Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Structural Estimation of Markov Decision Processes in High-Dimensional State Space with Finite-Time Guarantees

Siliang Zeng, Mingyi Hong, Alfredo Garcia

To cite this article:
Siliang Zeng, Mingyi Hong, Alfredo Garcia (2024) Structural Estimation of Markov Decision Processes in High-Dimensional State Space with Finite-Time Guarantees. Operations Research

Published online in Articles in Advance 19 Sep 2024

. https://doi.org/10.1287/opre.2022.0511

https://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Articles in Advance, pp. 1–18
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Crosscutting Areas

Structural Estimation of Markov Decision Processes in High-Dimensional State Space with Finite-Time Guarantees

Siliang Zeng,[a] Mingyi Hong,[a,*] Alfredo Garcia[b]

[a] Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, Minnesota 55455; [b] Department of Industrial and Systems Engineering, Texas A&M University College of Engineering, College Station, Texas 77843

*Corresponding author

Contact: zeng0176@umn.edu, https://orcid.org/0009-0006-0765-5028 (SZ); mhong@umn.edu, https://orcid.org/0000-0003-1263-9365 (MH); alfredo.garcia@tamu.edu, https://orcid.org/0000-0002-2761-7479 (AG)

Received: September 30, 2022
Revised: March 1, 2024
Accepted: July 4, 2024
Published Online in Articles in Advance: September 19, 2024

Area of Review: Machine Learning and Data Science

https://doi.org/10.1287/opre.2022.0511

Abstract. We consider the task of estimating a structural model of dynamic decisions by a human agent based on the observable history of implemented actions and visited states. This problem has an inherent nested structure: In the inner problem, an optimal policy for a given reward function is identified, whereas in the outer problem, a measure of fit is maximized. Several approaches have been proposed to alleviate the computational burden of this nested-loop structure, but these methods still suffer from high complexity when the state space is either discrete with large cardinality or continuous in high dimensions. Other approaches in the inverse reinforcement learning literature emphasize policy estimation at the expense of reduced reward estimation accuracy. In this paper, we propose a *single-loop* estimation algorithm with finite time guarantees that is equipped to deal with high-dimensional state spaces without compromising reward estimation accuracy. In the proposed algorithm, each policy improvement step is followed by a stochastic gradient step for likelihood maximization. We show the proposed algorithm converges to a stationary solution with a finite-time guarantee. Further, if the reward is parameterized linearly, the algorithm approximates the maximum likelihood estimator sublinearly.

Funding: M. Hong and S. Zeng are supported by the National Science Foundation [Grants EPCN-2311007 and CCF-1910385]. This work is also part of AI-CLIMATE: "AI Institute for Climate-Land Interactions, Mitigation, Adaptation, Tradeoffs and Economy" and is supported by the U.S. Department of Agriculture National Institute of Food and Agriculture and the National Science Foundation National AI Research Institutes [Competitive Award 2023-67021-39829]. A. Garcia is partially supported by the Army Research Office [Grant W911NF-22-1-0213].

Supplemental Material: The computer code and data that support the findings of this study are available within this article's supplemental material at https://doi.org/10.1287/opre.2022.0511.

Keywords: inverse reinforcement learning • dynamic discrete choice model

1. Introduction

We consider the task of estimating a structural model of dynamic decisions by a single human agent based on the observable history of implemented actions and visited states. This problem has been studied as the estimation of dynamic discrete choice (DDC) models in econometrics and inverse reinforcement learning (IRL) in artificial intelligence and machine learning research.

Rust (1987) is a seminal piece of literature on dynamic discrete choice estimation. In that paper, the estimation task is formulated as a bilevel optimization problem where the inner problem is a stochastic dynamic programming problem, and the outer problem is the likelihood maximization of observed actions and states. Rust (1987) proposed an iterative nested fixed-point algorithm in which the inner dynamic programming problem is solved repeatedly followed by maximum likelihood updates of the structural parameters. Over the years, a significant amount of literature on alternative estimation methods requiring less computational effort has been developed. For example, Hotz and Miller (1993) and Hotz et al. (1994) proposed two-step algorithms that avoid the repeated solution of the inner stochastic dynamic programming problem. In the first step, a nonparametric estimator of the policy (also referred to as conditional choice probabilities) is obtained, and the inverse of a map from differences in Bellman's value function for different states to randomized policies is computed. In the second step, a pseudo-log-likelihood is maximized. Two-step estimators may suffer from substantial finite sample bias if the estimated policies in the first step are of poor quality. Sequential estimators that are recursively obtained by alternating between pseudo-likelihood maximization

and improved policy estimation are considered in Aguirregabiria and Mira (2002). In general, the computational burden for all these methods is significant when the state space is either discrete with large cardinality, or they are continuous in high dimensions. Discretization may be avoided using forward Monte Carlo simulations (Bajari et al. 2007, Reich 2018), but this also becomes computationally demanding in high dimensions. A constrained optimization approach for maximum likelihood estimation of dynamic discrete choice models is considered in Su and Judd (2012). However, the number of constraints needed to represent Bellman's equation becomes a significant computational burden with discrete state space with large cardinality or continuous state space in high dimensions.

Recent work has addressed the computational challenges posed by high-dimensional state space. For example, in Adusumilli and Eckardt (2019), the authors extend the CCP estimator approach proposed in Hotz and Miller (1993) by considering a functional approximation approach coupled with a temporal difference (TD) algorithm to maximize pseudo-likelihood. In Chernozhukov et al. (2022), the authors consider an approach to adjust the Conditional choice probabilities (CCP) estimator to account for finite simple bias in high-dimensional settings.

The literature in IRL features the seminal work (Ziebart et al. 2008) in which a model for the expert's behavior is formulated as the policy that maximizes entropy subject to a constraint requiring that the expected features under such policy match the empirical averages in the expert's observation data set.[1] The algorithms developed for maximum entropy estimation (Ziebart et al. 2008, 2010; Wulfmeier et al. 2015) 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 value function approximation.

Recent works have developed algorithms to alleviate the computational burden of nested-loop estimation methods. For example, in Garg et al. (2021), the authors propose to transform the standard formulation of IRL into a single-level problem by estimating the Q-function rather than estimating the reward function and associated optimal policy separately. However, the implicit reward function in the Q-function identified is a poor estimate since it is not guaranteed to satisfy Bellman's equation. Finally, Ni et al. (2020) considers an approach called *f*-IRL 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. Although the results reported are based upon a single-loop implementation, the paper does not provide a convergence guarantee to support performance.

In contrast to the lines of works surveyed above, we focus our efforts in developing estimation algorithms with finite-time guarantees for computing high-quality estimators. Toward addressing this challenge, in this paper, we propose a class of new algorithms that only require a finite number of computational steps for a class of (nonlinearly parameterized) structural estimation problems assuming the environment dynamics are known (or samples from the environment dynamics are available in real time). Specifically, the proposed algorithm has a single-loop structure wherein a single-step update of the estimated policy is followed by an update of the reward parameter. We show that the algorithm has strong theoretical guarantees: To achieve certain ϵ-approximate stationary solution for a nonlinearly parameterized problem, it requires $\mathcal{O}(\epsilon^{-2})$ steps of policy and reward updates each. To our knowledge, it is the first algorithm that has finite-time guarantee for the structural estimation of an Markov Decision Process (MDP) under nonlinear parameterization of the reward function. 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 transferring to a new environment, the performance of state-of-the-art reinforcement learning (RL) algorithms, using estimated rewards, outperform those that use rewards recovered from existing IRL and imitation learning benchmarks.

Finally, we consider the extension to the offline case in which the estimation task also includes the environment dynamics. Referring to our recent work (Zeng et al. 2023), we consider a two-stage estimation approach. First, a maximum likelihood model of dynamics is identified. However, this first stage estimator of the environment dynamics may be inaccurate due to limited data coverage. Thus, in the second stage, a "conservative" reward estimator is obtained by introducing a penalty for model uncertainty.

The structure of this paper is as follows. In Section 2, we introduce the basic setting for structural estimation of MDPs. In Section 2.2, we introduce the problem formulation of the maximum likelihood IRL. In Section 3, we discuss the problem approximation in high-dimensional spaces. In Section 4, we introduce a single-loop algorithm for estimation and formalize a finite-time performance guarantee for high-dimensional states. In Section 5, we present the convergence results of our proposed single-loop algorithm. In Section 6, we consider the case with linearly parameterized rewards to show the proposed algorithm converges sublinearly to the maximum likelihood estimator. These results are proven by establishing a duality relationship between maximum entropy IRL and maximum likelihood IRL. In Section 7, we consider the case in which the agent's preferences can be represented by a

reward that is only a function of the state. In Section 8, we outline the extension of the proposed algorithms to the offline case. Finally, in Section 9, we present the numerical results.

2. Background

2.1. Dynamic Discrete Choice Model

We now review the basic setting for dynamic discrete choice model as given for example in Rust (1994). At time $t \geq 0$, the agent implements an action a_t from a finite (discrete) action space $\mathcal{A}$ and receives a reward $r(s_t, a_t; \theta) + \epsilon_t(a_t)$, where $s_t \in S$ is the state at time t, $r(s_t, a_t; \theta)$ is the reward associated to the state-action pair (s_t, a_t) with $\theta \in \mathbb{R}^p$ a parameter and $\epsilon_t(a_t) : \mathbb{R}^{|\mathcal{A}|} \to \mathbb{R}$ is a random perturbation that is observable by the agent (decision maker) but not by the modeler.

Upon implementing the action $a_t \in \mathcal{A}$, the state evolves according to a Markov process with kernel $P(s_{t+1}|s_t, a_t)$. Moreover, let $\mu(\epsilon_t|s_t)$ denote the probability distribution for the random perturbation, where the probability distribution is a function of the state.

Let $\pi(\cdot|s_t, \epsilon_t)$ denote a randomized policy, that is, $\pi(a_t|s_t, \epsilon_t)$ is the probability that action a_t is implemented when the state is s_t and the observed reward perturbation vector is ϵ_t.

The agent's optimal policy is characterized by the value function:

$$V_\theta(s_0, \epsilon_0) = \max_\pi \mathbb{E}_{s_0 \sim \rho, \tau \sim \pi}\left[\sum_{t=0}^{\infty} \gamma^t (r(s_t, a_t; \theta) + \epsilon_t(a_t)) \middle| s_0, \epsilon_0\right],$$

where the expectation is taken with respect to $a_t \sim \pi(\cdot|s_t, \epsilon_t), s_{t+1} \sim P(\cdot|s_t, a_t), \epsilon_{t+1} \sim \mu(\cdot|s_{t+1})$, and $\gamma \in [0, 1)$ is the discount factor. The Bellman equation is

$$\begin{aligned} V_\theta(s_t, \epsilon_t) &= \max_{a_t \in \mathcal{A}} [r(s_t, a_t; \theta) + \epsilon_t(a_t) \\ &\quad + \gamma \mathbb{E}_{s_{t+1} \sim P(\cdot|s_t, a_t), \epsilon_{t+1} \sim \mu(\cdot|s_{t+1})}[V_\theta(s_{t+1}, \epsilon_{t+1})]], \\ &= \max_{a_t \in \mathcal{A}} [Q_\theta(s_t, a_t) + \epsilon_t(a_t)], \end{aligned}$$

where $Q_\theta : S \times \mathcal{A} \mapsto \mathbb{R}$ is the fixed point of the *soft*-Bellman operator:

$$\Lambda_\theta(Q(s_t, a_t)) = r(s_t, a_t; \theta) + \gamma \mathbb{E}_{s_{t+1} \sim P(\cdot|s_t, a_t), \epsilon_{t+1} \sim \mu(\cdot|s_{t+1})} \left[\max_{a \in \mathcal{A}} (Q(s_{t+1}, a) + \epsilon_{t+1}(a))\right]. \quad (1)$$

As the realization of the reward perturbations is not observable by the modeler, a parametrized model of the agent's behavior is a map $\pi_\theta(\cdot|s_t)$ that satisfies Bellman's optimality as follows:

$$\pi_\theta(a_t|s_t) = P\left(a_t \in \arg\max_{a \in \mathcal{A}} [Q_\theta(s_t, a) + \epsilon_t(a)]\right). \quad (2)$$

Assume observations are in the form of expert state-action trajectories $\tau^E = \{(s_t, a_t)\}_{t \geq 0}$ drawn from a ground-truth (or "expert") policy π^E, that is, $a_t \sim \pi^E(\cdot|s_t)$, $s_{t+1} \sim P(\cdot|s_t, a_t)$ and $s_0 \sim \rho(\cdot)$, where $\rho(\cdot)$ denotes the initial distribution of the first state s_0. The expected discounted log-likelihood of observing such trajectory under model π_θ can be written as

$$\begin{aligned} &\mathbb{E}_{\tau^E \sim \pi^E}\left[\sum_{t=0}^{\infty} \gamma^t \log(P(s_{t+1}|s_t, a_t)\pi_\theta(a_t|s_t))\right] \\ &= \mathbb{E}_{\tau^E \sim \pi^E}\left[\sum_{t=0}^{\infty} \gamma^t \log \pi_\theta(a_t|s_t)\right] \\ &\quad + \mathbb{E}_{\tau^E \sim \pi^E}\left[\sum_{t=0}^{\infty} \gamma^t \log P(s_{t+1}|s_t, a_t)\right]. \end{aligned}$$

Given that the term $\mathbb{E}_{\tau^E \sim \pi^E}[\sum_{t=0}^{\infty} \gamma^t \log P(s_{t+1}|s_t, a_t)]$ is independent of the reward parameter θ, the maximum likelihood estimation problem can be formulated as follows:

$$\max_\theta \; L(\theta) := \mathbb{E}_{\tau^E \sim \pi^E}\left[\sum_{t=0}^{\infty} \gamma^t \log \pi_\theta(a_t|s_t)\right] \quad (3a)$$

$$\text{s.t.} \quad \pi_\theta(a_t|s_t) = P\left(a_t \in \arg\max_{a \in \mathcal{A}} [Q_\theta(s_t, a) + \epsilon_t(a)]\right), \quad (3b)$$

where Q_θ is the fixed point of the *soft*-Bellman operator in (1).

In the next section, we review the literature on the entropy-regularized RL model and then highlight the formal equivalence of entropy-regularized IRL with the dynamic discrete choice model just introduced in (3).

