Understanding Inverse Reinforcement Learning under Overparameterization: Non-Asymptotic Analysis and Global Optimality

Ruijia Zhang
Johns Hopkins University
rzhan127@jh.edu

Siliang Zeng
University of Minnesota
zeng0176@umn.edu

Chenliang Li
Texas A&M University
chenliangli@tamu.edu

Alfredo Garcia
Texas A&M University
alfredo.garcia@tamu.edu

Mingyi Hong
University of Minnesota
mhong@umn.edu

Abstract

The goal of the Inverse reinforcement learning (IRL) task is to identify the underlying reward function and the corresponding optimal policy from a set of expert demonstrations. While most IRL algorithms' theoretical guarantees rely on a linear reward structure, we aim to extend the theoretical understanding of IRL to scenarios where the reward function is parameterized by neural networks. Meanwhile, conventional IRL algorithms usually adopt a nested structure, leading to computational inefficiency, especially in high-dimensional settings. To address this problem, we propose the first two-timescale single-loop IRL algorithm under neural network parameterized reward and provide a non-asymptotic convergence analysis under overparameterization. Although prior optimality results for linear rewards do not apply, we show that our algorithm can identify the globally optimal reward and policy under certain neural network structures. This is the first IRL algorithm with a non-asymptotic convergence guarantee that provably achieves global optimality in neural network settings.

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

1 Introduction

Given the observed trajectories of states and actions executed by an expert, we consider the challenge of inferring the reinforcement learning reward function in which the expert was trained. This problem is commonly known as inverse reinforcement learning (IRL), as discussed in a recent survey by Osa et al. (2018). IRL involves the task of estimating both the reward function and the expert's policy that best aligns with the provided data. While there are certain limitations on the identifiability of rewards (Kim et al., 2021), the process of estimating rewards from expert trajectories offers valuable insights for estimating optimal policies under different environment dynamics and/or learning new tasks through reinforcement learning, which has been widely used in different applications such as autonomous driving (Kurach et al., 2018; Neu and Szepesvari, 2012; Phan-Minh et al., 2022), autonomous trading (Yang et al., 2018), and LLM alignment (Ouyang et al., 2022).

In the Maximum Entropy Inverse Reinforcement Learning (MaxEnt-IRL) formulation introduced by Ziebart et al. (2008), the expert's behavior is modeled as a policy that maximizes entropy while satisfying a constraint to ensure that the expected features align with the empirical averages in the expert's dataset. The algorithms developed for MaxEnt-IRL (Ziebart et al., 2008, 2010; Wulfmeier et al., 2015) usually employ a nested loop structure. These algorithms cycle through two main stages: an outer loop focused on the reward update, and an inner loop responsible for calculating precise policy estimates. Although these nested loop algorithms remain computationally feasible within tabular contexts, they pose significant computational challenges in high-dimensional settings where function approximation is necessary (Finn et al., 2016).

Most recently, a new formulation of IRL based on maximum likelihood (ML) estimation has been proposed (Zeng et al., 2022). The ML formulation is a bi-level optimization problem. In this framework, the upper-level problem maximizes the log-likelihood function given observed trajectories, while the lower-level problem seeks to identify the optimal policy given the current reward parameterization. To alleviate the computational burden associated with the recurrent optimization of the lower-level policy, ML-IRL carries out the policy improvement step and reward update alternately. The new policy at each round is generated by *soft policy iteration* (Haarnoja et al., 2017). This approach ensures that each iteration can be executed with relatively lower computational costs.

However, whether ML-IRL converges or not relies on its policy improvement step, which has not been thoroughly discussed before. An $\epsilon-$accurate estimation of soft Q-function for policy improvement is assumed to ensure global convergence. It's worth noting that this assumption remains intractable in practice. The single-loop implementation of ML-IRL (Zeng et al., 2022) carries out multiple soft Actor-Critic (SAC) steps for the policy update instead. It therefore lacks a convergence guarantee to substantiate its performance.

Moreover, existing ML-IRL works only consider the optimality analysis when the reward employs a linear structure. Zeng et al. (2023) establishes a strong duality relationship with maximum entropy Inverse Reinforcement Learning (IRL) when the reward can be expressed as a linear combination of features.

We want to advance IRL research by developing a theoretical understanding of IRL under neural network parameterization.

1.1 Challenges:

The main challenge is how to ensure the global convergence of our IRL algorithm when we don't have an accurate policy evaluation in practice. Besides, the nonconvex structure of neural network parametrization, coupled with bi-level optimization formulation, complicates the optimality analysis.

Therefore, this paper seeks to address these challenges and answer the following questions:
(i) Can we design an IRL framework that ensures convergence without requiring precise policy evaluation, even when the reward function is parameterized by neural networks?
(ii) Is it possible to develop a computationally efficient single-loop algorithm within this framework that avoids inner policy optimization, while still maintaining convergence?
(iii) Moreover, does this algorithm provably identify the ground truth reward function and the corresponding optimal policy?

In this work, we provide affirmative answers to the above questions. We summarize our major contributions here.

1.2 Contributions:

- We design the first theoretically guaranteed ML-IRL framework under neural network settings. This framework comprises two main stages: (i) In the policy improvement stage, we apply a finite-step neural soft Q-learning (Haarnoja et al., 2017; Cai et al., 2023) to update the policy. (ii) In the reward update stage, we first estimate the gradient of the maximum log-likelihood objective function with respect to the reward parameter through the sampled trajectory induced by the present policy. Subsequently, gradient ascent is employed to update the reward parameter based on this gradient estimate.
- To reduce the computational complexity of nested soft Q-learning in the policy improvement stage, we propose a single-loop ML-IRL algorithm for solving the bi-level IRL problem. Inspired by the two-timescale alternating update process of actor-critic (Hong et al., 2020; Borkar, 1997), our approach carefully differentiates the stepsizes for Temporal Difference (TD) updates and reward parameter updates, which ensures global convergence with just one iteration of soft Q-learning.
- Additionally, we show that the algorithm has strong theoretical guarantees: it requires $\mathcal{O}\left(\epsilon^{-8}\right)$ steps to identify an ϵ-approximate globally optimal reward estimator under overparameterization. To our knowledge, this is the first single-loop algorithm with finite-time convergence analysis and global optimality guarantees for IRL problems parameterized by neural networks.
- Finally, we conduct intensive numerical experiments on Mujoco tasks. We show that our methods have better performances than many previous state-of-the-art IRL algorithms for imitating the expert behavior.

2 Related Work

2.1 Single-loop IRL Algorithms

Towards more efficient training, some recent works have developed algorithms to alleviate the computational burden of nested-loop IRL structures. In Reddy et al. (2019), the authors attempt to model the IRL using certain maximum entropy RL problem with a very trivial reward function that assigns $r = +1$ for matching expert demonstrations and $r = 0$ for all other behaviors. Garg et al. (2021) proposes to transform the standard formulation of IRL into a single-level problem by directly estimating the soft Q-function, which encodes the information of the reward function and policy. However, the implicit reward function recovered by the soft Q-function is not very accurate since it is not strictly subject to satisfy Bellman's equation. Another approach called f-IRL (Ni et al., 2021) estimates rewards based on the minimization of several measures of divergence with respect to the expert's state visitation measure. However, it is restricted to certain scenarios where the reward function only depends on the state. Besides, no convergence guarantee is given to support its performance on the reported numerical results of the single-loop implementation.

2.2 Convergence Analysis of IRL algorithms

Syed and Schapire (2007); Syed et al. (2008); Abbeel and Ng (2004); Neu and Szepesvari (2012); Ross and Bagnell (2010); Rajaraman et al. (2020) first study the convergence of IRL but only in the tabular case. The convergence result of Behavior Cloning (BC) established by Rashidinejad et al. (2023) fails in continuous state and action spaces with a horizon $H \geq 2$, since BC is considered as a classification problem and always faces unseen states in the continuous state space. Liu et al. (2021) and Zhang et al. (2020) present a convergence guarantee for GAIL in the linear function and neural network approximation setting respectively. However, the problem formulation of GAIL is a min-max optimization, where the inner part maximizes the reward and the outer part minimizes the policy. Liu et al. (2021) and Zhang et al. (2020) only show that the learned policy cannot be distinguished from the expert policy in expectation, without mentioning the recovery of the expert reward function.

2.3 Neural Network Analysis

Our theoretical understanding of optimality guarantees is inspired by the literature on the local linearization property of the neural network parameterization (Maei et al., 2009; Bertsekas, 2018; Geist and Pietquin, 2013). Previous works on the optimality analysis of neural networks such as Li and Liang (2018); Du and Lee (2018) show that for a one-hidden-layer network with ReLU activation function using overparameterization and random initialization, GD and SGD can find the near global-optimal solutions in polynomial time. Zou et al. (2018); Allen-Zhu et al. (2020a); Gao et al. (2019) extends the result to L-hidden-layer-fully-connected neural network structure. Several recent works Cai et al. (2023); Gaur et al. (2023); Fu et al. (2021); Zhang et al. (2020) discusses optimality results in deep reinforcement learning. However, previous results cannot be easily generalized to the bi-level IRL optimization formulation, which is significantly more challenging.

3 Preliminaries

In RL, a discounted Markov decision process (MDP) is denoted by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \mu_0, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P(\cdot \mid s, a)$ is the transition probability kernel, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, $\mu_0(\cdot)$ is the initial state distribution and γ is the discount factor. A policy π takes state $s \in \mathcal{S}$ as an input and gives a distribution over actions $\mathcal{A}$. μ_π is the stationary state-action distribution associated with the policy π. We consider $\mathcal{S}$ to be continuous and $\mathcal{A}$ to be finite.

3.1 Entropy-Regularized Reinforcement Learning

Maximum entropy RL (Haarnoja et al., 2017) searches the optimal policy, which maximizes its entropy at each visited state:

$$\pi^*_{\text{MaxEnt}} = \arg\max_\pi \sum_{t=0}^{\infty} \mathbb{E}_{(\mathbf{s}_t, \mathbf{a}_t) \sim \mu_\pi} \left[\gamma^t \left(r\left(\mathbf{s}_t, \mathbf{a}_t\right) + \mathcal{H}\left(\pi\left(\cdot \mid \mathbf{s}_t\right)\right)\right)\right], \tag{1}$$

where $\mathcal{H}(\pi(\cdot \mid s)) := -\sum_{a \in \mathcal{A}} \pi(a \mid s) \log \pi(a \mid s)$ denotes the entropy of policy $\pi(\cdot \mid s)$. In this context, the soft V-function and soft Q-function are defined as follows:

$$V^{\text{soft}}_{r,\pi}(s) = \mathbb{E}_{\pi, s_0 = s}\left[\sum_{t=0}^{\infty} \gamma^t \left(r(s_t, a_t) + \mathcal{H}\left(\pi(\cdot \mid s_t)\right)\right)\right], \tag{2}$$

$$Q^{\text{soft}}_{r,\pi}(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim \mathcal{P}(\cdot \mid s, a)}\left[V^{\text{soft}}_{r,\pi}(s')\right]. \tag{3}$$

By soft policy iteration (Haarnoja et al., 2017), a greedy entropy-regularized policy at $(k+1)$th iteration π_{k+1} can be generated by π_k at kth step as follows: $\pi_{k+1}(\cdot \mid s) \propto \exp\left(Q^{soft}_{r,\pi_k}(s, \cdot)\right), \forall s \in \mathcal{S}$ (Ziebart et al., 2010; Haarnoja et al., 2017). Under a fixed reward function, it can be shown that the new policy π_{k+1} monotonically improves π_k, and it converges linearly to the optimal policy of (1); see Theorem 4 in Haarnoja et al. (2017) and Theorem 1 in Cen et al. (2022).

