\infty + 2\alpha L_q\|g_{k-1}\| \\
&\overset{(iii)}{\leq} \gamma\|Q^{\text{soft}}_{r_{\theta_{k-1}},\pi_{k-1}} - Q^{\text{soft}}_{r_{\theta_{k-1}},\pi_{\theta_{k-1}}}\|_\infty + 4\alpha L_q^2
\end{aligned} \tag{57}
$$

where (i) is from (55); (ii) follows the reward update scheme in (8); (iii) is from (56).

Summing the inequality (57) from $k = 1$ to $k = K$, it holds that

$$
\sum_{k=1}^{K} \|Q^{\text{soft}}_{r_{\theta_k},\pi_k} - Q^{\text{soft}}_{r_{\theta_k},\pi_{\theta_k}}\|_\infty \leq \gamma \sum_{k=0}^{K-1} \|Q^{\text{soft}}_{r_{\theta_k},\pi_k} - Q^{\text{soft}}_{r_{\theta_k},\pi_{\theta_k}}\|_\infty + 4\alpha K L_q^2 \tag{58}
$$

Rearranging the inequality (58) and divided (58) by K on both sides, it holds that

$$
\frac{1-\gamma}{K}\sum_{k=1}^{K} \|Q^{\text{soft}}_{r_{\theta_k},\pi_k} - Q^{\text{soft}}_{r_{\theta_k},\pi_{\theta_k}}\|_\infty \leq \frac{\gamma}{K}\Big(\|Q^{\text{soft}}_{r_{\theta_0},\pi_0} - Q^{\text{soft}}_{r_{\theta_0},\pi_{\theta_0}}\|_\infty - \|Q^{\text{soft}}_{r_{\theta_K},\pi_K} - Q^{\text{soft}}_{r_{\theta_K},\pi_{\theta_K}}\|_\infty\Big) + 4\alpha L_q^2 \tag{59}
$$

Dividing the constant $1-\gamma$ on both sides of (59), it holds that

$$
\frac{1}{K}\sum_{k=1}^{K} \|Q^{\text{soft}}_{r_{\theta_k},\pi_k} - Q^{\text{soft}}_{r_{\theta_k},\pi_{\theta_k}}\|_\infty \leq \frac{\gamma C_0}{K(1-\gamma)} + \frac{4L_q^2}{1-\gamma}\alpha
$$

where we denote $C_0 := \|Q^{\text{soft}}_{r_{\theta_0},\pi_0} - Q^{\text{soft}}_{r_{\theta_0},\pi_{\theta_0}}\|_\infty$. We could also write the inequality above as

$$
\begin{aligned}
&\frac{1}{K}\sum_{k=0}^{K-1} \|Q^{\text{soft}}_{r_{\theta_k},\pi_k} - Q^{\text{soft}}_{r_{\theta_k},\pi_{\theta_k}}\|_\infty \\
&\leq \frac{\gamma C_0}{T(1-\gamma)} + \frac{C_0}{T} - \frac{\|Q^{\text{soft}}_{r_{\theta_K},\pi_K} - Q^{\text{soft}}_{r_{\theta_K},\pi_{\theta_K}}\|_\infty}{K} + \frac{4L_q^2}{1-\gamma}\alpha \\
&\leq \frac{C_0}{T(1-\gamma)} + \frac{4L_q^2}{1-\gamma}\alpha.
\end{aligned}
$$

Recall the stepsize is defined as $\alpha = \frac{\alpha_0}{T^\sigma}$ where $\sigma > 0$. Then we have the following result:

$$
\frac{1}{K}\sum_{k=0}^{K-1} \|Q^{\text{soft}}_{r_{\theta_k},\pi_k} - Q^{\text{soft}}_{r_{\theta_k},\pi_{\theta_k}}\|_\infty = \mathcal{O}(K^{-1}) + \mathcal{O}(K^{-\sigma}). \tag{60}
$$

With the inequality (49), it follows that

$$
\frac{1}{K}\sum_{k=0}^{K-1} \|\log\pi_{k+1} - \log\pi_{\theta_k}\|_\infty \leq \frac{2}{K}\sum_{k=0}^{K-1} \|Q^{\text{soft}}_{r_{\theta_k},\pi_k} - Q^{\text{soft}}_{r_{\theta_k},\pi_{\theta_k}}\|_\infty = \mathcal{O}(K^{-1}) + \mathcal{O}(K^{-\sigma}).
$$

Therefore, we complete the proof of (11a) in Theorem 2. □

H.2 Proof of (11b)

Proof. In this part, we prove the convergence of reward parameters $\{\theta_k\}_{k\geq 0}$.