2.2. Maximum Likelihood Inverse Reinforcement Learning (ML-IRL)

A recent literature has considered MDP models with information processing costs (Tishby and Polani 2011, Ortega and Braun 2013, Matějka and McKay 2015, Hansen and Miao 2018). In these papers, optimal behavior is modeled as the solution to the following problem:

$$\max_{\pi \in \Pi} \; J_\theta(\pi; \rho) \triangleq \mathbb{E}_{s_0 \sim \rho, \tau^A \sim \pi}\left[\sum_{t=0}^{\infty} \gamma^t (r(s_t, a_t; \theta) - c(\pi(\cdot|s_t))\right],$$

where $\rho(\cdot)$ denotes the initial distribution of the first state s_0, τ^A is a trajectory generate from the agent policy π, and $c(\cdot)$ is a function representing the information processing cost. A common specification is $c(\pi(\cdot|s_t)) = \alpha D_{\mathrm{KL}}(\pi(\cdot|s_t)||\pi_0(\cdot|s_t))$, where $D_{\mathrm{KL}}(\pi(\cdot|s_t)||\pi_0(\cdot|s_t)) = \sum_{a \in \mathcal{A}} \pi(a|s_t) \log \frac{\pi(a|s_t)}{\pi_0(a|s_t)}$ is Kullback-Leibler divergence between $\pi(\cdot|s_t)$ and a reference (or default) policy $\pi_0(\cdot|s_t)$ and $\alpha \geq 0$ is a scale parameter. As the objective function above can be rescaled by $\frac{1}{\alpha}$, we can set $\alpha = 1$. To model no prior knowledge, the reference policy π_0

is the uniformly random policy, that is, $\pi_0(a) = \frac{1}{|\mathcal{A}|}$ for any $a \in \mathcal{A}$. In this case, we can further rewrite the problem as

$$\max_{\pi \in \Pi} J_\theta(\pi; \rho)$$
$$\triangleq \mathbb{E}_{s_0 \sim \rho, \tau^{\mathrm{A}} \sim \pi}\left[\sum_{t=0}^{\infty} \gamma^t (r(s_t, a_t; \theta) + \mathcal{H}(\pi(\cdot|s_t)))\right] + \frac{\log|\mathcal{A}|}{1-\gamma},$$

where $\mathcal{H}(\pi(\cdot|s_t)) = -\sum_{a \in A} \pi(a|s_t) \log \pi(a|s_t)$ is the entropy of $\pi(\cdot|s_t)$. This model has also been recently used in the RL literature (Haarnoja et al. 2017, 2018; Cayci et al. 2021; Cen et al. 2022) where it is commonly referred to as an *entropy regularized* MDP.

Denoting the expert policy as π^{E} and assuming an entropy-regularized MDP model for behavior. and the model for dynamic behavior described, the IRL problem can be formulated as follows:

$$\max_{\theta} \quad L(\theta) := \mathbb{E}_{\tau^{\mathrm{E}} \sim \pi^{\mathrm{E}}}\left[\sum_{t=0}^{\infty} \gamma^t \log \pi_\theta(a_t|s_t)\right] \tag{4a}$$

$$\text{s.t.} \quad \pi_\theta := \arg\max_{\pi} \mathbb{E}_{s_0 \sim \rho, \tau^{\mathrm{A}} \sim \pi}\left[\sum_{t=0}^{\infty} \gamma^t (r(s_t, a_t; \theta) + \mathcal{H}(\pi(\cdot|s_t)))\right]. \tag{4b}$$

When the reward perturbations $\epsilon_t(a)$ follow independent and identically distributed (i.i.d.) Gumbel distribution with zero mean and variance $\frac{\pi^2}{6}$ for $a \in A$, the models of Behavior (3b) and (4b) are equivalent (see proposition 1 in Mai and Jaillet 2020). Specifically, the fixed point $Q_\theta(s,a)$ of the Bellman operator Λ_θ in (1) and the optimal policy (2) are of the form:

$$Q_\theta(s,a) := r(s,a;\theta) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V_\theta(s')]. \tag{5a}$$

$$V_\theta(s) = \log\left(\sum_{\tilde{a} \sim \mathcal{A}} \exp Q_\theta(s, \tilde{a})\right), \tag{5b}$$

$$\pi_\theta(a|s) = \frac{\exp(Q_\theta(s,a))}{\sum_{\tilde{a} \in \mathcal{A}} \exp(Q_\theta(s,\tilde{a}))}. \tag{5c}$$

As has been shown in Haarnoja et al. (2018) and Cen et al. (2022), the policy described in (5c) corresponds to the optimal policy in (4b) so that

$$V_\theta(s) = \max_{\pi} \mathbb{E}_{s_0 \sim \rho, \tau^{\mathrm{A}} \sim \pi}\left[\sum_{t=0}^{\infty} \gamma^t (r(s_t, a_t; \theta) + \mathcal{H}(\pi(\cdot|s_t))) \middle| s_0 = s\right]. \tag{6}$$

2.3. Computational Effort and Estimation Quality of Existing Algorithms

The existing solution and approximation methodologies for solving (3) (or equivalently, (4)) are ill equipped for dealing with the high-dimensional state space. For example, the algorithms considered in Rust (1994), Ziebart et al. (2010), and Wulfmeier et al. (2015) rely on a nested-loop structure that requires the solution of a fixed-point problem in the inner loop before making any updates to the parameter estimates for the outer loop. Evidently, in high-dimensional environments. the inner-loop solution renders the nested-loop structure computationally intractable.

Similarly, with a high-dimensional continuous state, a discretization approach (Su and Judd 2012) to solving the inner problem (3b) (or equivalently, (4b)) is computationally intractable. Forward Monte Carlo simulations (Bajari et al. 2007, Reich 2018) are an alternative to discretization, but this is also computationally demanding in high dimensions.

Approximation algorithms (Hotz and Miller 1993, Hotz et al. 1994, Ni et al. 2020, Garg et al. 2021) reduce the computational burden of the nested-loop structure. However, the resulting estimates may be of poor quality. For example, the CCP estimator from Hotz and Miller (1993) and Hotz et al. (1994) may suffer from finite sample bias because in the high-dimensional state space, initial policy estimates (i.e., conditional choice probabilities) based on empirical frequencies are likely of poor quality. Sequential estimators (Aguirregabiria and Mira 2002) reduce bias at the expense of significant computational burden. The reward estimates in Garg et al. (2021) do not approximate a solution to the inner problem and are thus likely to be of poor quality. Recently, Adusumilli and Eckardt (2019) and Chernozhukov et al. (2022) have proposed approaches to account for finite-sample bias in CCP estimators in high-dimensional environments.

In the present paper, we introduce a new class of single-loop algorithms that exhibits finite-time guarantees of performance for solving (3)) (or more precisely, its approximated version to be introduced in the next section).

As many papers in the dynamic discrete choice (DDC) estimation literature (Rust 1994) rely on a two-stage approach to estimating dynamics (first stage) and rewards (second stage), the results obtained in this paper address the computational complexity of the *second*-stage estimation task. This issue was ignored in Rust (1987) due the scale of the problem. However, computational complexity is an important concern in high-dimensional environments.

In Section 8, we also discuss how to extend the proposed method to the two-stage problem/offline setting where estimating dynamics should be considered.

3. Problem Approximation in High-Dimensional State Space

In practice, the IRL problem (4) (and its equivalent (3)) can only be approximated with a *finite* set of observed trajectories because the ground-truth behavior model (or "expert" policy) π^{E} is not known. Let $\mathcal{D} := \{\tau^{\mathrm{E}}\}$ denote a finite data set of state-action trajectories

independently drawn from the expert policy and the environment dynamics. Let $\tau^{\mathrm{E}} \sim \mathcal{D}$ denote a uniformly sampled trajectory from $\mathcal{D}$. Using a finite data set, a natural choice for an *empirical* approximation to the estimation problem is the following:

$$\max_{\theta} \quad \tilde{L}(\theta;\mathcal{D}) := \mathbb{E}_{\tau^{\mathrm{E}}\sim\mathcal{D}}\left[\sum_{t=0}^{\infty}\gamma^t \log \pi_\theta(a_t|s_t)\right] \tag{7a}$$

$$\text{s.t.} \quad \pi_\theta(a_t|s_t) := \arg\max_{\pi} \mathbb{E}_{s_0\sim\rho,\tau^{\mathrm{A}}\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t (r(s_t,a_t;\theta)+\mathcal{H}(\pi(\cdot|s_t)))\right]. \tag{7b}$$

However, with high-dimensional state space, the above approximation $\tilde{L}(\theta,\mathcal{D})$ is likely to incur significant error because the observed transitions in the data may not adequately describe the ground-truth transition kernel. In what follows, we introduce a different *surrogate* empirical objective $\hat{L}(\theta,\mathcal{D})$, which provides a better approximation to the original likelihood function $L(\theta)$ given in (4) with a high-dimensional state.

To motivate the definition of $\hat{L}(\theta,\mathcal{D})$, let us start by expressing the likelihood function $L(\theta) := \mathbb{E}_{\tau^{\mathrm{E}}\sim\pi^{\mathrm{E}}}[\sum_{t=0}^{\infty}\gamma^t \log \pi_\theta(a_t|s_t)]$ in terms of the difference in expected value:

$$\begin{aligned}
&L(\theta)\\
&= \mathbb{E}_{\tau^{\mathrm{E}}\sim\pi^{\mathrm{E}}}\left[\sum_{t=0}^{\infty}\gamma^t \log \pi_\theta(a_t|s_t)\right]\\
&\overset{(i)}{=} \mathbb{E}_{\tau^{\mathrm{E}}\sim\pi^{\mathrm{E}}}\left[\sum_{t=0}^{\infty}\gamma^t \log\left(\frac{\exp Q_\theta(s_t,a_t)}{\sum_{a\in\mathcal{A}}\exp Q_\theta(s_t,a)}\right)\right]\\
&\overset{(ii)}{=} \mathbb{E}_{\tau^{\mathrm{E}}\sim\pi^{\mathrm{E}}}\left[\sum_{t=0}^{\infty}\gamma^t (Q_\theta(s_t,a_t)-V_\theta(s_t))\right]\\
&\overset{(iii)}{=} \mathbb{E}_{\tau^{\mathrm{E}}\sim\pi^{\mathrm{E}}}\left[\sum_{t=0}^{\infty}\gamma^t (r(s_t,a_t;\theta)+\gamma\mathbb{E}_{s_{t+1}\sim P(\cdot|s_t,a_t)}[V_\theta(s_{t+1})] - V_\theta(s_t))\right]\\
&= \mathbb{E}_{\tau^{\mathrm{E}}\sim\pi^{\mathrm{E}}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t;\theta)\right]\\
&\quad + \sum_{t=0}^{\infty}\gamma^{t+1}\mathbb{E}_{(s_t,a_t)\sim\tau^{\mathrm{E}}}[\mathbb{E}_{s_{t+1}\sim P(\cdot|s_t,a_t)}[V_\theta(s_{t+1})]]\\
&\quad - \mathbb{E}_{\tau^{\mathrm{E}}\sim\pi^{\mathrm{E}}}\left[\sum_{t=0}^{\infty}\gamma^t V_\theta(s_t)\right]\\
&= \mathbb{E}_{\tau^{\mathrm{E}}\sim\pi^{\mathrm{E}}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t;\theta)\right] - \mathbb{E}_{s_0\sim\rho(\cdot)}[V_\theta(s_0)],
\end{aligned} \tag{8}$$

where (i) follows the closed-form expression of the optimal policy π_θ in (5c), (ii) follows the expression of the soft value function V_θ in (5b), and (iii) follows from the fixed point definition in (5a).

Observe that in the above decomposition, the first term is related to the expert policy π^{E}, whereas the second term is related to the initial distribution ρ and the transition kernel P. Note that in practice, we only have limited observations of expert trajectories from a fixed data set $\mathcal{D}$, but cannot directly sample the trajectory from the expert policy π^{E} in an online manner. Hence, we need to construct an estimation problem that utilizes limited observations of expert trajectories to approximate the original maximum likelihood objective in (8). Because we assume the *online* setting in which the transition kernel P and the initial distribution ρ are either available for access or known, we can construct a *surrogate* approximation to the likelihood as follows:

$$\hat{L}(\theta;\mathcal{D}) := \mathbb{E}_{\tau\sim\mathcal{D}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t;\theta)\right] - \mathbb{E}_{s_0\sim\rho}[V_\theta(s_0)]. \tag{9}$$

In contrast, if we conduct the same analysis on the *empirical* approximation $\tilde{L}(\theta;\mathcal{D})$ presented in (7), we obtain

$$\begin{aligned}
&\tilde{L}(\theta;\mathcal{D})\\
&= \mathbb{E}_{\tau^{\mathrm{E}}\sim\mathcal{D}}\left[\sum_{t=0}^{\infty}\gamma^t \log \pi_\theta(a_t|s_t)\right]\\
&\overset{(i)}{=} \mathbb{E}_{\tau^{\mathrm{E}}\sim\mathcal{D}}\left[\sum_{t=0}^{\infty}\gamma^t \log\left(\frac{\exp Q_\theta(s_t,a_t)}{\sum_{a\in\mathcal{A}}\exp Q_\theta(s_t,a)}\right)\right]\\
&\overset{(ii)}{=} \mathbb{E}_{\tau^{\mathrm{E}}\sim\mathcal{D}}\left[\sum_{t=0}^{\infty}\gamma^t (Q_\theta(s_t,a_t)-V_\theta(s_t))\right]\\
&\overset{(iii)}{=} \mathbb{E}_{\tau\sim\mathcal{D}}\left[\sum_{t=0}^{\infty}\gamma^t (r(s_t,a_t;\theta)+\gamma\mathbb{E}_{s_{t+1}\sim P(\cdot|s_t,a_t)}[V_\theta(s_{t+1})] - V_\theta(s_t))\right]\\
&= \mathbb{E}_{\tau\sim\mathcal{D}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t;\theta)\right]\\
&\quad + \sum_{t=0}^{\infty}\gamma^{t+1}\mathbb{E}_{(s_t,a_t)\sim\mathcal{D}}[\mathbb{E}_{s_{t+1}\sim P(\cdot|s_t,a_t)}[V_\theta(s_{t+1})]]\\
&\quad - \mathbb{E}_{\tau\sim\mathcal{D}}\left[\sum_{t=0}^{\infty}\gamma^t V_\theta(s_t)\right]\\
&= \underbrace{\left(\mathbb{E}_{\tau\sim\mathcal{D}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t;\theta)\right] - \mathbb{E}_{s_0\sim\mathcal{D}}[V_\theta(s_0)]\right)}_{\text{T1: surrogate likelihood}}\\
&\quad + \left(\sum_{t=0}^{\infty}\gamma^{t+1}\mathbb{E}_{(s_t,a_t)\sim\mathcal{D},s_{t+1}\sim P(\cdot|s_t,a_t)}[V_\theta(s_{t+1})]\right.\\
&\quad \left. - \sum_{t=0}^{\infty}\gamma^{t+1}\mathbb{E}_{(s_t,a_t,s_{t+1})\sim\mathcal{D}}[V_\theta(s_{t+1})]\right),
\end{aligned} \tag{10}$$

where the second term is the error introduced by approximating the transition using finite data.