3.2 Problem Formulation: Maximum Log-Likelihood IRL (ML-IRL)

In this section, we review the fundamentals of maximum likelihood inverse reinforcement learning (ML-IRL).

Assume observations are in the form of expert state-action trajectories $\tau^E = \{(s_t, a_t)\}_{t\geq 0}$ drawn from an expert policy. A model of the expert's behavior under parameterized reward $\widehat{r}(s,a;\theta)$ is a randomized policy $\pi_\theta(\cdot \mid s)$ where θ is a parameter vector. Assuming the transition dynamics $\mathcal{P}(s_{t+1} \mid s_t, a_t)$ are known, the discounted log-likelihood of observing a sample trajectory τ^E under model π_θ can be written as follows:

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

Let $\mathcal{L}(\theta) := \mathbb{E}_{\tau^E\sim\pi^E}\left[\sum_{t>0}\gamma^t\log\pi_\theta(a_t \mid s_t)\right]$. The IRL problem can then be formulated as the following maximum log-likelihood IRL formulation (Zeng et al., 2022):

$$
\begin{aligned}
&\max_\theta\ \mathcal{L}(\theta) \\
&\text{s.t. } \pi_\theta = \arg\max_\pi \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t\left(\widehat{r}(s_t, a_t;\theta) + \mathcal{H}(\pi(\cdot \mid s_t))\right)\right],
\end{aligned}
\tag{4}
$$

where $\mathcal{H}(\pi(\cdot \mid s)) := -\sum_{a\in\mathcal{A}}\pi(a \mid s)\log\pi(a \mid s)$ denotes the entropy of policy $\pi(\cdot \mid s)$.

However, in practice, the ground truth policy π^E is unknown. Let $\mathcal{D} := \left\{\tau_i^E\right\}_{i=1}^N$ denote the demonstration dataset containing only finite observed trajectories. The empirical discounted log-likelihood $\widehat{\mathcal{L}}(\theta)$ is denoted by $\mathbb{E}_{\tau^E\sim\mathcal{D}}\left[\sum_{t>0}\gamma^t\log\pi_\theta(a_t \mid s_t)\right]$. Therefore, we consider an empirical version of (4) as follows.

$$
\begin{aligned}
&\max_\theta\ \widehat{\mathcal{L}}(\theta) \\
&\text{s.t. } \pi_\theta = \arg\max_\pi \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t\left(\widehat{r}(s_t, a_t;\theta) + \mathcal{H}(\pi(\cdot \mid s_t))\right)\right].
\end{aligned}
\tag{5}
$$

Note that the problem takes the form of a bi-level optimization problem, where the upper-level problem (ML-IRL) optimizes the reward parameter θ, while the lower-level problem describes the expert's policy as the solution to an entropy-regularized MDP (Haarnoja et al., 2017, 2018). The entropy regularization in (5) ensures the uniqueness of the optimal policy π_θ given the fixed reward function $\widehat{r}(s,a;\theta)$ (Haarnoja et al., 2017, 2018), even when the underlying MDP is high-dimensional and complex.

4 Proposed Algorithm

4.1 Neural Network parameterization of IRL

Definition 1 *The state-action pair $(s,a) \in \mathcal{S}\times\mathcal{A}$ is represented by a vector $x = \psi(s,a) \in \mathcal{X} \subseteq \mathbb{R}^d$ with $d > 2$, where ψ is a given one-to-one feature map.*

With a slight abuse of notation, we use (s,a) and x interchangeably. Without loss of generality, we assume that $\|x\|_2 = 1$ and $|\widehat{r}(x;\theta)|$ is upper bounded by a constant $\bar{r} > 0$ for any $x \in \mathcal{X}$. In this paper, we want to extend our understanding of IRL by replacing the commonly assumed linear reward function with a neural network parameterized one.

To parametrize the reward function and the soft Q value function, we use a two-layer neural network.

$$
\widehat{Q}(x;W) = \frac{1}{m}\sum_{j=1}^{m} b^j\sigma\left({W^j}^\top x\right), \tag{6a}
$$

$$
\widehat{r}(x;\theta) = \frac{1}{m}\sum_{j=1}^{m} b^j\sigma\left({\theta^j}^\top x\right). \tag{6b}
$$

Here σ is the rectified linear unit (ReLU) activation function. $\sigma(y) = \max\{0, y\}$. $b = \left(b^1, \ldots, b^m\right)$ is generated as $b^j \sim \mathrm{Unif}(\{-1, 1\})$, the initialization parameter W_0 is generated as $W_0^j \sim N(0, I_d/d)$, and the initialization parameter θ_0 is initialized as $\theta_0^j \sim N(0, I_d/d)$ for any $j \in [m]$ independently. During training, we only update $W = \left(W^1, \ldots, W^m\right) \in \mathbb{R}^{md}$ and $\theta = \left(\theta^1, \ldots, \theta^m\right) \in \mathbb{R}^{md}$, while keeping $b = \left(b^1, \ldots, b^m\right) \in \mathbb{R}^m$ fixed as the random initialization. W is restricted within a closed ball with radius B: $S_B = \left\{W \in \mathbb{R}^{md} : \|W - W_0\|_2 \leq B\right\}$ $(B > 0)$.This is a classical neural network structure for theoretical analysis, due to its local linearization property (Wang et al., 2019). It has been demonstrated to be capable of learning a class of infinite-order smooth functions (Arora et al., 2019; Allen-Zhu et al., 2020b).

To solve problem (5), Zeng et al. (2022) proposes an algorithm named Maximum Likelihood Inverse Reinforcement Learning (ML-IRL) based on certain stochastic algorithms for bi-level optimization (Hong et al., 2020; Ji et al., 2021; Khanduri et al., 2021), thereby avoiding the heavy computational burden from the traditional nested loop algorithms. Specifically, *soft policy iteration* (Haarnoja et al., 2017) is adopted to optimize the lower-level problem. Therefore, the soft Q-function value under the current step's policy and reward function is critical for policy improvement in ML-IRL.

However, a key limitation of ML-IRL is that the exact soft Q-function value is typically not accessible. In order to make the algorithm more practical, we decide to incorporate finite-step soft Q-learning (Haarnoja et al., 2017) to directly approximate the optimal soft Q value $\widehat{Q}^{soft}_{\widehat{r}_\theta,\pi_\theta}(s,a)$ under the current reward $\widehat{r}_\theta$. The detailed implementation is shown in Algorithm 1.

4.2 ML-IRL with Dynamically Truncated Soft Q-learning Nested

Typically, it requires $\mathcal{O}\left(\epsilon^{-2}\right)$ loops of Algorithm 1 to get an $\epsilon-$accurate approximation of the optimal soft Q-function

Algorithm 1 Neural Soft Q-learning

Require:
Exploration policy π_{exp} such that $\pi_{\text{exp}}(a|s) > 0$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, total iterations T, stepsize η, current reward $\widehat{r}_\theta$, $b^j \sim \text{Unif}(\{-1,1\})$, $W_0^j \sim N\left(0, I_d/d\right)(s \in [m])$, $\bar{W} = W_0$, $S_B = \left\{W \in \mathbb{R}^{md} : \|W - W_0\|_2 \leq B\right\} (B > 0)$,

for $t = 0, 1, \ldots, T-1$ **do**
- Sample a tuple $(s, a, \widehat{r}_\theta, s')$ from the stationary distribution μ_{exp} of the exploration policy π_{exp}
- Bellman residual calculation: $\delta \leftarrow \widehat{Q}(s,a;W_t) - \widehat{r}_\theta - \gamma \operatorname{softmax}_{a' \in \mathcal{A}} \widehat{Q}\left(s', a'; W_t\right)$;
- TD update: $\widetilde{W}_{t+1} \leftarrow W_t - \eta\delta \cdot \nabla_W \widehat{Q}(s,a;W_t)$;
- Projection: $W_{t+1} \leftarrow \operatorname{argmin}_{W \in S_B} \|W - \widetilde{W}_{t+1}\|_2$
- Averaging: $\bar{W} \leftarrow \frac{t+1}{t+2} \cdot \bar{W} + \frac{1}{t+2} \cdot W_{t+1}$

end for

Output: $\widehat{Q}_{\text{out}}(\cdot) = \widehat{Q}^{soft}_{\widehat{r}_\theta, \pi_\theta}(\cdot) \leftarrow \widehat{Q}(\cdot; \bar{W})$

for each policy improvement (Cai et al., 2023). However, it is not possible to do so in real practice. To address this, we present an improved and feasible Algorithm 2. The innovation in Algorithm 2 lies in the dynamic adjustment of the number of stochastic steps in the nested Algorithm 1. Specifically, at the $k+1$th round of policy improvement, Algorithm 1 is required to perform $k+2$ iterations to approximate the optimal soft Q value under the reward $\widehat{r}_{\theta_k}$ at kth iteration.

This design utilizes soft Q-learning to get a direct but imprecise estimation of optimal soft Q-function value in finite iterations. The previous convergence guarantee in ML-IRL (Zeng et al., 2022) may no longer hold since Algorithm 2 doesn't achieve an $\epsilon-$accurate policy estimation required for soft policy iteration. However, by forcing the number of nested iterations in Algorithm 1 to increase linearly with outer reward update rounds, we can still ensure global convergence. The convergence guarantee will be discussed in later Section 5. We now present the Algorithm 2 below.

4.3 Two-timescale Single Loop ML-IRL

As is shown above, Algorithm 2 still retains a nested inner loop structure due to the approximation of the optimal soft Q-function. Despite requiring only a finite number of iterations for each policy improvement round, it still results in an overall complexity of $\mathcal{O}\left(K^2\right)$. To reduce the computational complexity, we design a two-timescale single-loop Algorithm 3. While still relying on the approximate optimal soft Q-function (line 3 to line 8), Algorithm 3 achieves a significant improvement by just one iteration of stochastic update from Algorithm 1, under the current step's reward.

To ensure global convergence, we employ the two-timescale stochastic approximation method (Hong et al., 2020; Borkar, 1997) By carefully selecting two distinct stepsizes, the lower-level policy optimization converges more quickly than the upper-level reward updates. This approach enables the policy π_{k+1} to stay aligned with the optimal π_{θ_k}. In the policy improvement step of Algorithm 3 (line 9), the lower-level policy is iteratively updated by the *soft policy iteration*

Algorithm 2 Maximum Likelihood Inverse Reinforcement Learning (ML-IRL) with Dynamically Truncated Soft Q-learning nested

Require:
Reward parameter θ_0, stepsize of reward parameter α;

for $k = 0, 1, \ldots, K-1$ **do**
- Optimal Soft Q-function Approximation: Run Algorithm 1 under reward function $\widehat{r}\left(\cdot; \theta_k\right)$ for $T = k+2$ rounds
- Policy Improvement: $\pi_{k+1}(\cdot \mid s) \propto \exp\left(\widehat{Q}^{soft}_{\widehat{r}_{\theta_k}, \pi_{\theta_k}}(s, \cdot)\right), \forall s \in \mathcal{S}$.
- Data Sampling I: Sampling expert trajectory $\tau_k^E := \left\{s_t, a_t\right\}_{t \geq 0}$ from the dataset $\mathcal{D}$
- Data Sampling II: Sampling agent trajectory $\tau_k^A := \left\{s_t, a_t\right\}_{t \geq 0}$ from the policy π_{k+1}
- Estimating Gradient: $g_k := h\left(\theta_k; \tau_k^E\right) - h\left(\theta_k; \tau_k^A\right)$ where $h(\theta; \tau) := \sum_{t \geq 0} \gamma^t \nabla_\theta \widehat{r}\left(s_t, a_t; \theta\right)$
- Reward Parameter Update: $\theta_{k+1} := \theta_k + \alpha g_k$

end for

return π_K, θ_K

(Haarnoja et al., 2017), which converges linearly to the optimal policy under a fixed reward function (Cen et al., 2022). As a result, the lower-level updates occur on a faster timescale. In contrast, the upper-level reward parameter updates (line 13) progress more slowly because it does not exhibit such a nice linear convergence property. A more detailed discussion on choosing the appropriate stepsizes can be found in Supplementary D.2. We give the full implementation details of Algorithm 3 as follows.