We have the following result of the objective function $L(\theta)$:

$$
\begin{aligned}
L(\theta_{k+1}) &\overset{(i)}{\geq} L(\theta_k) + \langle \nabla L(\theta_k), \theta_{k+1} - \theta_k \rangle - \frac{L_c}{2}\|\theta_{k+1} - \theta_k\|^2 \\
&\overset{(ii)}{=} L(\theta_k) + \alpha \langle \nabla L(\theta_k), g_k \rangle - \frac{L_c \alpha^2}{2}\|g_k\|^2 \\
&= L(\theta_k) + \alpha \langle \nabla L(\theta_k), g_k - \nabla L(\theta_k) \rangle + \alpha \|\nabla L(\theta_k)\|^2 - \frac{L_c \alpha^2}{2}\|g_k\|^2 \\
&\overset{(iii)}{\geq} L(\theta_k) + \alpha \langle \nabla L(\theta_k), g_k - \nabla L(\theta_k) \rangle + \alpha \|\nabla L(\theta_k)\|^2 - 2L_c L_q^2 \alpha^2
\end{aligned} \tag{61}
$$

where (i) is from the Lipschitz smooth property in (10b) of Lemma 2; (ii) follows the update scheme (8); (iii) is from constant bound in (56).

Taking an expectation over the both sides of (61), it holds that

$$
\begin{aligned}
&\mathbb{E}\left[L(\theta_{k+1})\right] \\
&\geq \mathbb{E}\left[L(\theta_k)\right] + \alpha \mathbb{E}\Big[\langle \nabla L(\theta_k), g_k - \nabla L(\theta_k) \rangle\Big] + \alpha \mathbb{E}\Big[\|\nabla L(\theta_k)\|^2\Big] - 2L_c L_q^2 \alpha^2 \\
&= \mathbb{E}\left[L(\theta_k)\right] + \alpha \mathbb{E}\Big[\langle \nabla L(\theta_k), \mathbb{E}\big[g_k - \nabla L(\theta_k) \big| \theta_k\big] \rangle\Big] + \alpha \mathbb{E}\Big[\|\nabla L(\theta_k)\|^2\Big] - 2L_c L_q^2 \alpha^2 \\
&\overset{(i)}{=} \mathbb{E}\left[L(\theta_k)\right] + \alpha \mathbb{E}\bigg[\Big\langle \nabla L(\theta_k), \mathbb{E}_{\tau \sim \pi_{\theta_k}}\Big[\sum_{t \geq 0} \gamma^t \nabla_\theta r(s_t, a_t; \theta_t)\Big] - \mathbb{E}_{\tau \sim \pi_{k+1}}\Big[\sum_{t \geq 0} \gamma^t \nabla_\theta r(s_t, a_t; \theta_t)\Big] \Big\rangle\bigg] \\
&\quad + \alpha \mathbb{E}\Big[\|\nabla L(\theta_k)\|^2\Big] - 2L_c L_q^2 \alpha^2 \\
&\overset{(ii)}{\geq} \mathbb{E}\left[L(\theta_k)\right] - 2\alpha L_q \underbrace{\mathbb{E}\bigg[\Big\|\mathbb{E}_{\tau \sim \pi_{\theta_k}}\Big[\sum_{t \geq 0} \gamma^t \nabla_\theta r(s_t, a_t; \theta_k)\Big] - \mathbb{E}_{\tau \sim \pi_{k+1}}\Big[\sum_{t \geq 0} \gamma^t \nabla_\theta r(s_t, a_t; \theta_k)\Big]\Big\|\bigg]}_{\text{term A}} \\
&\quad + \alpha \mathbb{E}\Big[\|\nabla L(\theta_k)\|^2\Big] - 2L_c L_q^2 \alpha^2
\end{aligned} \tag{62}
$$

where (i) follows (6) and (7); (ii) is due to the fact that $\|\nabla L(\theta)\| \leq 2L_q$.

Then we further analyze the term A as below:

$$
\begin{aligned}
&\mathbb{E}\bigg[\Big\|\mathbb{E}_{\tau \sim \pi_{\theta_k}}\Big[\sum_{t \geq 0} \gamma^t \nabla_\theta r(s_t, a_t; \theta_k)\Big] - \mathbb{E}_{\tau \sim \pi_{k+1}}\Big[\sum_{t \geq 0} \gamma^t \nabla_\theta r(s_t, a_t; \theta_k)\Big]\Big\|\bigg] \\
&\overset{(i)}{=} \mathbb{E}\bigg[\Big\|\frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d(\cdot,\cdot;\pi_{\theta_k})}\big[\nabla_\theta r(s,a;\theta_k)\big] - \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d(\cdot,\cdot;\pi_{k+1})}\big[\nabla_\theta r(s,a;\theta_k)\big]\Big\|\bigg] \\
&\overset{(ii)}{\leq} \frac{2}{1-\gamma} \cdot \max_{s \in \mathcal{S}, a \in \mathcal{A}} \|\nabla_\theta r(s,a;\theta_k)\| \cdot \mathbb{E}\big[\|d(\cdot,\cdot;\pi_{\theta_k}) - d(\cdot,\cdot;\pi_{k+1})\|_{TV}\big] \\
&\overset{(iii)}{\leq} \frac{2L_r}{1-\gamma}\mathbb{E}\big[\|d(\cdot,\cdot;\pi_{\theta_k}) - d(\cdot,\cdot;\pi_{k+1})\|_{TV}\big] \\
&\overset{(iv)}{\leq} 2L_q C_d \mathbb{E}\Big[\|Q^{\text{soft}}_{r_{\theta_k},\pi_{\theta_k}} - Q^{\text{soft}}_{r_{\theta_k},\pi_k}\|\Big] \\
&\overset{(v)}{\leq} 2L_q C_d \sqrt{|\mathcal{S}| \cdot |\mathcal{A}|}\mathbb{E}\Big[\|Q^{\text{soft}}_{r_{\theta_k},\pi_{\theta_k}} - Q^{\text{soft}}_{r_{\theta_k},\pi_k}\|_\infty\Big]
\end{aligned} \tag{63}
$$