From the above analysis, we argue that the *surrogate* approximation $\hat{L}(\theta;\mathcal{D})$ in (9) is a more accurate objective function compared with the *empirical* likelihood $\tilde{L}(\theta;\mathcal{D})$ in (7a). Below, we show under a mild assumption, $\hat{L}(\theta;\mathcal{D})$ can well approximate $L(\theta)$ when data are large enough.

Assumption 1. *For any reward parameter θ, the following condition holds:*

$$0 \le r(s,a;\theta) \le C_r, \qquad \forall s \in S, a \in \mathcal{A}, \tag{11}$$

where $C_r > 0$ is a fixed constant.

Lemma 1. *Suppose Assumption 1 holds. Consider the likelihood function $L(\theta)$ in (4a) and its surrogate empirical version $\hat{L}(\theta;\mathcal{D})$ defined in (9). Then, with probability greater than $1-\delta$, we have*

$$|L(\theta) - \hat{L}(\theta;\mathcal{D})| \le \frac{C_r}{1-\gamma}\sqrt{\frac{\ln(2/\delta)}{2|\mathcal{D}|}}. \tag{12}$$

The proof of Lemma 1 can be found in the Online Appendix.

In the rest of this work, we will consider the following surrogate estimation problem:

$$\max_{\theta} \quad \hat{L}(\theta;\mathcal{D}) := \mathbb{E}_{\tau\sim\mathcal{D}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t;\theta)\right] - \mathbb{E}_{s_0\sim\rho}[V_\theta(s_0)] \tag{13a}$$

$$\text{s.t.} \quad \pi_\theta(a_t|s_t) := \arg\max_{\pi} \mathbb{E}_{s_0\sim\rho,\tau^{\mathrm{A}}\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t(r(s_t,a_t;\theta) + \mathcal{H}(\pi(\cdot|s_t)))\right]. \tag{13b}$$

4. Proposed Algorithm

The main idea in the proposed algorithm is to alternate between one step of policy update to improve the solution of the lower-level problem, and one step of the parameter update that improves the upper-level likelihood objective. 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.

In Sections 4 and 5, we will design an algorithm to solve the approximated maximum likelihood problem (13). We emphasize that, in Sections 4–7, we assume an *online* setting where the learner knows the transition kernel $P(s_{t+1}|s_t,a_t)$ or can sample from it. The motivation is that understanding how to develop efficient algorithms for the *online* setting is the basis for addressing the more challenging *offline* setting. In Section 8, we will briefly outline how to extend this work to the offline setting. Below we present the details of our algorithm at a given iteration k.

4.1. Policy Improvement Step

Let us consider optimizing the lower-level problem (4b), when the reward parameter θ_k is held fixed. Toward this end, we define the so-called soft Q-function and soft value functions under a given policy-reward pair (π_k,θ_k):

$$V_k(s) = \mathbb{E}_{\tau^{\mathrm{A}}\sim\pi_k}\left[\sum_{t=0}^{\infty}\gamma^t(r(s_t,a_t;\theta_k) + \mathcal{H}(\pi_k(\cdot|s_t)))\middle| s_0 = s\right], \tag{14}$$

$$Q_k(s,a) = r(s,a;\theta_k) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}[V_k(s')]. \tag{15}$$

Similarly, if the policy is *optimal* for a given parameter θ (as defined in (4b)), then we will denote the associated soft Q-function and soft value function as Q_θ and V_θ.

To obtain an estimate of the policy at iteration k, let us suppose that we have access to an estimate of the soft Q-function, denoted as $\hat{Q}_k(s,a)$, which satisfies $\|\hat{Q}_k - Q_k\|_\infty \le \epsilon_{\text{app}}$, with $\epsilon_{\text{app}} > 0$ being the approximation error. Then the estimated policy will be generated according to

$$\pi_{k+1}(a|s) \propto \exp(\hat{Q}_k(s,a)), \qquad \forall s \in S, a \in \mathcal{A}. \tag{16}$$

When $\epsilon_{\text{app}} = 0$, or equivalently when $\hat{Q}_k(s,a) = Q_k(s,a)$, $\forall s \in S, a \in \mathcal{A}$, and when $r(\cdot,\cdot;\theta_k)$ is fixed, the above update is referred to as the *soft policy iteration*; it is known that the policy will be monotonically improved by soft policy iteration and will converge linearly to the optimal policy (Cen et al. 2022, theorem 1). In practice, when we do not have direct access to the exact soft Q-function Q_k, one could use an *estimated* soft Q-function $\hat{Q}_k$ to perform the approximated soft policy iteration in (16), which can be obtained by following the update schemes in soft Q-learning (Haarnoja et al. 2017) or soft actor-critic (SAC) (Haarnoja et al. 2018).

4.2. Reward Optimization Step

We propose to use a stochastic gradient-type algorithm to optimize the reward parameter θ. Toward this end, let us first derive the exact gradient $\nabla_\theta L(\theta)$. See the supplementary material for detailed proof.

Lemma 2. *The gradient of the $L(\theta)$ and $\hat{L}(\theta;\mathcal{D})$, as defined in (4a) and (9), respectively, can be expressed as*

$$\nabla_\theta L(\theta) = \mathbb{E}_{\tau^{\mathrm{E}}\sim\pi^{\mathrm{E}}}\left[\sum_{t=0}^{\infty}\gamma^t\nabla_\theta r(s_t,a_t;\theta)\right] - \mathbb{E}_{\tau^{\mathrm{A}}\sim\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t\nabla_\theta r(s_t,a_t;\theta)\right], \tag{17a}$$

$$\nabla_\theta \hat{L}(\theta;\mathcal{D}) = \mathbb{E}_{\tau^{\mathrm{E}}\sim\mathcal{D}}\left[\sum_{t=0}^{\infty}\gamma^t\nabla_\theta r(s_t,a_t;\theta)\right] - \mathbb{E}_{\tau^{\mathrm{A}}\sim\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t\nabla_\theta r(s_t,a_t;\theta)\right]. \tag{17b}$$

We note that the gradient expression (17a) takes the same form as the one given in a recent work (Sanghvi et al. 2021,

equation (1)). However, our proof that focuses on the *infinite* horizon case is different. Moreover, we further derive the gradient expression of the sample-based estimation problem $\hat{L}(\theta;\mathcal{D})$, which has not been considered in Sanghvi et al. (2021).

In order to obtain stochastic estimators of the empirical gradient $\nabla_\theta \hat{L}(\theta_k;\mathcal{D})$, we take two approximation steps: (1) approximate the optimal policy π_{θ_k} by π_{k+1} in (16) because the optimal policy π_{θ_k} is not available throughout the algorithm and (2) sample the trajectory τ^{A} from the current policy π_{k+1}.

Following the approximation steps mentioned above, we construct a stochastic estimator g_k to approximate the empirical gradient $\nabla_\theta \hat{L}(\theta_k;\mathcal{D})$ in (17b) as follows:

$$g_k := h(\theta_k;\tau_k^{\mathrm{E}}) - h(\theta_k;\tau_k^{\mathrm{A}}), \tag{18}$$

where $h(\theta;\tau) := \sum_{t=0}^{\infty}\gamma^t \nabla_\theta r(s_t,a_t;\theta)$. With the stochastic gradient estimator g_k, the reward parameter θ_k is updated as

$$\theta_{k+1} = \theta_k + \alpha g_k, \tag{19}$$

where α is the step size in updating the reward parameter.

Algorithm 1 summarizes the proposed two-step approach for solving the IRL problem (4). It is worth mentioning that the proposed algorithm can also be used to solve the DDC problem (3) due to the equivalence between (3) and (4).

Algorithm 1 (ML-IRL)

Input: Initialize reward parameter θ_0 and policy π_0. Set the reward parameter's step size as α.

for $k = 0, 1, \ldots, K-1$ **do**

Policy Evaluation: Approximate the soft Q-function $Q_k(\cdot,\cdot)$ by $\hat{Q}_k(\cdot,\cdot)$.

Policy Improvement: $\pi_{k+1}(a|s) \propto \exp(\hat{Q}_k(s,\cdot))$, $\forall s \in S, a \in \mathcal{A}$. (Lower-Level Update)

Data Sampling I: Sample a trajectory τ_k^{E} from the data set $\mathcal{D}$.

Data Sample II: Sample a trajectory $\tau_k^{\mathrm{A}} := \{s_t,a_t\}_{t\geq 0}$ from the current policy π_{k+1}

Estimating Gradient: $g_k := h(\theta_k,\tau_k^{\mathrm{E}}) - h(\theta_k,\tau_k^{\mathrm{A}})$ where $h(\theta,\tau) := \sum_{t=0}^{\infty}\gamma^t \nabla_\theta r(s_t,a_t;\theta)$

Reward Parameter Update: $\theta_{k+1} := \theta_k + \alpha g_k$ (Upper-Level Update)

end for

Before closing this section, let us note that the generic alternating update strategy adopted by our algorithm is efficient, because completely solving the policy optimization subproblem all the time could be redundant and could induce heavy computation burden. Such a kind of strategy has been used in many other RL-related settings as well. For example, the well-known AC algorithm for policy optimization (Konda and Tsitsiklis 1999, Hong et al. 2020, Wu et al. 2020) alternates between one step of policy update and one step of critic parameter update. However, these types of algorithms are known to be challenging to analyze, partly because when the inner problem (e.g., the policy optimization problem (4b)) is not solved exactly, the update direction for the main parameter (e.g., θ in (4)) can be very far from the desired descent directions. That is, g_k in (18) can be a very coarse approximation of the exact gradient $\nabla_\theta \hat{L}(\theta_k;\mathcal{D})$ as expressed in (17b). In the subsequent sections, we develop techniques to address the above-mentioned changes.

5. Theoretical Analysis

Our analysis is based on the so-called *two-timescale* stochastic approximation (TTSA) approach (Borkar 1997, Hong et al. 2020), 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 π_{k+1} to track the optimal π_{θ_k}, so that the gradient estimate g_k will stay close to the gradient $\nabla_\theta \hat{L}(\theta_k)$. Indeed, Algorithm 1 has the desired two time scale phenomenon because the policy update (16) converges linearly to the optimal policy under a fixed reward function (Cen et al. 2022, theorem 2) (hence it is fast), whereas the reward parameter update does not have such linear convergence property (hence it is slow). To begin our analysis, let us first present a few technical assumptions.

Assumption 2 (Ergodic Dynamics). *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*

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

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

Assumption 2 assumes the Markov chain mixes at a geometric rate. It is a common assumption in the literature of RL (Bhandari et al. 2018, Zou et al. 2019, Wu et al. 2020), which holds for any time-homogeneous Markov chain with finite-state space or any uniformly ergodic Markov chain with general state space.

Assumption 3 (Lipschitz Reward). *For any $s \in S$, $a \in \mathcal{A}$, and any reward parameter θ, the following holds:*

$$|\nabla_\theta r(s,a;\theta)| \leq L_r, \tag{20a}$$

$$|\nabla_\theta r(s,a;\theta_1) - \nabla_\theta r(s,a;\theta_2)| \leq L_g \|\theta_1 - \theta_2\|, \tag{20b}$$

where L_r and L_g are positive constants.

Assumption 3 assumes that the parameterized reward function has bounded gradient and is Lipschitz smooth. Such assumptions in Lipschitz property are common in the literature of min-max/bilevel optimization (Hong et al. 2020, Jin et al. 2020, Chen et al. 2021, Guan et al.

2021, Khanduri et al. 2021). Based on Assumptions 1–3, we next provide the following Lipschitz properties.

Lemma 3. *Suppose Assumptions 1–3 hold. There are positive constant L_q and L_c such that the following results hold for any reward parameter θ_1 and θ_2:*

$$|Q_{\theta_1}(s,a) - Q_{\theta_2}(s,a)| \le L_q\|\theta_1 - \theta_2\|, \quad \forall s \in S, a \in \mathcal{A}, \tag{21a}$$

$$\|\nabla_\theta \hat{L}(\theta_1; \mathcal{D}) - \nabla_\theta \hat{L}(\theta_2; \mathcal{D})\| \le L_c\|\theta_1 - \theta_2\|, \tag{21b}$$

where $Q_\theta(\cdot,\cdot)$ denotes the soft Q-function under the reward parameter θ and the policy π_θ.

The full proof of the result is delegated to the Online Appendix.

Next, we present the main results, which show the convergence speed of the policy $\{\pi_k\}_{k\ge0}$ and the reward parameter $\{\theta_k\}_{k\ge0}$ in Algorithm 1. Please see the appendix for the detailed proof.

Theorem 1. *Suppose Assumptions 1 and 2 hold. Let K denote the total number of iterations to be run by the algorithm. Let us select $\alpha := \frac{\alpha_0}{K^\sigma}$ for the reward update step (19), where $\alpha_0 > 0$ and $\sigma \in (0,1)$ are some fixed constants. Then the following holds:*

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\,|\log\pi_{k+1} - \log\pi_{\theta_k}|_\infty] = \mathcal{O}(K^{-1}) + \mathcal{O}(K^{-\sigma}) + \mathcal{O}(\epsilon_{\text{app}}), \tag{22a}$$

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla_\theta \hat{L}(\theta_k; \mathcal{D})\|^2] = \mathcal{O}(K^{-\sigma}) + \mathcal{O}(K^{-1+\sigma}) + \mathcal{O}(K^{-1}) + \mathcal{O}(\epsilon_{\text{app}}), \tag{22b}$$

where

$$\|\log\pi_{k+1} - \log\pi_{\theta_k}\|_\infty := \max_{s\in S, a\in\mathcal{A}} |\log\pi_{k+1}(a|s) - \log\pi_{\theta_k}(a|s)|.$$

In particular, if setting $\sigma = 1/2$, then both quantities in (22a) and (22b) converge with the rate $\mathcal{O}(K^{-1/2}) + \mathcal{O}(\epsilon_{\text{app}})$.

In Theorem 1, we present the finite-time guarantee for the convergence of the Algorithm 1. We note that our theoretical guarantee is different from the existing works, such as Cen et al. (2022), who showed the convergence rate of soft policy iteration under a *fixed* reward function. Theorem 1 analyzes a more challenging setting where *both* the policy and reward parameter are kept changing. To our knowledge, this is the first result that characterizes the finite-time convergence for an algorithm developed for either the structural estimation problem (3) or the maximum likelihood IRL problem (4). In the following result, we characterize the dimension dependence of the performance of the policy estimated with Algorithm 1.