5 Theoretical Analysis

In this section, we mainly provide theoretical insights into our proposed algorithms.

5.1 Definition:

Recall that the soft Q function has the following overparameterization structure:

$$\widehat{Q}(x; W) = \frac{1}{m} \sum_{j=1}^m b^j \mathbb{1}\left\{{W^j}^\top x > 0\right\} {W^j}^\top x,$$

and $\nabla_{W^j} \widehat{Q}(x; W) = \frac{b^j}{m} \mathbb{1}\left\{{W^j}^\top x > 0\right\} x$ almost everywhere in $\mathbb{R}^d$.

Definition 2 *(Local Linearization of soft Q function) The neural network $\widehat{Q}(x;W)$ can be linearized locally at random initialization point W_0 by $\widehat{Q}_0(x;W)$ with respect to W:*

$$\widehat{Q}_0(x; W) = \Phi(x)^\top W, \tag{7}$$

where we define a feature map $\Phi(x) \in \mathbb{R}^{md}$ as follows: $\Phi(x) := \frac{1}{m} \cdot \left(\mathbb{1}\left\{{W_0^1}^\top x > 0\right\} x, \ldots, \mathbb{1}\left\{{W_0^m}^\top x > 0\right\} x\right)$.

Algorithm 3 Two-Timescale Single Loop Maximum Likelihood Inverse Reinforcement Learning (ML-IRL)

Require:
Reward parameter θ_0, explorative policy π_0, stepsize of reward parameter update α, stepsize of TD update η.

initialize $b^j \sim \text{Unif}(\{-1, 1\})$,
$W_0^j \sim N(0, I_d/d)\,(s \in [m]), \bar{W} = W_0$,
$S_B = \{W \in \mathbb{R}^{md} : \|W - W_0\|_2 \leq B\}\,(B > 0)$;
for $k = 0, 1, \ldots, K-1$ **do**
  Sample a tuple (s, a, r, s') from the stationary distribution μ_k of the current policy π_k
  Bellman residual calculation: $\delta \leftarrow \widehat{Q}(s, a; W_k) - \widehat{r}(s, a; \theta_k) - \gamma\, \text{softmax}_{a' \in \mathcal{A}}\, \widehat{Q}(s', a'; W_k)$;
  TD update: $\widetilde{W}_{k+1} \leftarrow W_k - \eta\delta \cdot \nabla_W \widehat{Q}(s, a; W_k)$;
  Projection: $W_{k+1} \leftarrow \text{argmin}_{W \in S_B} \|W - \widetilde{W}_{k+1}\|_2$
  Averaging: $\bar{W} \leftarrow \frac{k+1}{k+2} \cdot \bar{W} + \frac{1}{k+2} \cdot W_{k+1}$
  Output: $\widehat{Q}^{soft}_{\widehat{r}_{\theta_k}, \pi_{\theta_k}}(\cdot) = \widehat{Q}_{\text{out}}(\cdot) \leftarrow \widehat{Q}(\cdot; \bar{W})$
  Policy Improvement:
  $\pi_{k+1}(\cdot \mid s) \propto \exp\left(\widehat{Q}^{soft}_{\widehat{r}_{\theta_k}, \pi_{\theta_k}}(s, \cdot)\right), \forall s \in \mathcal{S}$.
  Data Sampling I: Sampling expert trajectory $\tau_k^E := \{s_t, a_t\}_{t \geq 0}$ from the dataset D
  Data Sampling II: Sampling agent trajectory $\tau_k^A := \{s_t, a_t\}_{t \geq 0}$ from the policy π_{k+1}
  Estimating Gradient:
  $g_k := h(\theta_k; \tau_k^E) - h(\theta_k; \tau_k^A)$, where $h(\theta; \tau) := \sum_{t \geq 0} \gamma^t \nabla_\theta \widehat{r}(s_t, a_t; \theta)$
  Reward Parameter Update: $\theta_{k+1} := \theta_k + \alpha g_k$
end for
return π_K, θ_K

$\widehat{Q}_0(x; W)$ is linear in the feature map $\Phi(x)$.

Remark: Local linearization of the neural network is a classic function class of nonlinear function approximation and is known to have strong representation power when the width m goes to infinity (Hofmann et al., 2008). It is commonly used in the literature of stationary point convergence analysis (Zou et al., 2018; Cai et al., 2023; Allen-Zhu et al., 2020b; Daniely, 2017).

5.2 Assumptions

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

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

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

Assumption 1 assumes the Markov chain mixes at a geometric rate. It is a common assumption in the literature of RL (Wu et al., 2022; Bhandari et al., 2018; Zou et al., 2019), which holds for any time-homogeneous Markov chain with finite-state space or any uniformly ergodic Markov chain with general state space.

Assumption 2 *(Regularity of Policy) There exists a constant $\nu' > 0$ such that for any $W_1, W_2 \in S_B$, it holds that*

$$(\gamma + \nu')^{-2}\, \mathbb{E}_{x \sim \mu}\left[\left(\widehat{Q}_0(x; W_1) - \widehat{Q}_0(x; W_2)\right)^2\right]$$
$$\geq \mathbb{E}_{s \sim \mu}\left[\left(\underset{a \in \mathcal{A}}{\text{softmax}} \widehat{Q}_0(s, a; W_1) - \underset{a \in \mathcal{A}}{\text{softmax}} \widehat{Q}_0(s, a; W_2)\right)^2\right].$$

Cai et al. (2023) proposes this strong regularity assumption to show that the soft Bellman operator is a contraction mapping and therefore guarantees the uniqueness of our approximate stationary point W_θ^*. Zou et al. (2019); Melo et al. (2008)'s discussions on the global convergence of linear Q-learning rely on a more strong assumption that implies Assumption 2.

Assumption 3 *(Regularity of Stationary Distribution) There exists a constant $c_1 > 0$ such that for any $\zeta \geq 0$ and $w \in \mathbb{R}^d$ with $\|w\|_2 = 1$, it holds that*

$$\mathbb{P}\left(\left|w^\top \psi(s, a)\right| \leq \zeta, \text{ for all } a \in \mathcal{A}\right) \leq c_1 \cdot \zeta,$$

where $(s, a) \sim \mu$.

Assumption 3 imposes a constraint on the density of μ with respect to the marginal distribution of x. This constraint is naturally satisfied when the marginal distribution of x has a uniformly bounded probability density across the unit sphere, as discussed in Cai et al. (2023).

5.3 Theorem

Theorem 1 *Suppose Assumptions 1, 2 and 3 hold. Selecting stepsize $\alpha := \frac{\alpha_0}{K^\sigma}$ for the reward update step and $\eta = \min\{K^{-\frac{1}{2}}, (1-\gamma)/8\}$ for the TD update step in Algorithm 1 where $\alpha_0 > 0$ and $\sigma \in (0, 1)$ are some fixed constants, and K is the total number of iterations of the Algorithm 2. Then the following result holds:*

$$\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}\left[\|\log \pi_{k+1} - \log \pi_{\theta_k}\|_\infty\right]$$
$$= \mathcal{O}\left(K^{-\frac{1}{4}}\right) + \mathcal{O}\left(B^{3/2} m^{-1/2} + B^{5/4} m^{-1/4}\right), \tag{8a}$$
$$\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}\left[\left\|\nabla \widehat{\mathcal{L}}(\theta_k)\right\|^2\right]$$
$$= \mathcal{O}\left(K^{-\sigma}\right) + \mathcal{O}\left(K^{-1+\sigma}\right) + \mathcal{O}\left(K^{-\frac{1}{4}}\right)$$
$$+ \mathcal{O}\left(m^{-1/2} + B^{3/2} m^{-1/2} + B^{5/4} m^{-1/4}\right), \tag{8b}$$

*where the expectation is over all randomness and $\|\log \pi_{k+1} - \log \pi_{\theta_k}\|_\infty$
$= \max_{s \in \mathcal{S}, a \in \mathcal{A}} |\log \pi_{k+1}(a \mid s) - \log \pi_{\theta_k}(a \mid s)|$.*

In Theorem 1, we present the non-asymptotic convergence guarantee for the Algorithm 2. In particular, by setting $\sigma = -\frac{1}{4}$ and the width m to infinity, it takes both reward and policy $\mathcal{O}(\epsilon^{-4})$ iterations to converge to an $\epsilon-$stationary point. The detailed proof of Theorem 1 is in Supplementary D.1.

Theorem 2 *Suppose Assumptions 1, 2 and 3 hold. Selecting stepsize $\alpha := \frac{\alpha_0}{K^\sigma}$ for the reward update step and $\eta = \min\{K^{-\frac{3}{4}}, (1-\gamma)/8\}$ for the TD update step where $\alpha_0 > 0$ and $\sigma \in (0,1)$ are some fixed constants, and K is the total number of iterations to be run by the Algorithm 3. Then the following result holds:*

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\|\log \pi_{k+1} - \log \pi_{\theta_k}\|_\infty\right] = \mathcal{O}\left(K^{-\frac{1}{8}}\right) + \mathcal{O}\left(B^{3/2}m^{-1/2} + B^{5/4}m^{-1/4}\right), \tag{9a}$$

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\widehat{\mathcal{L}}(\theta_k)\right\|^2\right] = \mathcal{O}\left(K^{-\sigma}\right) + \mathcal{O}\left(K^{-1+\sigma}\right) + \mathcal{O}\left(K^{-\frac{1}{8}}\right) + \mathcal{O}\left(m^{-1/2} + B^{3/2}m^{-1/2} + B^{5/4}m^{-1/4}\right), \tag{9b}$$

where the expectation is over all randomness and $\|\log \pi_{k+1} - \log \pi_{\theta_k}\|_\infty = \max_{s\in\mathcal{S}, a\in\mathcal{A}} |\log \pi_{k+1}(a \mid s) - \log \pi_{\theta_k}(a \mid s)|$.

Theorem 2 shows that our two-timescale single-loop Algorithm 3 can identify an approximate stationary point of our problem (5) up to a neural network error, which depends on the width of our neural network. Specifically, when $\sigma = -\frac{1}{8}$ and the neural network is sufficiently wide, both the reward and policy converge with the rate of $\mathcal{O}(K^{-\frac{1}{8}})$ iterations to an $\epsilon-$stationary point.

Remark: Our theoretical guarantee stands apart from the existing results in several key ways. Cen et al. (2022) only shows the convergence rate of soft policy iteration under a fixed reward function. Zeng et al. (2022) first considers the setting where both the policy and the reward parameter keep changing, but the convergence is only guaranteed when the reward function is linear. In contrast, we analyze a significantly more challenging scenario where the reward is parameterized by neural networks, providing a broader and more generalized convergence guarantee. We present our proof in Supplementary D.2.