where (i) follows the definition $d(s,a;\pi) = (1-\gamma)\pi(a|s)\sum_{t\geq 0}\gamma^t \mathcal{P}^\pi(s_t = s|s_0 \sim \eta)$; (ii) is due to distribution mismatch between two visitation measures; (iii) follows the inequality (9a) in Assumption 2; the inequality (iv) follows Lemma 4 and the fact that $\pi_{\theta_k}(\cdot|s) \propto \exp\big(Q^{\text{soft}}_{r_{\theta_k},\pi_{\theta_k}}(s,\cdot)\big)$, $\pi_{k+1}(\cdot|s) \propto \exp\big(Q^{\text{soft}}_{r_{\theta_k},\pi_k}(s,\cdot)\big)$ and the constant $L_q := \frac{L_r}{1-\gamma}$; (v) follows the conversion between

Frobenius norm and infinity norm. Through plugging the inequality (63) into (62), it leads to

$$
\begin{aligned}
&\mathbb{E}\left[L(\theta_{k+1})\right] \\
&\geq \mathbb{E}\left[L(\theta_k)\right] - 2\alpha L_q \mathbb{E}\left[\left\|\mathbb{E}_{\tau\sim\pi_{\theta_k}}\Big[\sum_{t\geq 0}\gamma^t \nabla_\theta r(s_t,a_t;\theta_k)\Big] - \mathbb{E}_{\tau\sim\pi_{k+1}}\Big[\sum_{t\geq 0}\gamma^t \nabla_\theta r(s_t,a_t;\theta_k)\Big]\right\|\right] \\
&\quad + \alpha\mathbb{E}\left[\|\nabla L(\theta_k)\|^2\right] - 2L_c L_q^2\alpha^2 \\
&\overset{(i)}{\geq} \mathbb{E}\left[L(\theta_k)\right] - 4\alpha C_d L_q^2\sqrt{|\mathcal{S}|\cdot|\mathcal{A}|}\,\mathbb{E}\left[\|Q^{\text{soft}}_{r_{\theta_k},\pi_{\theta_k}} - Q^{\text{soft}}_{r_{\theta_k},\pi_k}\|_\infty\right] + \alpha\mathbb{E}\left[\|\nabla L(\theta_k)\|^2\right] - 2L_c L_q^2\alpha^2
\end{aligned}
$$

where (i) follows the inequality (63).

Rearranging the inequality above and denote $C_1 := 4C_d L_q^2\sqrt{|\mathcal{S}|\cdot|\mathcal{A}|}$, it holds that

$$
\alpha\mathbb{E}\big[\|\nabla L(\theta_k)\|^2\big] \leq 2L_c L_q^2\alpha^2 + \alpha C_1\mathbb{E}\left[\|Q^{\text{soft}}_{r_{\theta_k},\pi_{\theta_k}} - Q^{\text{soft}}_{r_{\theta_k},\pi_k}\|_\infty\right] + \mathbb{E}\big[L(\theta_{k+1}) - L(\theta_k)\big]
$$

Summing the inequality above from $k = 0$ to $K-1$ and dividing both sides by αK, it holds that

$$
\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla L(\theta_k)\|^2\right] \leq 2L_c L_q^2\alpha + \frac{C_1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\|Q^{\text{soft}}_{r_{\theta_k},\pi_{\theta_k}} - Q^{\text{soft}}_{r_{\theta_k},\pi_k}\|_\infty\right] + \mathbb{E}\left[\frac{L(\theta_K) - L(\theta_0)}{K\alpha}\right] \tag{64}
$$

Note that the log-likelihood function $L(\theta_K)$ is negative and $L(\theta_0)$ is a bounded constant. Then we could plug (60) into (64), it holds that

$$
\frac{1}{K}\sum_{K=0}^{K-1}\mathbb{E}\left[\|\nabla L(\theta_K)\|^2\right] = \mathcal{O}(K^{-\sigma}) + \mathcal{O}(K^{-1}) + \mathcal{O}(K^{-1+\sigma}) \tag{65}
$$

which completes the proof for the inequality (11b). □