Remark 1. It is worth mentioning here that the Lipschitz constant L_c in (21b) is given by

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

where C_d is a constant given in (30). Hence, L_c and the subsequent convergence rate of the algorithm in Theorem 1 are dependent on the dimension of the problem (i.e., the size of the state and action space). However, the empirical evidence (to be presented in Section 9) strongly indicates that the proposed algorithm performs well with high-dimensional neural network representations. This is mainly because our formulation allows us to directly take (approximate) gradient steps on updating θ_k, and that for fixed reward parameterization θ_k, the lower-level policy optimization problem we are interested in has a closed-form solution (as a function of the corresponding Q_k). We believe that the extension of the analysis for our algorithm with function approximations (for the parameterized Q-function and the policy) will result in bounds that have less dependence on the dimension of the basis at the expense of additional approximation error term. The extension of our convergence analysis with function approximations is left for future research.

6. Linearly Parameterized Reward Function Case

The result in Theorem 1 can be further strengthened when rewards are a linear function of (possibly nonlinear) features, that is, $r(s,a;\theta) = \phi(s,a)^\top\theta$ with $\phi : \mathbb{R}^{|S|\times|A|} \to \mathbb{R}^p$, and the distribution of observations is consistent with optimal behavior for a ground truth parameter θ^*, $\pi^E = \pi_{\theta^*}$.

In this setting, the result in Theorem 1 can be strengthened to finite-time convergence to the optimal solution. To show this result, we first establish a duality relationship between the estimation problem in (13) and the maximum entropy estimator (Ziebart et al. 2013) that is the solution to the following problem:

$$\max_\pi \quad -\mathbb{E}_{\tau^A\sim\pi}\left[\sum_{t=0}^\infty \gamma^t \log\pi(a_t|s_t)\right] \tag{23a}$$

$$\text{s.t.} \quad \mathbb{E}_{\tau^A\sim\pi}\left[\sum_{t=0}^\infty \gamma^t\phi(s_t,a_t)\right] = \mathbb{E}_{\tau^E\sim\mathcal{D}}\left[\sum_{t=0}^\infty \gamma^t\phi(s_t,a_t)\right], \tag{23b}$$

$$\sum_{a_t\in\mathcal{A}} \pi(a_t|s_t) = 1, \quad \forall s_t \in S, t \ge 0, \tag{23c}$$

$$\pi(a_t|s_t) \ge 0, \quad \forall s_t \in S, a_t \in \mathcal{A}, t \ge 0, \tag{23d}$$

where (23b) requires that the expected discounted feature value under the model matches the expected discounted feature under the finite data set $\mathcal{D}$ of collected expert trajectories. When the expert policy is known or available for access, the maximum entropy estimation problem is defined as in (23) by replacing (23b) with

$$\mathbb{E}_{\tau^A\sim\pi}\left[\sum_{t=0}^\infty \gamma^t\phi(s_t,a_t)\right] = \mathbb{E}_{\tau^E\sim\pi^E}\left[\sum_{t=0}^\infty \gamma^t\phi(s_t,a_t)\right]. \tag{24}$$

The following result formalizes the relationship between the maximum entropy estimation problem (23) and the estimation problem (13).

Please see the detailed proof in the appendix.

Theorem 2. *Under linear parameterization for reward function $r(s,a;\theta)=\phi(s,a)^\top\theta$, the estimation problem defined in (13) (respectively, the maximum likelihood IRL problem (4)) is the Lagrangian dual of the maximum entropy estimation problem (23) (respectively, the problem defined by (23a), (24), (23c), (23d)). Moreover, strong duality holds between the two problems.*

Corollary 1. (i) *The surrogate objective defined in (13a) (dual objective) is a concave function of θ.* (ii) *If the ground-truth reward values $r(s,\tilde{a};\theta^*)$ for a reference action $\tilde{a}\in A$ and $s\in S$ are known, the optimal solution to (13) is unique.*

Proof. The first result is a direct consequence of Theorem 2 because the estimation problem (13) is a dual problem. Then we prove (ii) by contradiction. Let $\hat{\theta}_1,\hat{\theta}_2$ denote two distinct solutions of the estimation problem (13), which is the dual problem with respect to (w.r.t.) the maximum entropy IRL problem (23). From (17b), it follows that

$$\nabla_\theta\hat{L}(\hat{\theta}_i;\mathcal{D})=\mathbb{E}_{\tau^{\mathrm{E}}\sim\mathcal{D}}\left[\sum_{t=0}^{\infty}\gamma^t\phi(s_t,a_t;\theta)\right]-\mathbb{E}_{\tau^{\mathrm{A}}\sim\pi_{\hat{\theta}_i}}\left[\sum_{t=0}^{\infty}\gamma^t\phi(s_t,a_t;\theta)\right]=0,\quad i=1,2. \tag{25}$$

Let $Q_{\hat{\theta}_i}$ denote the unique fixed point of the soft-Bellman operator and $\tilde{Q}_{\hat{\theta}_i}(s,a):=Q_{\hat{\theta}_i}(s,a)-Q_{\hat{\theta}_i}(s,\tilde{a})$ for all $a\in A$. The following mapping (from the parameter space to the policy space)

$$\pi_{\hat{\theta}_i}(a|s):=\frac{\exp\tilde{Q}_{\hat{\theta}_i}(s,a)}{\sum_{a'\in A}\exp\tilde{Q}_{\hat{\theta}_i}(s,a')}$$

is one-to-one (see proposition 1 in Hotz and Miller 1993) and $\pi_{\hat{\theta}_1}\neq\pi_{\hat{\theta}_2}$. By Theorem 2 (strong duality), it holds that

$$\mathbb{E}_{\tau^{\mathrm{A}}\sim\pi_{\hat{\theta}_i}}\left[\sum_{t=0}^{\infty}\gamma^t\log\pi_{\hat{\theta}_i}(a_t|s_t)\right]=\mathbb{E}_{\tau^{\mathrm{A}}\sim\hat{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t\log\hat{\pi}(a_t|s_t)\right],$$

where $\hat{\pi}$ is an optimal solution to primal problem (23). This is a contradiction to the uniqueness of the optimal solution $\hat{\pi}$ because the maximum entropy objective (23a) is strictly concave. Hence, we can show that the optimal solution to (13) is unique. □

Note that the concavity property does not hold for the estimation objective in Rust (1994). For example, the undiscounted empirical likelihood for group 2 data in Rust (1987) can be shown to be nonconcave.

Moreover, we note that the bilevel formulations (4) and (13) are quite involved, and it is difficult to directly show the concavity of Problems (4) and (13) with nonlinear reward parameterization. Based on our observations under linear reward parameterization, as well as the finite sample guarantee given in Lemma 1, we have the following corollary.

Corollary 2. *Assume that the reward is linearly parameterized, that is, $r(s,a;\theta)=\phi(s,a)^\top\theta$ with $\theta\in\Theta\subset\mathbb{R}^p$ where Θ is a compact set. Assume the ground-truth reward value $r(s,\tilde{a};\theta^*)$ for a reference action $\tilde{a}\in A$ and $s\in S$ are known. Let $\hat{\theta}$ denote the optimal solution to (13). From Algorithm 1's output, define $\hat{\theta}_K:=\theta_{k^*(K)}$, where*

$$k^*(K):=\arg\min_{k\in\{0,K\}}\{\|\nabla\hat{L}(\theta_k,\mathcal{D})\|^2\},$$

then $\hat{\theta}_K\to\hat{\theta}$ in probability with finite-time guarantee $\mathbb{E}[\|\nabla\hat{L}(\hat{\theta}_K,\mathcal{D})\|^2]\le\mathcal{O}(K^{-1/2})$. Furthermore, if $|\mathcal{D}|\ge\frac{2C_r^2}{\epsilon^2(1-\gamma)^2}\ln\left(\frac{2}{\delta}\right)$, then with probability greater than $1-\delta$:

$$L(\theta^*)-L(\hat{\theta})\le\epsilon, \tag{26}$$

where θ^* is the ground truth parameter.

Proof. The finite-time guarantee $\mathbb{E}[\|\nabla\hat{L}(\hat{\theta}_K,\mathcal{D})\|^2]\le\mathcal{O}(K^{-1/2})$ implies $\|\nabla\hat{L}(\hat{\theta}_K,\mathcal{D})\|^2\to0$ in probability. By compactness, the set of accumulation points of the sequence $\{\hat{\theta}_K:K\in\mathbb{N}^+\}$ is nonempty. By Corollary 1(ii), the set of limit points is a singleton, hence $\hat{\theta}_K\to\hat{\theta}$ in probability. To prove the performance guarantee in (26), we can show the following decomposition of the error between the log likelihood objective evaluated at θ^* and $\hat{\theta}$, respectively. With probability greater than $1-\delta$, the following result holds:

$$\begin{aligned}&L(\theta^*)-L(\hat{\theta})\\&=(L(\theta^*)-\hat{L}(\theta^*;\mathcal{D}))+(\hat{L}(\theta^*;\mathcal{D})-\hat{L}(\hat{\theta};\mathcal{D}))+(\hat{L}(\hat{\theta};\mathcal{D})-L(\hat{\theta}))\\&\overset{(i)}{\le}\frac{C_r}{1-\gamma}\sqrt{\frac{\ln(2/\delta)}{2|\mathcal{D}|}}+(\hat{L}(\theta^*)-\hat{L}(\hat{\theta}))+\frac{C_r}{1-\gamma}\sqrt{\frac{\ln(2/\delta)}{2|\mathcal{D}|}}\\&=\frac{2C_r}{1-\gamma}\sqrt{\frac{\ln(2/\delta)}{2|\mathcal{D}|}}+(\hat{L}(\theta^*;\mathcal{D})-\hat{L}(\hat{\theta};\mathcal{D})),\end{aligned} \tag{27}$$

where (i) follows (12) in Lemma 1. Because we defined $\hat{\theta}$ as the optimal solution to $\hat{L}(\cdot;\mathcal{D})$, we know that $\hat{L}(\theta;\mathcal{D})-\hat{L}(\hat{\theta};\mathcal{D})\le0$ for any θ. Plugging this result into (27), the following result holds with probability greater than $1-\delta$:

$$\begin{aligned}&L(\theta^*)-L(\hat{\theta})\\&\le\frac{2C_r}{1-\gamma}\sqrt{\frac{\ln(2/\delta)}{2|\mathcal{D}|}}+(\hat{L}(\theta^*;\mathcal{D})-\hat{L}(\hat{\theta};\mathcal{D}))\\&\le\frac{2C_r}{1-\gamma}\sqrt{\frac{\ln(2/\delta)}{2|\mathcal{D}|}}.\end{aligned} \tag{28}$$

Hence, when the number of expert trajectories in the demonstration data set satisfies $|\mathcal{D}|\ge\frac{2C_r^2}{\epsilon^2(1-\gamma)^2}\ln\left(\frac{2}{\delta}\right)$,

then with probability greater than $1-\delta$, we obtain

$$L(\theta^*) - L(\hat{\theta}) \le \epsilon,$$

where θ^* is the ground truth parameter, which is optimal w.r.t. the log-likelihood objective $L(\cdot)$ defined in (4a). The corollary is proved. □

It is worth mentioning that when relaxing the assumption that the ground-truth reward value $r(s,\tilde{a};\theta^*)$ for a reference action $\tilde{a} \in A$ and $s \in S$ is known, we will no longer have a guarantee on parameter convergence. However, as shown below, the policy obtained by Algorithm 1 still converges to the expert policy.

By defining the state-action visitation measure $d^{\mathrm{E}}(s,a) := (1-\gamma)\pi^{\mathrm{E}}(a|s)\sum_{t=0}^{\infty}\gamma^t P^{\pi^{\mathrm{E}}}(s_t = s|s_0 \sim \rho)$ under the expert policy π^{E}, we can rewrite the expression of the log-likelihood objective $L(\cdot)$ in (4a) for any reward parameter θ as below:

$$\begin{aligned} L(\theta) &:= \mathbb{E}_{\tau^{\mathrm{E}}\sim\pi^{\mathrm{E}}}\left[\sum_{t=0}^{\infty}\gamma^t \log \pi_\theta(a_t|s_t)\right] \\ &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\mathrm{E}}(\cdot),a\sim\pi^{\mathrm{E}}(\cdot|s)}[\log \pi_\theta(a|s)]. \end{aligned}$$

Then the ϵ-optimal solution on the maximum likelihood IRL problem (4) implies

$$L(\theta^*) - L(\hat{\theta}) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\mathrm{E}}(\cdot),a\sim\pi^{\mathrm{E}}(\cdot|s)}\left[\log\left(\frac{\pi_{\theta^*}(a|s)}{\pi_{\hat{\theta}}(a|s)}\right)\right] \le \varepsilon,$$

where $d^{\mathrm{E}}(s,a) := (1-\gamma)\pi^{\mathrm{E}}(a|s)\sum_{t=0}^{\infty}\gamma^t P^{\pi^{\mathrm{E}}}(s_t = s|s_0 \sim \rho)$ denotes the state-action visitation measure under the expert policy π^{E}. Assume the expert behaviors are consistent with optimal behavior for a ground truth reward parameter θ^*, then it follows $\pi^{\mathrm{E}} = \pi_{\theta^*}$. Because of this property, we can obtain the following result:

$$\begin{aligned} &L(\theta^*) - L(\hat{\theta}) \\ &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\mathrm{E}}(\cdot),a\sim\pi^{\mathrm{E}}(\cdot|s)}\left[\log\left(\frac{\pi^{\mathrm{E}}(a|s)}{\pi_{\hat{\theta}}(a|s)}\right)\right] \\ &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\mathrm{E}}(\cdot)}[D_{KL}(\pi^{\mathrm{E}}(\cdot|s)\|\pi_{\hat{\theta}}(\cdot|s))] \\ &\le \varepsilon. \end{aligned}$$

Hence, Corollary 2 provides a formal guarantee that the recovered policy $\pi_{\hat{\theta}}$ solved from the empirical estimation problem (13) is ϵ-close to the expert policy π^{E} measured by the KL divergence.

Remark 2. We also believe the results for the linear reward parameterization case can be generalized to certain nonlinear parametric rewards representations. Such is the case, for example, of *overparameterized* neural networks. In this setting, under certain structural assumptions such as neural tangent kernel and local linearity (Jacot et al. 2018, Du et al. 2019), we expect that the resulting reward representation is approximately linear in the parameters. Hence, it would be possible to identify the global optimal reward estimator. These directions are left for future research.