Discussions: Allen-Zhu et al. (2019); Gao et al. (2019); Zou et al. (2018) go beyond the double-layer neural networks and establish high probability convergence guarantees for deep overparameterized neural networks. Cai et al. (2023) then apply these results to extend the analysis of neural soft Q-learning. Therefore, our Theorem 1 and 2 can be generalized to the setting where both the reward function and soft Q-function are parameterized by deep neural networks with slight modifications to network errors. We just refer to these findings here to support our experimental setup in a later section, since it is not the main focus of this work.

Next, we demonstrate that our Algorithm 3 can converge to the globally optimal solution of problem (5) under overparameterization. To study the global optimality of our algorithm, we reformulate the bi-level problem as the following equivalent saddle point problem (Zeng et al., 2024).

$$\max_\theta \min_\pi \quad \widehat{L}(\theta,\pi) := \mathbb{E}_{\tau^E\sim\mathcal{D}}\left[\sum_{t=0}^{\infty}\gamma^t\left(\widehat{r}\left(s_t,a_t;\theta\right)\right)\right] - \mathbb{E}_{\tau^{\mathrm{A}}\sim(\mu_0,\pi)}\left[\sum_{t=0}^{\infty}\gamma^t\left(\widehat{r}\left(s_t,a_t;\theta\right) + \mathcal{H}\left(\pi\left(\cdot\mid s_t\right)\right)\right)\right]. \tag{10}$$

Since π_θ is defined as the optimal policy under reward parameter θ in the bi-level formulation (5), the following relation holds:

$$\widehat{\mathcal{L}}(\theta) = \min_\pi \widehat{L}(\theta,\pi). \tag{11}$$

Theorem 3 *(Approximate Concavity) Given any fixed policy π and for any $\theta, \theta' \in \mathbb{R}^{md}$ we have:*

$$\left[\widehat{L}\left(\theta',\pi\right) - \widehat{L}\left(\theta,\pi\right)\right] \leq \left[\nabla_\theta \widehat{L}\left(\theta,\pi\right)^\top\left(\theta'-\theta\right)\right] + \mathcal{O}\left(m^{-1/2}\right).$$

Therefore, we note that the max-min objective $\widehat{L}(\theta,\pi)$ is concave in the reward parameter θ as the width m of our neural network tends to infinity. Besides, we establish the following relationship:$\nabla\widehat{\mathcal{L}}(\theta) = \nabla_\theta\widehat{L}(\theta,\pi_\theta)$ under overparameterization. We can utilize this relation and the concavity in Theorem 3 to show that any stationary point $\tilde{\theta}$ of (5) together with its corresponding optimal policy $\pi_{\tilde{\theta}}$ consists of a saddle point of (10) by checking the first-order condition. From (11), we derive the following result for a saddle point $\left(\tilde{\theta}, \pi_{\tilde{\theta}}\right)$:

$$\tilde{\theta} \in \arg\max_\theta \min_{\pi_\theta} \widehat{L}(\theta,\pi) = \arg\max_\theta \widehat{\mathcal{L}}(\theta).$$

Therefore, for any saddle point $\left(\tilde{\theta}, \pi_{\tilde{\theta}}\right)$ of the objective $\widehat{L}(\cdot,\cdot)$, the reward parameter $\tilde{\theta}$ is a global optimal solution of problem (5). The detailed proof can be found in Supplementary F.

Corollary 1 *Assume that the ground truth reward value $r\left(s,a;\theta^*\right)$ for an action $a \in A$ and $s \in S$ is known and that the set of parameters Θ is a compact set. Let $\widetilde{\theta}$ be the optimal solution to (5). If the number of expert trajectories in the demonstration data set satisfies $|\mathcal{D}| \geq \frac{2}{\epsilon^2 m^2(1-\gamma)^2}\ln\left(\frac{2}{\delta}\right)$, then with probability greater than $1-\delta$, we obtain the performance guarantee of the optimality gap in (4) that:*

$$\mathcal{L}\left(\theta^*\right) - \mathcal{L}(\tilde{\theta}) \leq \epsilon.$$

Theorem 1 and 2 have already shown that our learned reward converges to the staionary point of the empirical estimation (5), i.e. $\theta_K \to \tilde{\theta}$. Since the parameter set Θ is compact, therefore the set of limit points of the sequence $\left\{\theta_K : K \in \mathbb{N}^+\right\}$ is non-empty. The uniqueness of the global optimal point θ^* is established by the approximate concavity in Theorem 3 with the auxiliary Assumption 2. Therefore, the set of limit points is a singleton. To prove whether the global optimal reward from the empirical estimation problem (5) is close to the underlying true reward, we investigate the optimality gap $\mathcal{L}\left(\theta^*\right) - \mathcal{L}(\tilde{\theta})$. We give the proof sketch as follows.

Proof Sketch: With probability greater than $1-\delta$, the following relation holds:

$$\begin{aligned}
&\mathcal{L}\left(\theta^*\right) - \mathcal{L}(\tilde{\theta}) \\
&= \left(\mathcal{L}\left(\theta^*\right) - \widehat{\mathcal{L}}\left(\theta^*;\mathcal{D}\right)\right) + \left(\widehat{\mathcal{L}}\left(\theta^*;\mathcal{D}\right) - \widehat{\mathcal{L}}(\tilde{\theta};\mathcal{D})\right) \\
&\quad + (\widehat{\mathcal{L}}(\tilde{\theta};\mathcal{D}) - \mathcal{L}(\tilde{\theta})) \\
&\overset{(i)}{\leq} \frac{1}{m(1-\gamma)}\sqrt{\frac{\ln(2/\delta)}{2|\mathcal{D}|}} + \left(\widehat{\mathcal{L}}\left(\theta^*\right) - \widehat{\mathcal{L}}(\tilde{\theta})\right) \\
&\quad + \frac{1}{m(1-\gamma)}\sqrt{\frac{\ln(2/\delta)}{2|\mathcal{D}|}},
\end{aligned}$$

Task	IQ-Learn	BC	GAIL	f-IRL	Single Loop ML-IRL	Expert
Hopper	**3546.4***	20.53 ± 4.16	2757.88 ± 293.18	2993.54 ± 252.59	3262.89 ± 24.31	3592.63
Half-Cheetah	5043.26	-1.77 ±0.27	3254.51 ± 87.40	4650.49 ± 180.94	**5074.51 ± 249.70**	5098.30
Walker	5134.0*	-13.98 ± 0.22	1023.41 ± 81.64	4483.14 ± 155.72	**5182.09 ± 74.69**	5344.21
Ant	4362.9*	760.14 ± 0.53	1107.98 ± 531.90	4853.53 ± 213.72	**4919.80 ± 11.52**	5926.18

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

where (i) comes from hoefffing inequality.

Since $\tilde{\theta}$ is defined as the optimal solution to problem (5), we know that $\widehat{\mathcal{L}}(\theta;\mathcal{D}) - \widehat{\mathcal{L}}(\tilde{\theta};\mathcal{D}) \leq 0$ for any θ. Therefore, we yield that, with probability at least $1-\delta$,

$$\mathcal{L}(\theta^*) - \mathcal{L}(\tilde{\theta}) \leq \frac{2}{m(1-\gamma)}\sqrt{\frac{\ln(2/\delta)}{2|\mathcal{D}|}}.$$

When the number of expert trajectories in the demonstration data set satisfies $|\mathcal{D}| \geq \frac{2}{\epsilon^2 m^2(1-\gamma)^2}\ln\left(\frac{2}{\delta}\right)$, then with probability greater than $1-\delta$, we obtain the performance guarantee of the optimality gap that

$$\mathcal{L}(\theta^*) - \mathcal{L}(\tilde{\theta}) \leq \epsilon.$$

So far, we have demonstrated that our algorithm can theoretically recover the ground truth reward estimator under overparameterization.

6 Experiments

In this section, we mainly compare our single-loop ML-IRL with two classes of algorithms: (a) IRL algorithms that learn the reward function and its corresponding policy simultaneously, such as f-IRL (Ni et al., 2021) and IQ-Learn (Garg et al., 2021). (b) 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).

Experimental Setup: We test the performance of different algorithms on several high-dimensional robotics control tasks in Mujoco (Todorov et al., 2012). To ensure a fair comparison with benchmark algorithms under these settings, we choose soft Actor-Critic (Haarnoja et al., 2018) as the base RL algorithm due to its strong performance in environments with continuous action spaces. Besides, soft Actor-Critic is a special variant of the policy gradient algorithm and equivalent to soft Q-learning (Haarnoja et al., 2018; Schulman et al., 2018). Cai et al. (2023) extends the convergence analysis of neural soft Q-learning to soft Actor-Critic. Therefore, it is reasonable for us to use soft Actor-Critic to update our policy in the numerical implementation.

In SAC for the baseline algorithms, both policy network and Q-network are $(64,64)$ MLPs with ReLU activation function, and stepsizes as 3×10^{-3}. We use their open-source implementations in our experiments for all mentioned imitation learning / IRL benchmark algorithms[1][2]. In our proposed single-loop algorithm[3], we parameterize the reward function by a $(64,64)$ MLPs with ReLU activation function. For the reward network, we use Adam as the optimizer, and the stepsize is set to be 1×10^{-4}. At the beginning of each iteration, we warm-start the SAC algorithm by initializing both the policy and Q-networks with the neural networks trained in the previous iteration. We then run 10 episodes in the corresponding Mujoco environment to further train these two networks. Following this, both agent and expert trajectories are collected to update the reward network with a single gradient step.

Numerical Results: In each experiment, a limited data regime is considered where the expert dataset only contains a single expert trajectory. We use the expert data provided in the open-source implementation of f-IRL [4]. We train the algorithms until they converge. The accumulated reward scores reported in Table 1 are averaged over 3 random seeds. In each random seed, we train the algorithms from initialization and collect 20 trajectories to average their accumulated rewards after the algorithms converge. The results presented in Table 1 indicate that our proposed single loop algorithm have better performances than existing baseline methods on the majority of tasks.

As we can observe, BC has the poorest ability to imitate our expert's behavior. This aligns with the fact that BC relies on supervised learning, which fails to learn the policy well within a limited data regime. Furthermore, we note that the training process of IQ-Learn is unstable, potentially due to its imprecise approximation of the soft Q-function. As a result, in the Half-Cheetah task, we are unable to replicate the results reported in the original paper (Garg et al., 2021) and have to directly report the results from the original paper, which is indicated by an asterisk (*) in Table 1. The results of AIRL are excluded from Table 1 due to its consistent poor performance, despite our extensive parameter tuning efforts. (Ni et al., 2021; Liu et al., 2019) have encountered similar issues.

7 Conclusions

In this paper, we advance the understanding of the maximum likelihood IRL framework and propose two provably efficient algorithms, which can recover the neural network parameterized reward function and the policy simultaneously. To our knowledge, Algorithm 3 is the first single-loop algorithm to address the neural network parameterized IRL problem with solid theoretical guarantees. Besides, extensive numerical experiments on Mujoco tasks demonstrate the superiority of our single-loop ML-IRL algorithm over traditional methods. A limitation of our method is the requirement of a sufficiently large width of the parameterization neural network, so one future direction of this work is to further extend our theoretical analysis to the finite-width case.

[1] https://github.com/Div99/IQ-Learn

[2] https://github.com/KamyarGh/rl_swiss

[3] https://github.com/Cloud0723/ML-IRL

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

Acknowledgements

M. Hong and S. Zeng are supported partially by NSF under the grants EPCN-2311007, ECCS-2426064 and CCF-2414372, also by Minnesota Supercomputing Institute. A. Garcia and C. Li are partially supported by W911NF-22-1-0213, ECCS-2240789 and CCF-2414373.

References

Abbeel, P. and Ng, A. Y. (2004). Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1.

Allen-Zhu, Z., Li, Y., and Liang, Y. (2020a). Learning and generalization in overparameterized neural networks, going beyond two layers.

Allen-Zhu, Z., Li, Y., and Liang, Y. (2020b). Learning and generalization in overparameterized neural networks, going beyond two layers.

Allen-Zhu, Z., Li, Y., and Song, Z. (2019). A convergence theory for deep learning via over-parameterization.

Arora, S., Du, S. S., Hu, W., Li, Z., and Wang, R. (2019). Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks.

Bertsekas, D. P. (2018). Feature-based aggregation and deep reinforcement learning: A survey and some new implementations. *IEEE/CAA Journal of Automatica Sinica*, 6(1):1–31.

Bhandari, J., Russo, D., and Singal, R. (2018). A finite time analysis of temporal difference learning with linear function approximation.

Borkar, V. S. (1997). Stochastic approximation with two time scales. *Systems & Control Letters*, 29(5):291–294.

Cai, Q., Yang, Z., Lee, J. D., and Wang, Z. (2023). Neural temporal difference and q learning provably converge to global optima. *Mathematics of Operations Research*.

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

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

Daniely, A. (2017). Sgd learns the conjugate kernel class of the network.

Du, S. and Lee, J. (2018). On the power of over-parametrization in neural networks with quadratic activation. In *International conference on machine learning*, pages 1329–1338. PMLR.

Finn, C., Levine, S., and Abbeel, P. (2016). Guided cost learning: Deep inverse optimal control via policy optimization. In *International conference on machine learning*, pages 49–58. PMLR.

Fu, Z., Yang, Z., and Wang, Z. (2021). Single-timescale actor-critic provably finds globally optimal policy.

Gao, R., Cai, T., Li, H., Hsieh, C.-J., Wang, L., and Lee, J. D. (2019). Convergence of adversarial training in overparametrized neural networks. *Advances in Neural Information Processing Systems*, 32.

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

Gaur, M., Bedi, A. S., Wang, D., and Aggarwal, V. (2023). On the global convergence of natural actor-critic with two-layer neural network parametrization. *arXiv preprint arXiv:2306.10486*.

Geist, M. and Pietquin, O. (2013). Algorithmic survey of parametric value function approximation. *IEEE Transactions on Neural Networks and Learning Systems*, 24(6):845–867.

Guan, Z., Xu, T., and Liang, Y. (2021). When will generative adversarial imitation learning algorithms attain global convergence. In *International Conference on Artificial Intelligence and Statistics*, pages 1117–1125. PMLR.

Haarnoja, T., Tang, H., Abbeel, P., and Levine, S. (2017). Reinforcement learning with deep energy-based policies. In *International conference on machine learning*, pages 1352–1361. PMLR.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR.

Ho, J. and Ermon, S. (2016). Generative adversarial imitation learning. *Advances in neural information processing systems*, 29.

Hofmann, T., Schölkopf, B., and Smola, A. J. (2008). Kernel methods in machine learning.

Hong, M., Wai, H., Wang, Z., and Yang, Z. (2020). A two-timescale framework for bilevel optimization: Complexity analysis and application to actor-critic. arxiv e-prints, art. *arXiv preprint arXiv:2007.05170*.

Ji, K., Yang, J., and Liang, Y. (2021). Bilevel optimization: Convergence analysis and enhanced design. In *International conference on machine learning*, pages 4882–4892. PMLR.

Jin, C., Netrapalli, P., and Jordan, M. (2020). What is local optimality in nonconvex-nonconcave minimax optimization? In *International conference on machine learning*, pages 4880–4889. PMLR.

Khanduri, P., Zeng, S., Hong, M., Wai, H.-T., Wang, Z., and Yang, Z. (2021). A near-optimal algorithm for stochastic bilevel optimization via double-momentum. *Advances in neural information processing systems*, 34:30271–30283.

Kim, K., Garg, S., Shiragur, K., and Ermon, S. (2021). Reward identification in inverse reinforcement learning. In *International Conference on Machine Learning*, pages 5496–5505. PMLR.

Kurach, K., Lucic, M., Zhai, X., Michalski, M., and Gelly, S. (2018). The gan landscape: Losses, architectures, regularization, and normalization. *ArXiv*, abs/1807.04720.

Li, Y. and Liang, Y. (2018). Learning overparameterized neural networks via stochastic gradient descent on structured data. *Advances in neural information processing systems*, 31.

Liu, F., Ling, Z., Mu, T., and Su, H. (2019). State alignment-based imitation learning.

Liu, Z., Zhang, Y., Fu, Z., Yang, Z., and Wang, Z. (2021). Provably efficient generative adversarial imitation learning for online and offline setting with linear function approximation. *arXiv preprint arXiv:2108.08765*.

Maei, H., Szepesvari, C., Bhatnagar, S., Precup, D., Silver, D., and Sutton, R. S. (2009). Convergent temporal-difference learning with arbitrary smooth function approximation. *Advances in neural information processing systems*, 22.

Melo, F. S., Meyn, S. P., and Ribeiro, M. I. (2008). An analysis of reinforcement learning with function approximation. In *Proceedings of the 25th international conference on Machine learning*, pages 664–671.

Neu, G. and Szepesvari, C. (2012). Apprenticeship learning using inverse reinforcement learning and gradient methods.

Ni, T., Sikchi, H., Wang, Y., Gupta, T., Lee, L., and Eysenbach, B. (2021). f-irl: Inverse reinforcement learning via state marginal matching. In *Conference on Robot Learning*, pages 529–551. PMLR.

Osa, T., Pajarinen, J., Neumann, G., Bagnell, J. A., Abbeel, P., and Peters, J. (2018). An algorithmic perspective on imitation learning. *Foundations and Trends in Robotics*, 7(1-2):1–179.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. (2022). Training language models to follow instructions with human feedback.

Phan-Minh, T., Howington, F., Chu, T.-S., Lee, S. U., Tomov, M. S., Li, N., Dicle, C., Findler, S., Suarez-Ruiz, F., Beaudoin, R., Yang, B., Omari, S., and Wolff, E. M. (2022). Driving in real life with inverse reinforcement learning.

Pomerleau, D. A. (1988). Alvinn: An autonomous land vehicle in a neural network. *Advances in neural information processing systems*, 1.

Rajaraman, N., Yang, L. F., Jiao, J., and Ramachandran, K. (2020). Toward the fundamental limits of imitation learning.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. (2023). Bridging offline reinforcement learning and imitation learning: A tale of pessimism.

Reddy, S., Dragan, A. D., and Levine, S. (2019). Sqil: Imitation learning via reinforcement learning with sparse rewards. *arXiv preprint arXiv:1905.11108*.

Ross, S. and Bagnell, D. (2010). Efficient reductions for imitation learning. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 661–668. JMLR Workshop and Conference Proceedings.

Scaman, K. and Virmaux, A. (2019). Lipschitz regularity of deep neural networks: analysis and efficient estimation.

Schulman, J., Chen, X., and Abbeel, P. (2018). Equivalence between policy gradients and soft q-learning.

Syed, U., Bowling, M., and Schapire, R. E. (2008). Apprenticeship learning using linear programming. In *Proceedings of the 25th international conference on Machine learning*, pages 1032–1039.

Syed, U. and Schapire, R. E. (2007). A game-theoretic approach to apprenticeship learning. *Advances in neural information processing systems*, 20.

Todorov, E., Erez, T., and Tassa, Y. (2012). Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 5026–5033. IEEE.

Wang, L., Cai, Q., Yang, Z., and Wang, Z. (2019). Neural policy gradient methods: Global optimality and rates of convergence. *arXiv preprint arXiv:1909.01150*.

Wu, Y., Zhang, W., Xu, P., and Gu, Q. (2022). A finite time analysis of two time-scale actor critic methods.

Wulfmeier, M., Ondruska, P., and Posner, I. (2015). Maximum entropy deep inverse reinforcement learning. *arXiv preprint arXiv:1507.04888*.

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

Yang, S. Y., Yu, Y., and Almahdi, S. (2018). An investor sentiment reward-based trading system using gaussian inverse reinforcement learning algorithm. *Expert Systems with Applications*, 114:388–401.

Zeng, S., Li, C., Garcia, A., and Hong, M. (2022). Maximum-likelihood inverse reinforcement learning with finite-time guarantees. *Advances in Neural Information Processing Systems*, 35:10122–10135.

Zeng, S., Li, C., Garcia, A., and Hong, M. (2023). Understanding expertise through demonstrations: A maximum likelihood framework for offline inverse reinforcement learning. *arXiv preprint arXiv:2302.07457*.

Zeng, S., Li, C., Garcia, A., and Hong, M. (2024). When demonstrations meet generative world models: A maximum likelihood framework for offline inverse reinforcement learning.

Zhang, Y., Cai, Q., Yang, Z., and Wang, Z. (2020). Generative adversarial imitation learning with neural networks: Global optimality and convergence rate. *arXiv preprint arXiv:2003.03709*.

Ziebart, B. D., Bagnell, J. A., and Dey, A. K. (2010). Modeling interaction via the principle of maximum causal entropy. In *International Conference on Machine Learning*.

Ziebart, B. D., Maas, A. L., Bagnell, J. A., Dey, A. K., et al. (2008). Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pages 1433–1438. Chicago, IL, USA.

Zou, D., Cao, Y., Zhou, D., and Gu, Q. (2018). Stochastic gradient descent optimizes over-parameterized deep relu networks. arxiv e-prints, art. *arXiv preprint arXiv:1811.08888*.

Zou, S., Xu, T., and Liang, Y. (2019). Finite-sample analysis for sarsa with linear function approximation.

Checklist

1. For all models and algorithms presented, check if you include:
    (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]

   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [No]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Function Approximation of Soft Q-learning Cai et al. (2023)

Definition A *The soft Bellman optimality operator is defined as:*

$$\mathcal{T}(Q^{\text{soft}}_{\widehat{r}_{\theta_k},\pi})(s,a) = \mathbb{E}\left[r(s,a) + \gamma \cdot \underset{a' \in \mathcal{A}}{\text{softmax}} Q^{\text{soft}}_{\widehat{r}_{\theta_k},\pi}\left(s', a'\right) \mid s' \sim \mathcal{P}(\cdot \mid s,a)\right], \tag{12}$$

for which $Q^{\text{soft}}_{\widehat{r}_{\theta_k},\pi}$ is the fixed point. In other words, $Q^{\text{soft}}_{\widehat{r}_{\theta_k},\pi} = \mathcal{T}(Q^{\text{soft}}_{\widehat{r}_{\theta_k},\pi})$, which minimizes the following Mean Squared Bellman Error (MSBE) problem: $\mathbb{E}_{(s,a,s')\sim\mu}\left[(Q(x) - \mathcal{T}Q(x))^2\right]$.

In this work, we utilize the neural network as a nonlinear function approximation. An approximate function class $\mathcal{F}_W$ is defined as $\left\{\widehat{Q}(\cdot;W) + \nabla_W \widehat{Q}(\cdot;W)^\top (W' - W) : W' \in S_B\right\}$. $S_B = \left\{W' \in \mathbb{R}^{m\times d} : \|W' - W\|_2 \leq B\right\}$ is the parameter space of feasible W, which consists of the local linearization of $\widehat{Q}(\cdot;W')$ at point W (Cai et al., 2023). The bounded domain limits the distance between the parameters W' and the initial parameters W, enhancing training stability. Since we know the distribution of initialization W_0, we take $W = W_0$ here. Recall that we adopt a double-layer neural network with ReLU activation, the approximate function class can be defined as follows.