7. Case with State-Only Dependent Rewards

In this section, we consider the IRL problems when the reward is 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 that are only defined in terms of states is less sensitive to the specific environment dynamics in the expert's demonstration data set. Moreover, in applications such as healthcare (Yu et al. 2021) and autonomous driving (Kiran et al. 2021), simply imitating the expert policy can potentially result in poor performance because the learner and the expert may have different transition dynamics. Similar points have also been argued in recent works (Gangwani and Peng 2020, Ni et al. 2020, Viano et al. 2021).

Next, let us briefly discuss how we can understand (4) and Algorithm 1, when the reward is parameterized as a state-only function. First, it turns out that there is an equivalent formulation of (4a), when the expert trajectories only contain the visited states.

Lemma 4. *Suppose the reward is parameterized as a state-only function $r(s;\theta)$. Then (4) is equivalent to the following*:

$$\min_{\theta} \ \mathbb{E}_{s_0\sim\rho(\cdot)}[V_\theta(s_0)] - \mathbb{E}_{s_0\sim\rho(\cdot)}[V_\theta^{\mathrm{E}}(s_0)] \tag{29a}$$

$$\text{s.t.} \ \ \pi_\theta := \arg\max_{\pi} \ \mathbb{E}_{\tau^{\mathrm{A}}\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t(r(s_t;\theta) + \mathcal{H}(\pi(\cdot|s_t)))\right], \tag{29b}$$

where $V_\theta^{\mathrm{E}}(\cdot)$ denotes the soft value function under reward parameter θ and the expert policy π^{E}.

Please see Section 18 in the supplementary material for detailed derivation. Intuitively, the above lemma says that, when dealing with the state-only IRL, (29a) minimizes the gap between the soft value functions of the optimal policy π_θ and the expert policy π^{E}. Moreover, Algorithm 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 (18). Therefore, even under the state-only IRL setting where the expert data set only contains visited states, our formulation and the proposed algorithm still work if we parameterize the reward as a state-only function.

Moreover, it is straightforward to show that the convergence results in Theorem 1 also hold under the state-only IRL setting.

8. Extension to the Offline Setting

Throughout this paper, we focused on the *online* setting where the transition kernel $P(s_{t+1}|s_t,a_t)$ is known or alternatively, samples from such kernel are available to the learner in an online fashion. However, in many applications, this assumption does not hold, and the available data are fixed. In such an *offline* setting, one strategy to deal with the problem is to estimate *both* the transition kernel and the reward function based on the finite data set of state-action sequences. In our follow-up work (Zeng et al. 2023) to the present paper, we extended Algorithm 1 to the offline setting. In particular, a two-stage estimation procedure has been proposed, where in the first stage a maximum likelihood estimate of the transition kernel is obtained from transition triples (s,a,s') in a transition data set denoted as $\mathcal{D}^T$, that is, $\hat{P} := \arg\max_{\tilde{P}} \mathbb{E}_{(s,a,s')\sim\mathcal{D}^T}[\log \tilde{P}(s'|s,a)]$. Given that finite-data estimation of high-dimensional environment dynamics likely leads to an inaccurate model, in the second stage, a "conservative" reward estimator is obtained using $\hat{P}$ by introducing a regularization term $U(s,a)$ to account model uncertainty:

$$\max_{\theta} \quad \hat{L}(\theta) := \mathbb{E}_{\tau^{\mathrm{E}}\sim\mathcal{D}}\left[\sum_{t\geq 0}\gamma^t \log \pi_\theta(a_t|s_t)\right] \tag{30a}$$

$$\text{s.t.} \quad \pi_\theta := \arg\max_{\pi} \mathbb{E}_{\tau^{\mathrm{A}}\sim(\rho,\pi,\hat{P})}\left[\sum_{t\geq 0}\gamma^t(r(s_t,a_t;\theta)+\mathcal{H}(\pi(\cdot|s_t))-U(s_t,a_t))\right]. \tag{30b}$$

The regularization term in the lower-level problem (30b) induces *conservative* policies that assign low probability to state-action pairs in which $\hat{P}$ cannot provide an accurate prediction on the dynamics. Clearly, the second stage is closely related to the online setting discussed in this work. Therefore, algorithms and intuitions developed in the present work for the online setting is crucial for the offline setting as well.

There are many other outstanding issues to be resolved for the offline setting. For example, how well the estimated transition function can be recovered, how the error will propagate to the error of the reward estimation, and how to compute (stochastic) gradient for the new formulation (30a) and (30b). Because these investigations are out of the scope of this paper, we refer the readers to Zeng et al. (2024) for more details.

9. Testbed

In this section, we test the performance of our algorithm with limited expert trajectories 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.

9.1. Mujoco Tasks for IRL

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 (Todorov et al. 2012). Two classes 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) (Pomerleau 1988) and generative adversarial imitation learning (GAIL) (Ho and Ermon 2016); (2) IRL algorithms that learn a reward function *and* a policy simultaneously, including adversarial inverse reinforcement learning (AIRL) (Fu et al. 2017), *f*-IRL (Ni et al. 2020), and IQ-learn (Garg et al. 2021). To ensure fair comparison, all imitation learning/IRL algorithms use soft AC (Haarnoja et al. 2018) as the base RL algorithm. For the expert data set, 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 and ML-IRL (state-only) that utilizes the state-only reward function. In Table 1, we present the simulation results under a limited data regime where only five expert trajectories are collected. The scores (cumulative rewards) reported in the table is averaged over five random seeds. In each random seed, we train the algorithm from initialization and collect 20 trajectories to average their cumulative rewards after the algorithms converge. According to the results reported in Table 1 where we run the experiments with only five expert trajectories in the demonstration data set $\mathcal{D}$, it shows that our

Table 1. MuJoCo Results

Task	BC	GAIL	IQ-learn	*f*-IRL	ML-IRL (state-only)	ML-IRL (state-action)	Expert
Hopper	102.74	2,762.77	3,039.21	3,116.02	3,131.45	**3,290.02**	3,530.63
Half-cheetah	155.64	3,085.18	4,562.51	4,751.63	4,661.04	**4,846.43**	5,072.53
Walker	283.43	3,610.49	4,361.27	4,562.48	4,367.81	**4,703.35**	5,471.58
Ant	961.58	2,971.57	4,362.90	5,124.13	4,832.38	**5,157.03**	5,856.84
Humanoid	547.62	3,174.66	5,227.10	**5,399.67**	5,149.39	5,281.93	5,339.12

Notes. The performance of benchmark algorithms under five expert trajectories. Bold entries indicate the best performance for a specific task.

Table 2. Transfer Learning

Setting	IQ-learn	AIRL	*f*-IRL	ML-IRL (state-only)	Ground 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

Notes. The performance of benchmark algorithms under a single expert trajectory. The scores in the table are obtained similarly as in Table 1. Bold entries indicate the best performance for a specific task.

proposed algorithms outperform the baselines on most tasks.

We observe that BC fails to imitate the expert's behavior. It is likely because 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, likely due to its inaccurate approximation to the soft Q-function. Therefore, in the Mujoco tasks where IQ-learn does not perform well, we cannot match the results presented in the original paper (Garg et al. 2021). For those cases, we directly report results from the original paper. The results of AIRL are not presented in Table 1 because it performs poorly even after spending significant efforts in parameter tuning; note that similar observations have been made in Liu et al. (2020) and Ni et al. (2020).

9.2. Transfer Learning Across Changing Dynamics

We further evaluate IRL algorithms on the transfer learning setting. We follow the environment setup in Fu et al. (2017), where two environments with different dynamics are considered: Custom-Ant versus Disabled-Ant. We compare ML-IRL (state-only) with several existing IRL methods: (1) AIRL Fu et al. (2017), (2) *f*-IRL Ni et al. (2020), and (3) IQ-learn (Garg et al. 2021).

We consider two transfer learning settings: (1) data transfer and (2) reward transfer. For both settings, the expert data set/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 reported 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 on the environment dynamics. Therefore, the reward function recovered by IQ-Learn cannot be disentangled from the expert actions and environment dynamics, which leads to its failures in the transfer learning tasks.

10. Conclusions

The nested structure of the structural estimation of MDPs entails a significant computational burden in environments with a high-dimensional continuous state or discrete state with large cardinality. To alleviate such burden several approaches have been proposed in both the econometrics (dynamic discrete choice estimation) and artificial intelligence (inverse reinforcement learning) literature. For example, the approximation algorithms in Hotz and Miller (1993) and Hotz et al. (1994) reduce the computational burden, but the resulting estimates suffer from finite sample bias because in high-dimensional state space, initial policy estimates are likely of poor quality. Recent approaches in inverse reinforcement learning that lessen the computational burden (Ni et al. 2020, Garg et al. 2021) do so either at the expense of reward estimation accuracy or lack theoretical guarantees.

In this paper, we introduce a class of single-loop algorithms for the structural estimation of MDPs with nonlinear parametrization. In each iteration a policy improvement step is followed by a stochastic gradient step for likelihood maximization. We show that the proposed algorithm converges to a stationary solution with a finite-time guarantee. Further, if the reward is parameterized linearly, we show that the algorithm approximates the maximum likelihood estimator in sublinear time. Extensive experimentation in standard testbeds for robotics control problems show that the proposed algorithm achieves superior performance compared with other IRL and imitation learning approaches. In future work, we will consider extensions of the proposed algorithm when a model of the state dynamics is not available and thus must also be estimated.

Appendix

A.1. Auxiliary Lemmas

Before starting the proof of the main theorems in this paper, we first introduce several supporting lemmas in this section. Throughout this section, we assume Assumptions 2 and 3 hold true.

Lemma A.1 (Xu et al. 2020, Lemma 3). *Consider the initialization distribution $\rho(\cdot)$ and transition kernel $P(\cdot|s,a)$. Under $\rho(\cdot)$ and $P(\cdot|s,a)$, denote $d_w(\cdot,\cdot)$ as the state-action visitation distribution of MDP with the softmax policy parameterized*

by parameter w. Suppose Assumption 2 holds, for all policy parameter w and w', we have

$$\|d_w(\cdot,\cdot) - d_{w'}(\cdot,\cdot)\|_{TV} \le C_d\|w - w'\| \tag{A.1}$$

where C_d is a positive constant.

Lemma A.2. *Suppose Assumption 3 holds. Under the approximated soft policy iteration in (15), denote the soft Q-function under reward parameter θ_k and policy π_{k+1} as $Q_{k+\frac{1}{2}}$; further note that Q_{k+1} has been defined as the soft Q-function under the reward parameter θ_{k+1} and policy π_{k+1}. Then for any $s \in S$, $a \in \mathcal{A}$ and $k \ge 0$, the following inequality holds:*

$$|Q_{k+\frac{1}{2}}(s,a) - Q_{k+1}(s,a)| \le L_q\|\theta_k - \theta_{k+1}\|, \tag{A.2}$$

where $L_q := \frac{L_r}{1-\gamma}$ and L_r is the positive constant defined in Assumption 3.

Lemma A.3. *Using approximated soft policy iteration (15), the following holds for any iteration $k \ge 0$:*

$$Q_k(s,a) \le Q_{k+\frac{1}{2}}(s,a) + \frac{2\gamma\epsilon_{\text{app}}}{1-\gamma}, \quad \forall s \in S, a \in \mathcal{A}, \tag{A.3}$$

$$\|Q_{\theta_k} - Q_{k+\frac{1}{2}}\|_\infty \le \gamma\|Q_{\theta_k} - Q_k\|_\infty + \frac{2\gamma\epsilon_{\text{app}}}{1-\gamma}, \tag{A.4}$$

where $Q_{k+\frac{1}{2}}(\cdot,\cdot)$ denotes the soft Q-function under reward parameter θ_k and updated policy π_{k+1}, and $\|Q_{\theta_k} - Q_{k+\frac{1}{2}}\|_\infty = \max_{s\in S} \max_{a\in\mathcal{A}} |Q_{\theta_k}(s,a) - Q_{k+\frac{1}{2}}(s,a)|$.

A.2. Proof of Theorem 1

In this section, we prove (21a) and (21b), respectively, to show the convergence of the lower-level problem and the upper-level problem.

A.2.1. Proof of Relation (21a). In this proof, we first show the convergence of the lower-level variable $\{\pi_k\}_{k\ge 0}$. Recall that we approximate the optimal policy π_{θ_k} by π_{k+1} at each iteration k. Moreover, the policy π_{k+1} is generated as below:

$$\pi_{k+1}(a|s) \propto \exp(\hat{Q}_k(s,a)), \text{ where } \|\hat{Q}_k - Q_k\|_\infty \le \epsilon_{\text{app}}. \tag{A.5}$$

We first analyze the approximation error between π_{θ_k} and π_{k+1}. Recall that both policies π_{k+1} and π_{θ_k} are in the softmax form parameterized by $\hat{Q}_k$ and Q_{θ_k}, then it holds

$$\begin{aligned}\|\log\pi_{k+1} - \log\pi_{\theta_k}\|_\infty &\overset{(i)}{\le} 2\|\hat{Q}_k - Q_{\theta_k}\|_\infty \\ &= 2\|\hat{Q}_k - Q_k + Q_k - Q_{\theta_k}\|_\infty \le 2\epsilon_{\text{app}} + 2\|Q_k - Q_{\theta_k}\|_\infty,\end{aligned} \tag{A.6}$$

where (i) follows the Lipschitz property of softmax policy, which is shown in proof of Lemma 3.

Based on Inequality (A.6), we further analyze $\|Q_k - Q_{\theta_k}\|_\infty$ to show the convergence of the policy estimates. Here, we use an auxiliary sequence $\{Q_{k+\frac{1}{2}}\}_{k\ge 0}$, where $Q_{k+\frac{1}{2}}$ is defined as the soft Q-function under reward parameter θ_k and the policy π_{k+1}, its expression follows its

$$Q_{k+\frac{1}{2}}(s,a) := r(s,a;\theta_k) + \mathbb{E}_{\tau^{\text{A}}\sim\pi_{k+1}}\left[\sum_{t=1}^{\infty}\gamma^t(r(s_t,a_t;\theta_k) + \mathcal{H}(\pi_{k+1}(\cdot|s_t)))\middle|(s_0,a_0) = (s,a)\right]. \tag{A.7}$$

Then, the following relations hold:

$$\begin{aligned}\|Q_k - Q_{\theta_k}\|_\infty &= \|Q_k - Q_{\theta_k} + Q_{\theta_{k-1}} - Q_{\theta_{k-1}} + Q_{k-\frac{1}{2}} - Q_{k-\frac{1}{2}}\|_\infty \\ &\le \|Q_{\theta_k} - Q_{\theta_{k-1}}\|_\infty + \|Q_{k-\frac{1}{2}} - Q_{\theta_{k-1}}\|_\infty + \|Q_k - Q_{k-\frac{1}{2}}\|_\infty \\ &\overset{(i)}{\le} L_q\|\theta_k - \theta_{k-1}\| + \|Q_{k-\frac{1}{2}} - Q_{\theta_{k-1}}\|_\infty + \|Q_k - Q_{k-\frac{1}{2}}\|_\infty \\ &\overset{(ii)}{\le} \|Q_{k-\frac{1}{2}} - Q_{\theta_{k-1}}\|_\infty + 2L_q\|\theta_k - \theta_{k-1}\|,\end{aligned} \tag{A.8}$$

where (i) is from (20a) in Lemma 3; (ii) follows Lemma 2. Based on (A.8), we further analyze the two terms in (A.7) as below.