Definition B *(Approximate Function Class)*

$$\mathcal{F}_{W_0} = \left\{\frac{1}{\sqrt{m}} \sum_{r=1}^{m} b^j \mathbb{1}\left\{{W_0^j}^\top x > 0\right\} {W^j}^\top x : W \in S_B\right\}.$$

To find the best neural network approximation of $Q^{\text{soft}}_{\widehat{r}_{\theta_k},\pi}$, we usually turn to seek the stationary point $\widehat{Q}_0(\cdot;W^*_{\theta_k})$ of a surrogate measure for the MSBE on $\mathcal{F}_{W_0}$, which is referred to as the Mean Squared Projected Bellman Error (MSPBE).

$$\min_W \text{MSPBE}(W) = \mathbb{E}_{(s,a)\sim\mu}\left[\left(\widehat{Q}(s,a;W) - \Pi_{\mathcal{F}_W}\mathcal{T}\widehat{Q}(s,a;W)\right)^2\right], \tag{13}$$

where the soft Q-function is parametrized by $\widehat{Q}(s,a;W)$ with parameter W. In this context, μ signifies the stationary distribution of (s,a) induced by the policy π. The projection onto a function class $\mathcal{F}_W$ is represented as $\Pi_{\mathcal{F}_W}$. We then introduce two lemmas to support the feasibility of focusing on the stationary point of MSPBE for convergence analysis.

Lemma 1 *(Lemma 4.2 in Cai et al. (2023)) (Existence, Uniqueness, and Optimality of $\widehat{Q}_0\left(\cdot;W^*_{\theta_k}\right)$). There exists a stationary point $W^*_{\theta_k}$ for any $b \in \mathbb{R}^m$ and $W_0 \in \mathbb{R}^{m\times d}$. Also, $\widehat{Q}_0\left(\cdot;W^*_{\theta_k}\right)$ is unique almost everywhere and is the global optimum of the MSPBE that corresponds to the projection onto $\mathcal{F}_{W_0}$.*

This lemma establishes the unique existence of an approximate stationary point, as it corresponds to the fixed point of the operator $\Pi_{\mathcal{F}_{W_0}}\mathcal{T}$. This is because the projection operator is an ℓ_2-norm contraction, associated with the stationary distribution μ.

Lemma 2 *(Proposition 4.7 in Cai et al. (2023))*

$$\left\|\widehat{Q}_0\left(\cdot;W^*_{\theta_k}\right) - Q^{\text{soft}}_{\widehat{r}_{\theta_k},\pi}(\cdot)\right\|_\mu \leq (1-\gamma)^{-1} \cdot \left\|\Pi_{\mathcal{F}_{B,m}} Q^{\text{soft}}_{\widehat{r}_{\theta_k},\pi} - Q^{\text{soft}}_{\widehat{r}_{\theta_k},\pi}(\cdot)\right\|_\mu, \tag{14}$$

where $Q^{\text{soft}}_{\widehat{r}_{\theta_k},\pi}$ is the fixed point of soft Bellman optimality operator defined in (12).

When the neural network is overparameterized with width $m \to \infty$ and sufficiently large S_B, the projection $\Pi_{\mathcal{F}_{B,\infty}}$ reduces to the identity operator Hofmann et al. (2008). By Lemma 2, $\widehat{Q}_0\left(\cdot;W^*_{\theta_k}\right) = Q^{\text{soft}}_{\widehat{r}_{\theta_k},\pi}(\cdot)$. Therefore, $\widehat{Q}_0\left(\cdot;W^*_{\theta_k}\right)$ can be seen as the global optimum of the MSBE.

Therefore, it suffices to consider $\mathcal{F}_{W_0}$ in place of $\mathcal{F}_{W^\dagger}$ for our convergence analysis.

B Auxiliary Lemmas

Assumptions 1, 2 and 3 hold throughout the entire section. Necessary proofs of some auxiliary lemmas can be found in the next Section C.

Due to the fact that the optimal policy has the closed form $\pi_\theta(\cdot \mid s) \propto \exp\left(Q^{\text{soft}}_{r_\theta,\pi_\theta}(s,\cdot)\right)$ (Haarnoja et al., 2017), we can establish the following lemma to express the objective function $\widehat{\mathcal{L}}(\theta)$ in an equivalent form.

Lemma 3

$$\widehat{\mathcal{L}}(\theta) = \mathbb{E}_{\tau^E \sim D}\left[\sum_{t=0}^{\infty} \gamma^t \widehat{r}(s_t, a_t; \theta)\right] - \mathbb{E}_{\tau^{\mathrm{A}} \sim \pi_\theta}\left[\sum_{t=0}^{\infty} \gamma^t \left(\widehat{r}(\mathbf{s}_t, \mathbf{a}_t; \theta) + \mathcal{H}\left(\pi\left(\cdot \mid \mathbf{s}_t\right)\right)\right)\right]. \tag{15}$$

The equivalent formulation of the objective directly leads to the following lemma without proof, providing a closed-form expression for the gradient of the objective function.

Lemma 4 *(Lemma 4.1 in Zeng et al. (2022)) The gradient of the likelihood objective $\widehat{\mathcal{L}}(\theta)$ in (3.2) can be equivalently expressed as follows:*

$$\nabla \widehat{\mathcal{L}}(\theta) = \mathbb{E}_{\tau^E \sim \mathcal{D}}\left[\sum_{t \geq 0} \gamma^t \nabla_\theta \widehat{r}(s_t, a_t; \theta)\right] - \mathbb{E}_{\tau^{\mathrm{A}} \sim \pi_\theta}\left[\sum_{t \geq 0} \gamma^t \nabla_\theta \widehat{r}(s_t, a_t; \theta)\right]. \tag{16}$$

Before we further investigate the theoretical properties of likelihood objective $\widehat{\mathcal{L}}(\theta)$, we continue introducing some related lemmas:

Lemma 5 *(Proposition 1 in Scaman and Virmaux (2019), Lemma 5.3 in Zeng et al. (2022)) Let soft Q-function and reward function be defined as (6a) (6b) respectively. For any $s \in \mathcal{S}, a \in \mathcal{A}$ and any reward parameter θ, the following holds:*

$$|\widehat{r}(s, a; \theta_1) - \widehat{r}(s, a; \theta_2)| \leq \frac{1}{\sqrt{m}} \|\theta_1 - \theta_2\|, \tag{17}$$

$$\left|\widehat{Q}(s, a; W_1) - \widehat{Q}_(s, a; W_2)\right| \leq \frac{1}{\sqrt{m}} \|W_1 - W_2\|, \tag{18}$$

$$\left|Q^{soft}_{r_{\theta_1}, \pi_{\theta_1}}(s, a) - Q^{soft}_{r_{\theta_2}, \pi_{\theta_2}}(s, a)\right| \leq \frac{1}{\sqrt{m}(1-\gamma)} \|\theta_1 - \theta_2\|. \tag{19}$$

This lemma shows that the parameterized reward and soft Q-function are Lipschitz continuous, and therefore have a bounded gradient. We note that it's a natural result of our specific neural network structure, which relaxes the commonly used Lispchitz continuity assumption in the literature.

Lemma 6 *(Xu et al. (2020),Lemma 3) Consider the initialization distribution $\eta(\cdot)$ and transition kernel $\mathcal{P}(\cdot \mid s, a)$. Under $\mu_0(\cdot)$ and $\mathcal{P}(\cdot \mid s, a)$, denote $\mu_\pi(\cdot, \cdot)$ as the state-action visitation distribution of MDP with the Boltzmann policy parameterized by parameter θ. For all policy parameter θ and θ', we have*

$$\|\mu_\pi(\cdot, \cdot) - \mu_{\pi'}(\cdot, \cdot)\|_{TV} \leq C_d \left\|\theta - \theta'\right\|,$$

where C_d is a positive constant.

Lemma 7 *Under the same reward function, we can establish the following bound of accumulated rewards under different policies:*

$$\begin{aligned} &\mathbb{E}\left[\left\|\mathbb{E}_{\tau^A \sim \pi_{\theta_k}}\left[\sum_{t \geq 0} \gamma^t \nabla_\theta \widehat{r}(s_t, a_t; \theta_k)\right] - \mathbb{E}_{\tau^{A'} \sim \pi_{k+1}}\left[\sum_{t \geq 0} \gamma^t \nabla_\theta \widehat{r}(s_t, a_t; \theta_k)\right]\right\|\right] \\ &\leq 2L_q C_d \mathbb{E}\left[\left\|\widehat{Q}^{\text{soft}}_{\widehat{r}_{\theta_k}, \pi_{\theta_k}} - \widehat{Q}^{\text{soft}}_{\widehat{r}_{\theta_k}, \pi_k}\right\|\right], \end{aligned} \tag{20}$$

where the constant $L_q := \frac{1}{\sqrt{m}(1-\gamma)}$.

Similar to the local linearization of the soft Q function, we have the following definition for the reward function.

Definition C *(Local Linearization of reward function) The neural network $\widehat{r}(x; \theta)$ can be locally linearized at random initialization point θ_0 by $\widehat{r}_0(x; \theta)$ with respect to θ:*

$$\widehat{r}_0(x; \theta) = \Psi(x)^\top \theta, \tag{21}$$

where $\Psi(x) = \frac{1}{\sqrt{m}} \cdot \left(\mathbb{1}\left\{\theta_0^{1\top} x > 0\right\} x, \ldots, \mathbb{1}\left\{\theta_0^{m\top} x > 0\right\} x\right) \in \mathbb{R}^{m \times d}$.

Lemma 8 *Let $V(\theta) = \widehat{r}(s_t, a_t; \theta) - \widehat{r}_0(s_t, a_t; \theta)$ be the error of local linearization of the reward function, then its gradient can be bounded as:*

$$\nabla V(\theta) = \mathcal{O}\left(m^{-1/2}\right), \tag{22}$$

where $\widehat{r}_0(s_t, a_t; \theta)$ is defined in 21

We establish this lemma to characterize the gradient error bound of local linearization of the reward function, which is critical to the next proposition.

Proposition 1 *Approximate Lipschitz Smoothness*

$$\left\|\nabla \widehat{\mathcal{L}}\left(\theta_1\right)-\nabla \widehat{\mathcal{L}}\left(\theta_2\right)\right\| \leq \mathcal{O}\left(m^{-1/2}\right)+L_c\left\|\theta_1-\theta_2\right\|, \tag{23}$$

where $L_c=2 L_q C_d \frac{1}{\sqrt{m}(1-\gamma)}$.

The Lipschitz smoothness property is common in the literature of min-max / bi-level optimization Hong et al. (2020); Khanduri et al. (2021); Jin et al. (2020); Guan et al. (2021); Chen et al. (2021). We get an approximate Lipschitz smoothness property by studying the local linearization of the neural network parametrized reward function.

Lemma 9 *(Lemma 4.5 in Cai et al. (2023)) Recall that Bellman residual is calculated as* $\delta_k := z(k) = \widehat{Q}(s,a;W_k) - \widehat{r}\left(s,a;\theta_k\right)-\gamma \operatorname{softmax}_{a' \in \mathcal{A}} \widehat{Q}\left(s', a'; W_k\right)$. *There exists* $\sigma_z^2 = O\left(B^2\right)$ *such that the variance of the TD updates can be upper bounded as follows:*

$$\mathbb{E}_{init, \mu_k}\left[\|z(k)\|_2^2\right] \leq \sigma_z^2, \tag{24}$$

$$\mathbb{E}_{init, \mu_k}\left[\left\|z(k)-\mathbb{E}_{\mu_k}[z(k)]\right\|_2^2\right] \leq \sigma_z^2, \tag{25}$$

where the expectation is over initialization of parameter and the induced state-action distribution of current π_k.

Lemma 10 *The stochastic gradient estimation* $g_k := h\left(\theta_k; \tau_k^E\right)-h\left(\theta_k; \tau_k^A\right)$, *where* $h(\theta; \tau^E) := \sum_{t \geq 0} \gamma^t \nabla_\theta \widehat{r}\left(s_t, a_t; \theta\right)$

$$\left\|g_k\right\| \leq 2 L_q, \tag{26}$$

This lemma provides a constant upper bound of stochastic updates of reward.
The following two lemmas are first established in Cai et al. (2023) on studying the convergence of neural soft Q-learning. We'll use them in our lower-level analysis.

Lemma 11 *Descent Lemma in soft Q-function (C.27 in Cai et al. (2023))*

$$\begin{aligned} & \mathbb{E}\left[\left(\widehat{Q}_0(x;W_k)-\widehat{Q}_0\left(x;W_{\theta_k}^*\right)\right)^2\right] \\ & \leq \frac{\mathbb{E}\left[\left\|W_k-W_{\theta_k}^*\right\|_2^2\right]-\mathbb{E}\left[\left\|W_{k+1}-W_{\theta_k}^*\right\|_2^2\right]+\eta^2 \sigma_z^2}{\left(2 \eta_k(1-\gamma)-8 \eta^2\right)}+\mathcal{O}\left(B^3 m^{-1}+B^{5/2} m^{-1/2}\right), \end{aligned} \tag{27}$$

where the expectation is taken over all randomness and σ_z *is defined in Lemma 9.* $W_{\theta_k}^*$ *is the parameter of the optimal soft Q-function under reward* $\widehat{r}_{\theta_k}$.

Lemma 12 *Convergence bound of soft Q-learning (Theorem E.2 in Cai et al. (2023)) We set* η *to be of order* $T^{-1/2}$ *in Algorithm 1. Under Assumptions 2 and 3, the output* $\widehat{Q}_{out}$ *of Algorithm 1 satisfies*

$$\mathbb{E}\left[\left(\widehat{Q}_{out}(x)-\widehat{Q}_0\left(x;W_\theta^*\right)\right)^2\right]=\mathcal{O}\left(B^2 T^{-1/2}+B^3 m^{-1}+B^{5/2} m^{-1/2}\right), \tag{28}$$

where T *represents the number of iterations of Algorithm 1 and the expectation is over all randomness.*

Note that the network error term here is slightly different from the original one in Theorem E.2 in Cai et al. (2023). This is because we adopt a different neural network parameterization.

C Proof of Auxiliary Lemmas

C.1 Proof of Lemma 3:

First, our objective function $\widehat{\mathcal{L}}(\theta)$ in (3.2) is defined as below:

$$\widehat{\mathcal{L}}(\theta):=\mathbb{E}_{\tau^E \sim D}\left[\sum_{t=0}^{\infty} \gamma^t \log \pi_\theta\left(a_t \mid s_t\right)\right] \stackrel{(i)}{=} \mathbb{E}_{\tau^E \sim D}\left[\sum_{t=0}^{\infty} \gamma^t \log \left(\frac{\exp\left(\widehat{Q}_{\widehat{r}_\theta, \pi_\theta}^{\text{soft}}\left(s_t, a_t\right)\right)}{\sum_a \exp\left(\widehat{Q}_{\widehat{r}_\theta, \pi_\theta}^{\text{soft}}\left(s_t, a\right)\right)}\right)\right],$$

where (i) is due to the fact that the optimal policy has the closed form $\pi_\theta(\cdot \mid s) \propto \exp\left(\widehat{Q}^{\text{soft}}_{\widehat{r}_\theta,\pi_\theta}(s,\cdot)\right)$. Therefore, the objective function can be expressed in the following form:

$$\begin{aligned}
\widehat{\mathcal{L}}(\theta) &:= \mathbb{E}_{\tau^E \sim D}\left[\sum_{t=0}^{\infty} \gamma^t \left(\widehat{Q}^{\text{soft}}_{\widehat{r}_\theta,\pi_\theta}(s_t,a_t) - \log\left(\sum_a \exp\left(\widehat{Q}^{\text{soft}}_{\widehat{r}_\theta,\pi_\theta}(s_t,a)\right)\right)\right)\right] \\
&\overset{(i)}{=} \mathbb{E}_{\tau^E \sim D}\left[\sum_{t=0}^{\infty} \gamma^t \left(\widehat{Q}^{\text{soft}}_{\widehat{r}_\theta,\pi_\theta}(s_t,a_t) - V^{\text{soft}}_{\widehat{r}_\theta,\pi_\theta}(s_t)\right)\right] \\
&\overset{(ii)}{=} \mathbb{E}_{\tau^E \sim D}\left[\sum_{t=0}^{\infty} \gamma^t \left(\widehat{r}(s_t,a_t;\theta) + \gamma V^{\text{soft}}_{\widehat{r}_\theta,\pi_\theta}(s_{t+1}) - V^{\text{soft}}_{\widehat{r}_\theta,\pi_\theta}(s_t)\right)\right] \\
&= \mathbb{E}_{\tau^E \sim D}\left[\sum_{t=0}^{\infty} \gamma^t \widehat{r}(s_t,a_t;\theta)\right] + \mathbb{E}_{\tau^E \sim D}\left[\sum_{t=1}^{\infty} \gamma^t V^{\text{soft}}_{\widehat{r}_\theta,\pi_\theta}(s_t)\right] - \mathbb{E}_{\tau^E \sim D}\left[\sum_{t=0}^{\infty} \gamma^t V^{\text{soft}}_{\widehat{r}_\theta,\pi_\theta}(s_t)\right] \\
&= \mathbb{E}_{\tau^E \sim D}\left[\sum_{t=0}^{\infty} \gamma^t \widehat{r}(s_t,a_t;\theta)\right] - \mathbb{E}_{s_0 \sim \mu_0(\cdot)}\left[V^{\text{soft}}_{\widehat{r}_\theta,\pi_\theta}(s_0)\right] \\
&\overset{(iii)}{=} \mathbb{E}_{\tau^E \sim D}\left[\sum_{t=0}^{\infty} \gamma^t \widehat{r}(s_t,a_t;\theta)\right] - \mathbb{E}_{\tau^{\mathrm{A}} \sim \pi_\theta}\left[\sum_{t=0}^{\infty} \gamma^t \left(\widehat{r}(\mathbf{s}_t,\mathbf{a}_t;\theta) + \mathcal{H}(\pi(\cdot \mid \mathbf{s}_t))\right)\right],
\end{aligned}$$

where (i) follows the fact that the optimal soft value function could be expressed as $V^{\text{soft}}_{\widehat{r}_\theta,\pi_\theta}(s) = \log\left(\sum_a \exp\left(\widehat{Q}^{\text{soft}}_{\widehat{r}_\theta,\pi_\theta}(s,a)\right)\right)$. (ii) and (iii) is derived from the definition of soft Q value (3) and soft V value (2).

C.2 Proof of Lemma 5

We only prove the soft Q-function case here and the same reasoning also applies to reward function, since they employ the same parametrization structure. $\nabla_{W^j}\widehat{Q}(x;W) = \frac{1}{m} b_j \mathbb{1}\left\{{W^j}^\top x > 0\right\} x$ almost everywhere, the l_2 norm is bounded as follows:

$$\left\|\nabla_W \widehat{Q}(x;W)\right\|_2 \le \frac{1}{m}\left(\sum_{j=1}^{m} \mathbb{1}\left\{{W^j}^\top x > 0\right\} \|b_j\|_2^2 \|x\|_2^2\right)^{\frac{1}{2}} \le \frac{1}{\sqrt{m}},$$

Therefore, we can prove the lemma by the mean value theorem. Proof of inequality (19) is the same as Zeng et al. (2022) Lemma 5.3.

C.3 Proof of Lemma 7

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbb{E}_{\tau \sim \pi_{\theta_k}}\left[\sum_{t \ge 0} \gamma^t \nabla_\theta \widehat{r}(s_t,a_t;\theta_k)\right] - \mathbb{E}_{\tau \sim \pi_{k+1}}\left[\sum_{t \ge 0} \gamma^t \nabla_\theta \widehat{r}(s_t,a_t;\theta_k)\right]\right\|\right] \\
&\overset{(i)}{=} \mathbb{E}\left[\left\|\frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d(\cdot,\cdot;\pi_{\theta_k})}\left[\nabla_\theta \widehat{r}(s,a;\theta_k)\right] - \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d(\cdot,\cdot;\pi_{k+1})}\left[\nabla_\theta \widehat{r}(s,a;\theta_k)\right]\right\|\right] \\
&\overset{(ii)}{\le} \frac{2}{1-\gamma} \cdot \max_{s \in \mathcal{S}, a \in \mathcal{A}} \|\nabla_\theta \widehat{r}(s,a;\theta_k)\| \cdot \mathbb{E}\left[\|d(\cdot,\cdot;\pi_{\theta_k}) - d(\cdot,\cdot;\pi_{k+1})\|_{TV}\right] \\
&\overset{(iii)}{\le} \frac{2}{\sqrt{m}(1-\gamma)}\mathbb{E}\left[\|d(\cdot,\cdot;\pi_{\theta_k}) - d(\cdot,\cdot;\pi_{k+1})\|_{TV}\right] \\
&\overset{(iv)}{\le} 2L_q C_d \mathbb{E}\left[\left\|\widehat{Q}^{\text{soft}}_{\widehat{r}_{\theta_k},\pi_{\theta_k}} - \widehat{Q}^{\text{soft}}_{\widehat{r}_{\theta_k},\pi_k}\right\|\right],
\end{aligned} \tag{29}$$

where (i) follows the definition $d(s,a;\pi) = (1-\gamma)\pi(a \mid s)\sum_{t \ge 0}\gamma^t \mathcal{P}^\pi(s_t = s)$; (ii) is due to distribution mismatch between two visitation measures; (iii) follows the inequality (17) in Lemma 5; the inequality (iv) follows Lemma 6 and the fact that $\pi_{\theta_k}(\cdot \mid s) \propto \exp\left(\widehat{Q}^{\text{soft}}_{\widehat{r}_{\theta_k},\pi_{\theta_k}}(s,\cdot)\right)$, $\pi_{k+1}(\cdot \mid s) \propto \exp\left(\widehat{Q}^{\text{soft}}_{\widehat{r}_{\theta_k},\pi_k}(s,\cdot)\right)$, where the constant $L_q := \frac{1}{\sqrt{m}(1-\gamma)}$.

C.4 Proof of Lemma 8

First, we prove $V(\theta)$ is Lipschitz continuous.

$$\begin{aligned}
\left\|V(\theta) - V(\theta')\right\|_2 &\le \left\|\widehat{r}(s_t,a_t;\theta) - \widehat{r}(s_t,a_t;\theta')\right\|_2 + \left\|r_0(s_t,a_t;\theta) - r_0(s_t,a_t;\theta')\right\|_2 \\
&\overset{(\mathrm{i})}{\le} \frac{2}{\sqrt{m}}\left\|\theta - \theta'\right\|_2,
\end{aligned}$$

where (i) is using the Lipshitz continuity property from (17) in Lemma 5.
Now we need to show $\nabla V(\theta)$ is bounded:

$$|\nabla V(\theta)| = \lim_{\omega\to 0}\frac{|V(\theta+\omega)-V(\theta)|}{|\omega|} \leq \lim_{\omega\to 0}\frac{\frac{2}{\sqrt{m}}|\theta+\omega-\theta|}{|\omega|} = \frac{2}{\sqrt{m}},$$

Therefore, the lemma is shown.