Recall that we have already shown the following relation in (A.4):

$$\|Q_{\theta_k} - Q_{k+\frac{1}{2}}\|_\infty \le \gamma\|Q_{\theta_k} - Q_k\|_\infty + \frac{2\gamma\epsilon_{\text{app}}}{1-\gamma}. \tag{A.9}$$

Through plugging (A.9) into (A.8), we have the following result:

$$\begin{aligned}\|Q_k - Q_{\theta_k}\|_\infty &\le \|Q_{k-\frac{1}{2}} - Q_{\theta_{k-1}}\|_\infty + 2L_q\|\theta_k - \theta_{k-1}\| \\ &\le \gamma\|Q_{\theta_{k-1}} - Q_{k-1}\|_\infty + \frac{2\gamma\epsilon_{\text{app}}}{1-\gamma} + 2L_q\|\theta_k - \theta_{k-1}\|.\end{aligned} \tag{A.10}$$

To show the convergence of the soft Q-function based on (A.10), we further analyze the error between the reward parameters θ_k and θ_{k-1}. Recall that in Algorithm 1, the reward parameter is updated as

$$\theta_k = \theta_{k-1} + \alpha g_{k-1} = \theta_{k-1} + \alpha(h(\theta_{k-1},\tau^{\text{E}}_{k-1}) - h(\theta_{k-1},\tau^{\text{A}}_{k-1})),$$

where we denote $\tau := \{(s_t,a_t)\}_{t=0}^{\infty}$, $h(\theta,\tau) := \sum_{t=0}^{\infty}\gamma^t\nabla_\theta r(s_t,a_t;\theta)$ 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 data set 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 Inequality (19) in Assumption 3, we could show that

$$\|g_{k-1}\| \le \|h(\theta_{k-1},\tau^{\text{E}}_{k-1})\| + \|h(\theta_{k-1},\tau^{\text{A}}_{k-1})\| \le 2L_r\sum_{t=0}^{\infty}\gamma^t = \frac{2L_r}{1-\gamma} = 2L_q, \tag{A.11}$$

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

$$\begin{aligned}\|Q_k - Q_{\theta_k}\|_\infty &\overset{(i)}{\le} \gamma\|Q_{\theta_{k-1}} - Q_{k-1}\|_\infty + \frac{2\gamma\epsilon_{\text{app}}}{1-\gamma} + 2L_q\|\theta_k - \theta_{k-1}\| \\ &\overset{(ii)}{=} \gamma\|Q_{\theta_{k-1}} - Q_{k-1}\|_\infty + \frac{2\gamma\epsilon_{\text{app}}}{1-\gamma} + 2\alpha L_q\|g_{k-1}\| \\ &\overset{(iii)}{\le} \gamma\|Q_{\theta_{k-1}} - Q_{k-1}\|_\infty + \frac{2\gamma\epsilon_{\text{app}}}{1-\gamma} + 4\alpha L_q^2,\end{aligned} \tag{A.12}$$

where (i) is from (A.10); (ii) follows the reward update scheme in (18); and (iii) is from (A.11).

Summing Inequality (A.12) from $k = 1$ to $k = K$, it holds that

$$\sum_{k=1}^{K}\|Q_k - Q_{\theta_k}\|_\infty \le \gamma\sum_{k=0}^{K-1}\|Q_k - Q_{\theta_{k-1}}\|_\infty + K\frac{2\gamma\epsilon_{\text{app}}}{1-\gamma} + 4\alpha KL_q^2. \tag{A.13}$$

Rearranging Inequality (A.13) and dividing (A.13) by K on both sides, it holds that

$$\frac{1-\gamma}{K}\sum_{k=1}^{K}\|Q_k - Q_{\theta_k}\|_\infty \le \frac{\gamma}{K}(\|Q_0 - Q_{\theta_0}\|_\infty - \|Q_K - Q_{\theta_K}\|_\infty) + \frac{2\gamma\epsilon_{\text{app}}}{1-\gamma} + 4\alpha L_q^2. \tag{A.14}$$

Dividing the constant $1-\gamma$ on both sides of (A.14), it holds that

$$\frac{1}{K}\sum_{k=1}^{K}\|Q_k - Q_{\theta_k}\|_\infty \le \frac{\gamma C_0}{K(1-\gamma)} + \frac{2\gamma\epsilon_{\text{app}}}{(1-\gamma)^2} + \frac{4L_q^2}{1-\gamma}\alpha, \tag{A.15}$$

where we denote $C_0 := \|Q_0 - Q_{\theta_0}\|_\infty$. Add $\|Q_0 - Q_{\theta_0}\|_\infty$ and subtract $\|Q_K - Q_{\theta_K}\|_\infty$ on both sides of (A.15), and it follows that

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K-1}\|Q_k - Q_{\theta_k}\|_\infty &\le \frac{\gamma C_0}{K(1-\gamma)} + \frac{C_0}{K} - \frac{\|Q_K - Q_{\theta_K}\|_\infty}{K} + \frac{2\gamma\epsilon_{\text{app}}}{(1-\gamma)^2} + \frac{4L_q^2}{1-\gamma}\alpha \\ &\le \frac{C_0}{K(1-\gamma)} + \frac{2\gamma\epsilon_{app}}{(1-\gamma)^2} + \frac{4L_q^2}{1-\gamma}\alpha.\end{aligned}$$

Recall the step size is defined as $\alpha = \frac{\alpha_0}{K^\sigma}$ where $\sigma > 0$. Then we have

$$\frac{1}{K}\sum_{k=0}^{K-1}\|Q_k - Q_{\theta_k}\|_\infty = \mathcal{O}(K^{-1}) + \mathcal{O}(K^{-\sigma}) + \mathcal{O}(\epsilon_{\text{app}}). \tag{A.16}$$

Summing Inequality (6) from $k = 0$ to $K-1$, it holds that

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K-1}\|\log\pi_{k+1} - \log\pi_{\theta_k}\|_\infty &\le \frac{2}{K}\sum_{k=0}^{K-1}(\epsilon_{\text{app}} + \|Q_k - Q_{\theta_k}\|_\infty) \\ &= \mathcal{O}(K^{-1}) + \mathcal{O}(K^{-\sigma}) + \mathcal{O}(\epsilon_{\text{app}}).\end{aligned}$$

Therefore, we complete the proof of (21a) in Theorem 1. □

A.2.2. Proof of Relation (21b). In this part, we prove the convergence of reward parameters $\{\theta_k\}_{k\ge0}$.

We have the following result of the empirical estimation objective $\hat{L}(\theta;\mathcal{D})$:

$$\begin{aligned}\hat{L}(\theta_{k+1};\mathcal{D}) &\overset{(i)}{\ge} \hat{L}(\theta_k;\mathcal{D}) + \langle\nabla_\theta\hat{L}(\theta_k;\mathcal{D}), \theta_{k+1} - \theta_k\rangle - \frac{L_c}{2}\|\theta_{k+1} - \theta_k\|^2 \\ &\overset{(ii)}{=} \hat{L}(\theta_k;\mathcal{D}) + \alpha\langle\nabla_\theta\hat{L}(\theta_k;\mathcal{D}), g_k\rangle - \frac{L_c\alpha^2}{2}\|g_k\|^2 \\ &= \hat{L}(\theta_k;\mathcal{D}) + \alpha\langle\nabla_\theta\hat{L}(\theta_k;\mathcal{D}), g_k - \nabla_\theta\hat{L}(\theta_k;\mathcal{D})\rangle \\ &\quad + \alpha\|\nabla_\theta\hat{L}(\theta_k;\mathcal{D})\|^2 - \frac{L_c\alpha^2}{2}\|g_k\|^2 \\ &\overset{(iii)}{\ge} \hat{L}(\theta_k;\mathcal{D}) + \alpha\langle\nabla_\theta\hat{L}(\theta_k;\mathcal{D}), g_k - \nabla_\theta\hat{L}(\theta_k;\mathcal{D})\rangle \\ &\quad + \alpha\|\nabla_\theta\hat{L}(\theta_k;\mathcal{D})\|^2 - 2L_cL_q^2\alpha^2,\end{aligned} \tag{A.17}$$

where (i) is from the Lipschitz smooth property in (20b) of Lemma 3; (ii) follows the reward update scheme (18); and (iii) is from constant bound of the gradient estimator g_k in (A.11).

Taking an expectation over both sides of (A.17), it holds that

$$\begin{aligned}&\mathbb{E}[\hat{L}(\theta_{k+1};\mathcal{D})] \\ &\ge \mathbb{E}[\hat{L}(\theta_k;\mathcal{D})] + \alpha\mathbb{E}[\langle\nabla_\theta\hat{L}(\theta_k;\mathcal{D}), g_k - \nabla_\theta\hat{L}(\theta_k;\mathcal{D})\rangle] \\ &\quad + \alpha\mathbb{E}[\|\nabla_\theta\hat{L}(\theta_k;\mathcal{D})\|^2] - 2L_cL_q^2\alpha^2 \\ &= \mathbb{E}[\hat{L}(\theta_k;\mathcal{D})] + \alpha\mathbb{E}[\langle\nabla_\theta\hat{L}(\theta_k;\mathcal{D}), \mathbb{E}[g_k - \nabla_\theta\hat{L}(\theta_k;\mathcal{D})|\theta_k]\rangle] \\ &\quad + \alpha\mathbb{E}[\|\nabla_\theta\hat{L}(\theta_k;\mathcal{D})\|^2] - 2L_cL_q^2\alpha^2 \\ &\overset{(i)}{=} \mathbb{E}[\hat{L}(\theta_k;\mathcal{D})] + \alpha\mathbb{E}\left[\left\langle\nabla_\theta\hat{L}(\theta_k;\mathcal{D}), \mathbb{E}_{\tau^A\sim\pi_{\theta_k}}\left[\sum_{t\ge0}\gamma^t\nabla_\theta r(s_t,a_t;\theta_k)\right]\right.\right. \\ &\quad \left.\left. - \mathbb{E}_{\tau^A\sim\pi_{k+1}}\left[\sum_{t\ge0}\gamma^t\nabla_\theta r(s_t,a_t;\theta_k)\right]\right\rangle\right] \\ &\quad + \alpha\mathbb{E}\left[\|\nabla_\theta\hat{L}(\theta_k;\mathcal{D})\|^2\right] - 2L_cL_q^2\alpha^2 \\ &\overset{(ii)}{\ge} \mathbb{E}[\hat{L}(\theta_k;\mathcal{D})] - 2\alpha L_q \\ &\quad \underbrace{\mathbb{E}\left[\left\|\mathbb{E}_{\tau^A\sim\pi_{\theta_k}}\left[\sum_{t=0}^{\infty}\gamma^t\nabla_\theta r(s_t,a_t;\theta_k)\right] - \mathbb{E}_{\tau^A\sim\pi_{k+1}}\left[\sum_{t=0}^{\infty}\gamma^t\nabla_\theta r(s_t,a_t;\theta_k)\right]\right\|\right]}_{\text{term A}} \\ &\quad + \alpha\mathbb{E}[\|\nabla_\theta\hat{L}(\theta_k;\mathcal{D})\|^2] - 2L_cL_q^2\alpha^2,\end{aligned} \tag{A.18}$$

where (i) follows the expressions of $\nabla_\theta\hat{L}(\theta;\mathcal{D})$ in (16b) and the gradient estimator g_k in (17); and (ii) is due to the fact $\|\nabla_\theta\hat{L}(\theta;\mathcal{D})\| \le 2L_q$ according to (A.11).

Then we further analyze term A as below:

$$\begin{aligned}&\mathbb{E}\left[\left\|\mathbb{E}_{\tau^A\sim\pi_{\theta_k}}\left[\sum_{t=0}^{\infty}\gamma^t\nabla_\theta r(s_t,a_t;\theta_k)\right] - \mathbb{E}_{\tau^A\sim\pi_{k+1}}\left[\sum_{t=0}^{\infty}\gamma^t\nabla_\theta r(s_t,a_t;\theta_k)\right]\right\|\right] \\ &\overset{(i)}{=} \mathbb{E}\left[\left\|\frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d(\cdot,\cdot;\pi_{\theta_k})}[\nabla_\theta r(s,a;\theta_k)]\right.\right. \\ &\quad \left.\left. - \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d(\cdot,\cdot;\pi_{k+1})}[\nabla_\theta r(s,a;\theta_k)]\right\|\right] \\ &= \frac{1}{1-\gamma}\mathbb{E}\left[\left\|\sum_{s\in S,a\in\mathcal{A}}\nabla_\theta r(s_t,a_t;\theta_k)(d(s,a;\pi_{\theta_k}) - d(s,a;\pi_{k+1}))\right\|\right] \\ &\le \frac{1}{1-\gamma}\mathbb{E}\left[\sum_{s\in S,a\in\mathcal{A}}\|\nabla_\theta r(s_t,a_t;\theta_k)\|\cdot|d(s,a;\pi_{\theta_k}) - d(s,a;\pi_{k+1})|\right] \\ &\overset{(ii)}{\le} \frac{2L_r}{1-\gamma}\mathbb{E}[\|d(\cdot,\cdot;\pi_{\theta_k}) - d(\cdot,\cdot;\pi_{k+1})\|_{TV}] \\ &= 2L_q\mathbb{E}[\|d(\cdot,\cdot;\pi_{\theta_k}) - d(\cdot,\cdot;\pi_{k+1})\|_{TV}] \\ &\overset{(iii)}{\le} 2L_qC_d\mathbb{E}[\|Q_{\theta_k} - \hat{Q}_k\|] \\ &\overset{(iv)}{\le} 2L_qC_d\sqrt{|S|\cdot|\mathcal{A}|}\mathbb{E}[\|Q_{\theta_k} - \hat{Q}_k\|_\infty] \\ &\le 2L_qC_d\sqrt{|S|\cdot|\mathcal{A}|}\mathbb{E}[\epsilon_{\text{app}} + \|Q_{\theta_k} - Q_k\|_\infty],\end{aligned} \tag{A.19}$$

where (i) follows the definition of the state-action visitation measure $d(s,a;\pi) = (1-\gamma)\pi(a|s)\sum_{t=0}^{\infty}\gamma^t P^\pi(s_t = s|s_0\sim\rho)$; (ii) follows Inequality (19) in Assumption 3 and the definition

of the total variation norm $\|\cdot\|_{TV}$; (iii) follows the definition of the constant $L_q := \frac{L_r}{1-\gamma}$; and (iv) follows Lemma A.2 and the fact that $\pi_{\theta_k}(\cdot|s) \propto \exp(Q_{\theta_k}(s,\cdot))$, $\pi_{k+1}(\cdot|s) \propto \exp(\hat{Q}_k(s,\cdot))$ follows the conversion between Frobenius norm and infinity norm.

Through plugging Inequality (A.19) into (A.18), this leads to

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

where (i) follows Inequality (A.19).

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

$$\begin{aligned}
\alpha\mathbb{E}[\|\nabla_\theta \hat{L}(\theta_k;\mathcal{D})\|^2] &\leq 2L_c L_q^2\alpha^2 + \alpha C_1 \mathbb{E}[\|Q_{\theta_k} - Q_k\|_\infty + \epsilon_{app}] \\
&\quad + \mathbb{E}[\hat{L}(\theta_{k+1};\mathcal{D}) - \hat{L}(\theta_k;\mathcal{D})].
\end{aligned}$$

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

$$\begin{aligned}
\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla_\theta \hat{L}(\theta_k;\mathcal{D})\|^2] &\leq 2L_c L_q^2\alpha + \frac{C_1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|Q_{\theta_k} - Q_k\|_\infty + \epsilon_{\text{app}}] \\
&\quad + \mathbb{E}\left[\frac{\hat{L}(\theta_K;\mathcal{D}) - \hat{L}(\theta_0;\mathcal{D})}{K\alpha}\right].
\end{aligned}$$

According to Assumption 1, we assume that the reward function is bounded. Based on this assumption, we know that the empirical estimation objective $\hat{L}(\cdot;\mathcal{D})$ is bounded. Then we could plug (A.16) into the inequality above, and we obtain

$$\frac{1}{K}\sum_{K=0}^{K-1}\mathbb{E}[\|\nabla_\theta \hat{L}(\theta_K;\mathcal{D})\|^2] = \mathcal{O}(K^{-\sigma}) + \mathcal{O}(K^{-1}) + \mathcal{O}(K^{-1+\sigma}) + \mathcal{O}(\epsilon_{\text{app}}). \tag{A.20}$$

This completes the proof of this result. □

A.2. Proof of Theorem 2

In this section, we prove the duality between the estimation problem (13) and the maximum entropy IRL problem (22). To state the proof, we first write down the *partial* Lagrangian function, when only dualizing the constraint (22b) and (22c). After we derive the dual form for the problem with Constraint (22b) and (22c), we will make sure that Constraint (22d) is satisfied.

Let θ and C_{s_t} denote the dual variables of Constraints (22b) and (22c), respectively; define $\phi(\pi^E;\mathcal{D}) := \mathbb{E}_{\tau^E\sim\mathcal{D}}[\sum_{t=0}^{\infty}\gamma^t\phi(s_t,a_t)]$. Then the partial Lagrangian can be expressed as

$$\begin{aligned}
\mathcal{L}(\pi,\theta) &:= -\mathbb{E}_{\tau^A\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t \log\pi(a_t|s_t)\right] \\
&\quad + \theta^\top\left(\mathbb{E}_{\tau^A\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t\phi(s_t,a_t)\right] - \phi(\pi^E;\mathcal{D})\right) \\
&\quad + \sum_{t\geq 0, s_t\in S} C_{s_t}\left(\sum_{a\in\mathcal{A}}\pi(a|s_t) - 1\right).
\end{aligned} \tag{A.21}$$

Our plan is to show that the dual function, as defined by $\bar{\mathcal{L}}(\theta) := \max_\pi \mathcal{L}(\pi,\theta)$, has the following expression:

$$\bar{\mathcal{L}}(\theta) = \mathbb{E}_{s_0\sim\rho}[V_\theta(s_0)] - \mathbb{E}_{\tau^E\sim\mathcal{D}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t;\theta)\right], \tag{A.22}$$

so that the dual problem can be shown to be equivalent to Problem (22), as follows:

$$\begin{aligned}
\min_\theta \bar{\mathcal{L}}(\theta) &= \min_\theta \mathbb{E}_{s_0\sim\rho}[V_\theta(s_0)] - \mathbb{E}_{\tau^E\sim\mathcal{D}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t;\theta)\right] \\
&= \max_\theta \mathbb{E}_{\tau^E\sim\mathcal{D}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t;\theta)\right] - \mathbb{E}_{s_0\sim\rho}[V_\theta(s_0)].
\end{aligned}$$

Toward this end, let us compute the gradient of $\mathcal{L}(\pi,\theta)$ with respect to the policy $\pi(a|s_t = s)$:

$$\begin{aligned}
&\nabla_{\pi(a|s_t=s)}\mathcal{L}(\pi,\theta) \\
&= \nabla_{\pi(a|s_t=s)}\left(-\mathbb{E}_{\tau^A\sim\pi}\left[\sum_{\kappa=0}^{\infty}\gamma^\kappa\log\pi(a_\kappa|s_\kappa)\right]\right. \\
&\qquad \left. + \theta^\top\mathbb{E}_{\tau^A\sim\pi}\left[\sum_{\kappa=0}^{\infty}\gamma^\kappa\phi(s_\kappa,a_\kappa)\right]\right) \\
&\quad + \nabla_{\pi(a|s_t=s)}\left(-\theta^\top\phi(\pi^E;\mathcal{D}) + \sum_{\kappa\geq 0, s\in S} C_{s_\kappa=s}\left(\sum_{a\in\mathcal{A}}\pi(a|s_\kappa) - 1\right)\right) \\
&\overset{(i)}{=} \nabla_{\pi(a|s_t=s)}\left(-\mathbb{E}_{\tau^A\sim\pi}\left[\sum_{\kappa=t}^{\infty}\gamma^\kappa\log\pi(a_\kappa|s_\kappa)\right]\right. \\
&\qquad \left. + \theta^\top\mathbb{E}_{\tau^A\sim\pi}\left[\sum_{\kappa=t}^{\infty}\gamma^t\phi(s_t,a_t)\right]\right) + C_{s_t=s} \\
&= \nabla_{\pi(a|s_t=s)}\left(-\sum_{s\in S,a\in\mathcal{A}} P^\pi(s_t = s|s_0\sim\rho)\pi(a|s_t = s)\right. \\
&\qquad \left.\mathbb{E}_{\tau^A\sim\pi}\left[\sum_{\kappa=t}^{\infty}\gamma^\kappa\log\pi(a_\kappa|s_\kappa)\middle|(s_t,a_t) = (s,a)\right]\right) \\
&\quad + \nabla_{\pi(a|s_t=s)}\left(\sum_{s\in S,a\in\mathcal{A}} P^\pi(s_t = s|s_0\sim\rho)\pi(a|s_t = s)\right. \\
&\qquad \left.\mathbb{E}_{\tau^A\sim\pi}\left[\sum_{\kappa=t}^{\infty}\gamma^\kappa\theta^\top\phi(s_\kappa,a_\kappa)\middle|(s_t,a_t) = (s,a)\right]\right) + C_{s_t=s} \\
&= P^\pi(s_t = s|s_0\sim\rho)\left(-\gamma^t(\log\pi(a|s_t = s) + 1)\right. \\
&\quad + \mathbb{E}_{\tau^A\sim\pi}\left[\sum_{\kappa=t}^{\infty}-\gamma^{\kappa+1}\log\pi(a_{\kappa+1}|s_{\kappa+1})\middle|(s_t,a_t) = (s,a)\right] \\
&\quad \left. + \mathbb{E}_{\tau^A\sim\pi}\left[\sum_{\kappa=t}^{\infty}\gamma^\kappa\theta^\top\phi(s_\kappa,a_\kappa)\middle|(s_t,a_t) = (s,a)\right]\right) + C_{s_t=s},
\end{aligned} \tag{A.23}$$

where (i) follows the fact that the probability $\pi(a|s_t=s)$ has no effect on the trajectory generated before time t. Setting $\nabla_{\pi(a|s_t=s)}\mathcal{L}(\pi,\theta)=0$, we obtain the following first-order condition:

$$\begin{aligned}\log\pi(a|s_t=s)=&\left(\frac{C_{s_t=s}}{\gamma^t\cdot P^\pi(s_t=s|s_0\sim\rho)}-1\right)\\&-\mathbb{E}_{\tau^A\sim\pi}\left[\sum_{\kappa=t}^{\infty}\gamma^{\kappa+1-t}\log\pi(a_{\kappa+1}|s_{\kappa+1})\middle|(s_t,a_t)=(s,a)\right]\\&+\mathbb{E}_{\tau^A\sim\pi}\left[\sum_{\kappa=t}^{\infty}\gamma^{\kappa-t}\theta^\top\phi(s_\kappa,a_\kappa)\middle|(s_t,a_t)=(s,a)\right].\end{aligned}$$

Then, we can express $\pi(a|s_t=s)$ as below:

$$\begin{aligned}&\pi(a|s_t=s)\\&=\exp\left(-\mathbb{E}_{\tau^A\sim\pi}\left[\sum_{\kappa=t}^{\infty}\gamma^{\kappa+1-t}\log\pi(a_{\kappa+1}|s_{\kappa+1})\middle|(s_t,a_t)=(s,a)\right]\right.\\&\quad+\mathbb{E}_{\tau^A\sim\pi}\left[\sum_{\kappa=t}^{\infty}\gamma^{\kappa-t}\theta^\top\phi(s_\kappa,a_\kappa)\middle|(s_t,a_t)=(s,a)\right]\\&\quad\left.+\frac{C_{s_t=s}}{\gamma^t\cdot P^\pi(s_t=s)}-1\right).\end{aligned}\tag{A.24}$$

Note that $\left(\frac{C_{s_t=s}}{\gamma^t\cdot P^\pi(s_t=s|s_0\sim\rho)}-1\right)$ is independent of the action a_t. Hence, the following result holds:

$$\begin{aligned}&\pi(a|s_t=s)\\&\propto\exp\left(\mathbb{E}_{\tau^A\sim\pi}\left[\sum_{\kappa=t}^{\infty}\gamma^{\kappa-t}(\theta^T\phi(s_\kappa,a_\kappa)-\gamma\log\pi(a_{\kappa+1}|s_{\kappa+1}))\middle|(s_t,a_t)=(s,a)\right]\right)\\&=\exp\left(\mathbb{E}_{\tau^A\sim\pi}\left[\sum_{\kappa=0}^{\infty}\gamma^{\kappa}(\theta^\top\phi(s_\kappa,a_\kappa)-\gamma\log\pi(a_{\kappa+1}|s_{\kappa+1}))\middle|(s_0,a_0)=(s,a)\right]\right).\end{aligned}\tag{A.25}$$

According to (A.25), we could conclude that $\pi(a|s_t=s)$ only depends on the state-action pair (s,a) and is independent of the time index $t\geq 0$. Hence, we have shown that the policy π is a stationary policy and $\pi(a|s_t=s)=\pi(a|s)$ for any $t\geq 0$.

Therefore, we can rewrite (A.25) with $t=0$ as follows:

$$\begin{aligned}&\pi(a|s)\\&\propto\exp\left(\mathbb{E}_{\tau^A\sim\pi}\left[\sum_{\kappa=0}^{\infty}\gamma^{\kappa}(\theta^T\phi(s_\kappa,a_\kappa)-\gamma\log\pi(a_{\kappa+1}|s_{\kappa+1}))\middle|(s_0,a_0)=(s,a)\right]\right)\\&\overset{(i)}{=}\exp\left(r(s_0,a_0;\theta)\right.\\&\quad\left.+\mathbb{E}_{\tau^A\sim\pi}\left[\sum_{\kappa=0}^{\infty}\gamma^{\kappa+1}(r(s_{\kappa+1},a_{\kappa+1};\theta)-\log\pi(a_{\kappa+1}|s_{\kappa+1}))\middle|(s_0,a_0)=(s,a)\right]\right)\\&\overset{(ii)}{=}\exp(Q^\pi(s,a)),\end{aligned}\tag{A.26}$$

where (i) follows the linear approximation of the reward function that $r(s,a;\theta):=\theta^T\phi(s,a)$. Clearly, the right-hand side of (i) is the soft Q-function under reward parameter θ and the stationary policy π; therefore in (ii), we use $Q^\pi(s,a)$ to denote such a soft Q-function.

Recall that we have defined V_θ, Q_θ as the soft value function, soft Q-function under reward parameter θ, and the optimal policy π_θ. For any $s\in S$ and $a\in\mathcal{A}$, it follows that

$$V_\theta(s):=\mathbb{E}_{\tau^A\sim\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t(r(s_t,a_t;\theta)+\mathcal{H}(\pi_\theta(\cdot|s_t)))|s_0=s\right],\tag{A.27a}$$

$$Q_\theta(s,a):=r(s,a;\theta)+\gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}[V_\theta(s')].\tag{A.27b}$$

According to Haarnoja et al. (2017) and Cen et al. (2022), the optimal policy π_θ in the entropy-regularized MDP satisfies the following expression for any $s\in S$ and $a\in\mathcal{A}$:

$$\pi_\theta(a|s)=\frac{\exp(Q_\theta(s,a))}{\sum_{\tilde{a}\in\mathcal{A}}\exp(Q_\theta(s,\tilde{a}))}.\tag{A.28}$$

Therefore, we know the policy in (A.26) is the optimal policy π_θ. Using π_θ to replace the policy π in the Lagrangian function $L(\pi,\theta)$ as given by (A.21), we can express the dual function as

$$\begin{aligned}&\bar{\mathcal{L}}(\theta)\\&=-\mathbb{E}_{\tau^A\sim\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t\log\pi_\theta(a_t|s_t)\right]\\&\quad+\theta^\top\left(\mathbb{E}_{\tau^A\sim\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t\phi(s_t,a_t)\right]-\mathbb{E}_{\tau^E\sim\mathcal{D}}\left[\sum_{t=0}^{\infty}\gamma^t\phi(s_t,a_t)\right]\right)\\&\quad+\sum_{t\geq0,s_t\in S}C_{s_t}\left(\sum_{a\in\mathcal{A}}\pi_\theta(a|s_t)-1\right)\\&\overset{(i)}{=}-\mathbb{E}_{\tau^A\sim\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t\log\left(\frac{\exp Q_\theta(s_t,a_t)}{\sum_{a\in\mathcal{A}}\exp Q_\theta(s_t,a)}\right)\right]\\&\quad+\mathbb{E}_{\tau^A\sim\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t;\theta)\right]-\mathbb{E}_{\tau^E\sim\mathcal{D}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t;\theta)\right]\\&=-\mathbb{E}_{\tau^A\sim\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t\left(Q_\theta(s_t,a_t)-\log\left(\sum_{a\in\mathcal{A}}\exp Q_\theta(s_t,a)\right)\right)\right]\\&\quad+\mathbb{E}_{\tau^A\sim\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t;\theta)\right]-\mathbb{E}_{\tau^E\sim\mathcal{D}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t;\theta)\right]\\&\overset{(ii)}{=}-\mathbb{E}_{\tau^A\sim\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t(r(s_t,a_t;\theta)+\gamma V_\theta(s_{t+1})-V_\theta(s_t))\right]\\&\quad+\mathbb{E}_{\tau^A\sim\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t;\theta)\right]-\mathbb{E}_{\tau^E\sim\mathcal{D}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t;\theta)\right]\\&=-\mathbb{E}_{\tau^A\sim\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t(\gamma V_\theta(s_{t+1})-V_\theta(s_t))\right]-\mathbb{E}_{\tau^E\sim\mathcal{D}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t;\theta)\right]\\&=\mathbb{E}_{s_0\sim\rho}[V_\theta(s_0)]-\mathbb{E}_{\tau^E\sim\mathcal{D}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t;\theta)\right],\end{aligned}\tag{A.29}$$

where (i) follows the fact that $\pi_\theta(a_t|s_t)=\frac{\exp Q_\theta(s_t,a_t)}{\sum_{a\in\mathcal{A}}\exp Q_\theta(s_t,a)}$ (see (A.28)) and $r(s,a;\theta):=\theta^T\phi(s,a)$, and (ii) follows (A.27b) and (A.27a). Then we can show the equivalence between (A.28)

and (13a):

$$\min_\theta \bar{\mathcal{L}}(\theta) = \min_\theta \mathbb{E}_{s_0\sim\rho}[V_\theta(s_0)] - \mathbb{E}_{\tau^E\sim\mathcal{D}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t;\theta)\right]$$
$$= \max_\theta \mathbb{E}_{\tau^E\sim\mathcal{D}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t;\theta)\right] - \mathbb{E}_{s_0\sim\rho}[V_\theta(s_0)].$$

Hence, we proved that (13a) and (13b) is the dual form of (22a)–(22c) and Constraint (22d) is satisfied due to the closed form of the optimal policy π_θ in (A.28).

Note Objective (22a) is concave, and (22b) and (22c) are affine. In addition, the interior of the feasible region is not empty (i.e., Slater's condition). Hence, under linear parameterization of the reward function, there is strong duality (no gap) between the solutions of (13) and (22).

When the expert policy is known or available for access, following the derivations in (A.29), we show the dual problem of the maximum entropy estimation problem ((22a), (23), (22c), (22d)) as follows:

$$\max_\theta \quad \mathbb{E}_{\tau^E\sim\pi^E}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t;\theta)\right] - \mathbb{E}_{s_0\sim\rho}[V_\theta(s_0)]$$
$$\text{s.t} \quad \pi_\theta(a_t|s_t) := \arg\max_\pi \mathbb{E}_{s_0\sim\rho,\tau^A\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t(r(s_t,a_t;\theta)+\mathcal{H}(\pi(\cdot|s_t)))\right]. \tag{A.30}$$

Then based on our derivations in (8), we obtain the equivalence between (A.30) and (4a):

$$\mathbb{E}_{\tau^E\sim\pi^E}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t;\theta)\right] - \mathbb{E}_{s_0\sim\rho}[V_\theta(s_0)]$$
$$= \mathbb{E}_{\tau^E\sim\pi^E}\left[\sum_{t=0}^{\infty}\gamma^t \ln\pi_\theta(a_t|s_t)\right].$$

Therefore, we obtain the duality between the maximum likelihood estimation problem (4) and the maximum entropy estimation problem ((22a), (23), (22c), (22d)). □

Endnotes

[1] In Section 6, we show that if the reward is linearly parametrized, the maximum entropy formulation in Ziebart et al. (2008) is the dual of the maximum likelihood formulation of the estimation problem.

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

References

Adusumilli K, Eckardt D (2019) Temporal-difference estimation of dynamic discrete choice models. Preprint, submitted December 19, https://arxiv.org/abs/1912.09509.

Aguirregabiria V, Mira P (2002) Swapping the nested fixed point algorithm: A class of estimators for discrete Markov decision models. *Econometrica* 70(4):1519–1543.

Bajari P, Benkard CL, Levin J (2007) Estimating dynamic models of imperfect competition. *Econometrica* 75(5):1331–1370.

Bhandari J, Russo D, Singal R (2018) A finite time analysis of temporal difference learning with linear function approximation. *Proc. Conf. Learn. Theory* (PMLR, New York), 1691–1692.

Borkar VS (1997) Stochastic approximation with two time scales. *Systems Control Lett.* 29(5):291–294.

Cayci S, He N, Srikant R (2021) Linear convergence of entropy-regularized natural policy gradient with linear function approximation. Preprint, submitted June 8, https://arxiv.org/abs/2106.04096.

Cen S, Cheng C, Chen Y, Wei Y, Chi Y (2022) Fast global convergence of natural policy gradient methods with entropy regularization. *Oper. Res.* 70(4):2563–2578.

Chen T, Sun Y, Yin W (2021) Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems. *Adv. Neural Inform. Processing Systems* 34:25294–25307.

Chernozhukov V, Escanciano JC, Ichimura H, Newey WK, Robins JM (2022) Locally robust semiparametric estimation. *Econometrica* 90(4):1501–1535.

Du SS, Zhai X, Poczos B, Singh A (2019) Gradient descent provably optimizes over-parameterized neural networks. *Proc. Internat. Conf. Learn. Representations* (OpenReview.net).

Fu J, Luo K, Levine S (2017) Learning robust rewards with adversarial inverse reinforcement learning. Preprint, submitted October 30, https://arxiv.org/abs/1710.11248.

Gangwani T, Peng J (2020) State-only imitation with transition dynamics mismatch. *Proc. Internat. Conf. Learn. Representations* (OpenReview.net).

Garg D, Chakraborty S, Cundy C, Song J, Ermon S (2021) Iq-learn: Inverse soft-q learning for imitation. *Adv. Neural Inform. Processing Systems* 34:4028–4039.

Guan Z, Xu T, Liang Y (2021) When will generative adversarial imitation learning algorithms attain global convergence. *Proc. Internat. Conf. Artificial Intelligence Statist.* (PMLR, New York), 1117–1125.

Haarnoja T, Tang H, Abbeel P, Levine S (2017) Reinforcement learning with deep energy-based policies. *Proc. Internat. Conf. Machine Learn.* (PMLR, New York), 1352–1361.

Haarnoja T, Zhou A, Abbeel P, Levine S (2018) Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *Proc. Internat. Conf. Machine Learn.* (PMLR, New York), 1861–1870.

Hansen LP, Miao J (2018) Aversion to ambiguity and model misspecification in dynamic stochastic environments. *Proc. Natl. Acad. Sci. USA* 115(37):9163–9168.

Ho J, Ermon S (2016) Generative adversarial imitation learning. *Proc. 30th Internat. Conf. Neural Inform. Processing Systems* (Curran Associates Inc., Red Hook, NY), 4572–4580.

Hong M, Wai HT, Wang Z, Yang Z (2020) A two-timescale framework for bilevel optimization: Complexity analysis and application to actor-critic. Preprint, submitted July 10, https://arxiv.org/abs/2007.05170.

Hotz VJ, Miller RA (1993) Conditional choice probabilities and the estimation of dynamic models. *Rev. Econom. Stud.* 60(3):497–529.

Hotz VJ, Miller RA, Sanders S, Smith J (1994) A simulation estimator for dynamic models of discrete choice. *Rev. Econom. Stud.* 61:265–289.

Jacot A, Gabriel F, Hongler C (2018) Neural tangent kernel: Convergence and generalization in neural networks. *Proc. 32nd Internat. Conf. Neural Inform. Processing Systems* (Curran Associates Inc., Red Hook, NY), 8580–8589.

Jin C, Netrapalli P, Jordan M (2020) What is local optimality in nonconvex-nonconcave minimax optimization? *Proc. Internat. Conf. Machine Learn.* (PMLR, New York), 4880–4889.

Khanduri P, Zeng S, Hong M, Wai HT, Wang Z, Yang Z (2021) A near-optimal algorithm for stochastic bilevel optimization via double-momentum. *Adv. Neural Inform. Processing Systems* 34:30271–30283.

Kiran BR, Sobh I, Talpaert V, Mannion P, Al Sallab AA, Yogamani S, Pérez P (2021) Deep reinforcement learning for autonomous driving: A survey. *IEEE Trans. Intelligent Transportation Systems* 23(6):4909–4926.

Konda V, Tsitsiklis J (1999) Actor-critic algorithms. Solla S, Leen T, Müller K, eds. *Advances in Neural Information Processing Systems*, vol. 12 (MIT Press, Cambridge, MA).

Liu F, Ling Z, Mu T, Su H (2020) State alignment-based imitation learning. *Proc. Internat. Conf. Learn. Representations* (OpenReview.net).

Mai T, Jaillet P (2020) A relation analysis of Markov decision process frameworks. Preprint, submitted August 18, https://arxiv.org/abs/2008.07820.

Matějka F, McKay A (2015) Rational inattention to discrete choices: A new foundation for the multinomial logit model. *Amer. Econom. Rev.* 105(1):272–298.

Ni T, Sikchi H, Wang Y, Gupta T, Lee L, Eysenbach B (2020) f-irl: Inverse reinforcement learning via state marginal matching. Preprint, submitted November 9, https://arxiv.org/abs/2011.04709.

Ortega PA, Braun DA (2013) Thermodynamics as a theory of decision-making with information-processing costs. *Proc. A* 469(2153):20120683.

Pomerleau DA (1988) ALVINN: An autonomous land vehicle in a neural network. *Proc. 1st Internat. Conf. Neural Inform. Processing Systems* (MIT Press, Cambridge, MA), 305–313.

Reich G (2018) Divide and conquer: Recursive likelihood function integration for hidden Markov models with continuous latent variables. *Oper. Res.* 66(6):1457–1470.

Rust J (1987) Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher. *Econometrica* 55(5):999–1033.

Rust J (1994) Structural estimation of Markov decision processes. *Handbook of Econometrics*, vol. 4 (Elsevier, Amsterdam), 3081–3143.

Sanghvi N, Usami S, Sharma M, Groeger J, Kitani K (2021) Inverse reinforcement learning with explicit policy estimates. *Proc. Conf. AAAI Artificial Intelligence* 35:9472–9480.

Su CL, Judd KL (2012) Constrained optimization approaches to estimation of structural models. *Econometrica* 80(5):2213–2230.

Tishby N, Polani D (2011) Information theory of decisions and actions. *Perception-Action Cycle* (Springer, Berlin), 601–636.

Todorov E, Erez T, Tassa Y (2012) Mujoco: A physics engine for model-based control. *Proc. IEEE/RSJ Internat. Conf. Intelligent Robots Systems* (IEEE, Piscataway, NJ), 5026–5033.

Viano L, Huang YT, Kamalaruban P, Weller A, Cevher V (2021) Robust inverse reinforcement learning under transition dynamics mismatch. *Advances in Neural Information Processing Systems*, vol. 34 (Curran Associates Inc., Red Hook, NY), 25917–25931.

Wu YF, Zhang W, Xu P, Gu Q (2020) A finite-time analysis of two time-scale actor-critic methods. *Adv. Neural Inform. Processing Systems* 33:17617–17628.

Wulfmeier M, Ondruska P, Posner I (2015) Maximum entropy deep inverse reinforcement learning. Preprint, submitted July 17, https://arxiv.org/abs/1507.04888.

Xu T, Zhe W, Yingbin L (2020) Improving sample complexity bounds for (natural) actor-critic algorithms. *Adv. Neural Inform. Processing Sys.* 33:4358–4369.

Yu C, Liu J, Nemati S, Yin G (2021) Reinforcement learning in healthcare: A survey. *ACM Comput. Survey* 55(1):1–36.

Zeng S, Li C, Garcia A, Hong M (2023) When demonstrations meet generative world models: A maximum likelihood framework for offline inverse reinforcement learning. *Proc. 37th Internat. Conf. Neural Inform. Processing Systems* (Curran Associates Inc., Red Hook, NY), 65531–65565.

Ziebart BD, Bagnell JA, Dey AK (2010) Modeling interaction via the principle of maximum causal entropy. *Proc. Internat. Conf. Machine Learn.* (Omnipress, Madison, WI), 1255–1262.

Ziebart BD, Bagnell JA, Dey AK (2013) The principle of maximum causal entropy for estimating interacting processes. *IEEE Trans. Inform. Theory* 59(4):1966–1980.

Ziebart BD, Maas AL, Bagnell JA, Dey AK, et al. (2008) Maximum entropy inverse reinforcement learning. *Proc. Conf. AAAI Artificial Intelligence* 8:1433–1438.

Zou S, Xu T, Liang Y (2019) Finite-sample analysis for sarsa with linear function approximation. *Adv. Neural Inform. Processing Systems*, vol. 32 (Curran Associates Inc., Red Hook, NY).

Siliang Zeng is a PhD student in the department of electrical and computer engineering at the University of Minnesota, Twin Cities. His research interests focus on machine learning, reinforcement learning, and sequential decision making under uncertainty.

Mingyi Hong is an associate professor with the department of electrical and computer engineering, University of Minnesota. His research interests include optimization theory and applications in signal processing and machine learning. His work received two IEEE Signal Processing Society (SPS) Best Paper awards in 2021 to 2022, an International Consortium of Chinese Mathematicians Best Paper Award in 2020, he received Pierre-Simon Laplace Early Career Technical Achievement Award from IEEE SPS.

Alfredo Garcia is Michael and Sugar Barnes professor in the department of industrial and system engineering, Texas A&M University. His research interests include game theory and dynamic optimization with applications in communications and energy networks.