C.5 Proof of Proposition 1

By the triangle inequality,

$$\begin{aligned}\|\nabla_\theta \widehat{r}(x;\theta_1)-\nabla_\theta \widehat{r}(x;\theta_2)\| \leq{}& \underbrace{\|\nabla_\theta \widehat{r}(x;\theta_1)-\nabla_\theta \widehat{r}_0(x;\theta_1)\|}_{\text{Term I}} + \underbrace{\|\nabla_\theta \widehat{r}_0(x;\theta_2)-\nabla_\theta \widehat{r}(x;\theta_2)\|}_{\text{Term II}} \\ &+ \underbrace{\|\nabla_\theta \widehat{r}_0(x;\theta_1)-\nabla_\theta \widehat{r}_0(x;\theta_2)\|}_{\text{Term III}}.\end{aligned} \tag{30}$$

Both Term I and Term II are bounded by $\mathcal{O}(m^{-\frac{1}{2}})$ in Lemma 8.

For Term III, recall that $\nabla_\theta \widehat{r}_0(x;\theta) = \frac{1}{m}b\mathbb{1}\left\{{\theta_0}^\top x > 0\right\}x$. The following inequality holds:

$$\|\nabla_\theta \widehat{r}_0(x;\theta_1)-\nabla_\theta \widehat{r}_0(x;\theta_2)\| = 0. \tag{31}$$

We then plug (31) into (30) and yield the following inequality:

$$\|\nabla_\theta \widehat{r}(x;\theta_1)-\nabla_\theta \widehat{r}(x;\theta_2)\| \leq \mathcal{O}\left(m^{-1/2}\right). \tag{32}$$

By Lemma 4, we have:
$\nabla\widehat{\mathcal{L}}(\theta) = \mathbb{E}_{\tau^E\sim\mathcal{D}}\left[\sum_{t\geq 0}\gamma^t\nabla_\theta\widehat{r}(s_t,a_t;\theta)\right] - \mathbb{E}_{\tau^{\mathrm{A}}\sim\pi_\theta}\left[\sum_{t\geq 0}\gamma^t\nabla_\theta\widehat{r}(s_t,a_t;\theta)\right].$

Therefore,

$$\begin{aligned}&\left\|\nabla\widehat{\mathcal{L}}(\theta_1)-\nabla\widehat{\mathcal{L}}(\theta_2)\right\| \\ \leq{}& \underbrace{\left\|\mathbb{E}_{\tau^E\sim D}\left[\sum_{t\geq 0}\gamma^t\nabla_\theta\widehat{r}(s_t,a_t;\theta_1)\right]-\mathbb{E}_{\tau^E\sim D}\left[\sum_{t\geq 0}\gamma^t\nabla_\theta\widehat{r}(s_t,a_t;\theta_2)\right]\right\| +}_{:=\text{term A}} \\ & \underbrace{\left\|\mathbb{E}_{\tau^A\sim\pi_{\theta_1}}\left[\sum_{t\geq 0}\gamma^t\nabla_\theta\widehat{r}(s_t,a_t;\theta_1)\right]-\mathbb{E}_{\tau^{A'}\sim\pi_{\theta_2}}\left[\sum_{t\geq 0}\gamma^t\nabla_\theta\widehat{r}(s_t,a_t;\theta_2)\right]\right\|}_{:=\text{term B}}.\end{aligned} \tag{33}$$

For term A, it follows that:

$$\begin{aligned}&\left\|\mathbb{E}_{\tau^E\sim D}\left[\sum_{t\geq 0}\gamma^t\nabla_\theta\widehat{r}(s_t,a_t;\theta_1)\right]-\mathbb{E}_{\tau^E\sim D}\left[\sum_{t\geq 0}\gamma^t\nabla_\theta\widehat{r}(s_t,a_t;\theta_2)\right]\right\| \\ \overset{(i)}{\leq}{}& \mathbb{E}_{\tau^E\sim D}\left[\sum_{t\geq 0}\gamma^t\left\|\nabla_\theta\widehat{r}(s_t,a_t;\theta_1)-\nabla_\theta\widehat{r}(s_t,a_t;\theta_2)\right\|\right] \\ \overset{(ii)}{\leq}{}& \mathcal{O}(m^{-1/2}),\end{aligned}$$

where (i) follows the Jensen's inequality and (ii) is from inequality (32).

For term B,

$$
\begin{aligned}
&\left\|\mathbb{E}_{\tau\sim\pi_{\theta_1}}\left[\sum_{t\geq 0}\gamma^t\nabla_\theta\widehat{r}(s_t,a_t;\theta_1)\right]-\mathbb{E}_{\tau\sim\pi_{\theta_2}}\left[\sum_{t\geq 0}\gamma^t\nabla_\theta\widehat{r}(s_t,a_t;\theta_2)\right]\right\|\\
\overset{(i)}{\leq}&\left\|\mathbb{E}_{\tau\sim\pi_{\theta_1}}\left[\sum_{t\geq 0}\gamma^t\nabla_\theta\widehat{r}(s_t,a_t;\theta_1)\right]-\mathbb{E}_{\tau\sim\pi_{\theta_2}}\left[\sum_{t\geq 0}\gamma^t\nabla_\theta\widehat{r}(s_t,a_t;\theta_1)\right]\right\|\\
&+\left\|\mathbb{E}_{\tau\sim\pi_{\theta_2}}\left[\sum_{t\geq 0}\gamma^t\nabla_\theta\widehat{r}(s_t,a_t;\theta_1)\right]-\mathbb{E}_{\tau\sim\pi_{\theta_2}}\left[\sum_{t\geq 0}\gamma^t\nabla_\theta\widehat{r}(s_t,a_t;\theta_2)\right]\right\|\\
\overset{(ii)}{\leq}&\left\|\mathbb{E}_{\tau\sim\pi_{\theta_1}}\left[\sum_{t\geq 0}\gamma^t\nabla_\theta\widehat{r}(s_t,a_t;\theta_1)\right]-\mathbb{E}_{\tau\sim\pi_{\theta_2}}\left[\sum_{t\geq 0}\gamma^t\nabla_\theta\widehat{r}(s_t,a_t;\theta_1)\right]\right\|\\
&+\mathbb{E}_{\tau\sim\pi_{\theta_2}}\left[\sum_{t\geq 0}\gamma^t\left\|\nabla_\theta\widehat{r}(s_t,a_t;\theta_1)-\nabla_\theta\widehat{r}(s_t,a_t;\theta_2)\right\|\right]\\
\overset{(iii)}{\leq}&\ 2L_qC_d\left\|\widehat{Q}^{\text{soft}}_{\widehat{r}_{\theta_1},\pi_{\theta_1}}-\widehat{Q}^{\text{soft}}_{\widehat{r}_{\theta_2},\pi_{\theta_2}}\right\|_2+\mathcal{O}(m^{-1/2})\\
\overset{(iv)}{\leq}&\ 2L_qC_d\frac{1}{\sqrt{m}(1-\gamma)}\left\|\theta_1-\theta_2\right\|_2+\mathcal{O}(m^{-1/2}),
\end{aligned}
$$

where (i) follows the triangle inequality, (ii) is from Jensen's inequality ; (iii) is from (32) and Lemma 7; (iv) follows inequality (19) in Lemma 5.

Combining term A and term B, we obtain the approximate L-smoothness property:

$$
\left\|\nabla\widehat{\mathcal{L}}(\theta_1)-\nabla\widehat{\mathcal{L}}(\theta_2)\right\|\leq\mathcal{O}\left(m^{-1/2}\right)+L_c\left\|\theta_1-\theta_2\right\|, \tag{34}
$$

where $L_c=2L_qC_d\frac{1}{\sqrt{m}(1-\gamma)}$.

C.6 Proof of Lemma 10

$\|g_k\|\leq\left\|h\left(\theta_k,\tau_k^E\right)\right\|+\left\|h\left(\theta_k,\tau_k^A\right)\right\|\leq 2\frac{1}{\sqrt{m}}\sum_{t\geq 0}\gamma^t=\frac{2}{\sqrt{m}(1-\gamma)}=2L_q$.

D Lower-level Analysis

Lemma 13 *Given that the policies π_{k+1} and π_{θ_k} are in the softmax parameterization, we can establish an upper bound for the difference between π_{k+1} and π_{θ_k} as follows:*

$$
\left\|\log\pi_{k+1}-\log\pi_{\theta_k}\right\|_\infty\leq 2\left\|\widehat{Q}^{soft}_{\widehat{r}_{\theta_k},\pi_k}-\widehat{Q}^{soft}_{\widehat{r}_{\theta_k},\pi_{\theta_k}}\right\|_2 \tag{35}
$$

Proof D.1 *We first denote an operator $\log\left(\|\exp(v)\|_1\right):=\log\left(|\sum_{\tilde{a}\in\mathcal{A}}\exp\left(v_{\tilde{a}}\right)|\right)$, where the vector $v\in\mathbb{R}^{|\mathcal{A}|}$ and $v=\left[v_1,v_2,\cdots,v_{|\mathcal{A}|}\right]$. Then for any $v',v''\in\mathbb{R}^{|\mathcal{A}|}$, we have the following relation:*

$$
\begin{aligned}
\log\left(\left\|\exp\left(v'\right)\right\|_1\right)-\log\left(\left\|\exp\left(v''\right)\right\|_1\right)&\overset{(i)}{=}\left\langle v'-v'',\nabla_v\log\left(\|\exp(v)\|_1\right)|_{v=v_c}\right\rangle\\
&\leq\left\|v'-v''\right\|_\infty\cdot\left\|\nabla_v\log\left(\|\exp(v)\|_1\right)|_{v=v_c}\right\|_1\\
&\overset{(ii)}{=}\left\|v'-v''\right\|_\infty,
\end{aligned} \tag{36}
$$

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

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

For any $s \in \mathcal{S}$ and $a \in \mathcal{A}$, we have the following relationship:

$$
\begin{aligned}
&\left|\log\left(\pi_{k+1}(a \mid s)\right) - \log\left(\pi_{\theta_k}(a \mid s)\right)\right| \\
&\stackrel{(i)}{=} \left|\log\left(\frac{\exp\left(\widehat{Q}^{soft}_{\widehat{r}_{\theta_k},\pi_k}(s,a)\right)}{\sum_{\tilde{a}}\exp\left(\widehat{Q}^{\text{soft}}_{\widehat{r}_{\theta_k},\pi_k}(s,\tilde{a})\right)}\right) - \log\left(\frac{\exp\left(\widehat{Q}^{soft}_{\widehat{r}_{\theta_k},\pi_{\theta_k}}(s,a)\right)}{\sum_{\tilde{a}}\exp\left(\widehat{Q}^{\text{soft}}_{\widehat{r}_{\theta_k},\pi_{\theta_k}}(s,\tilde{a})\right)}\right)\right| \\
&\stackrel{(ii)}{\leq} \left|\log\left(\sum_{\tilde{a}}\exp\left(\widehat{Q}^{\text{soft}}_{\widehat{r}_{\theta_k},\pi_k}(s,\tilde{a})\right)\right) - \log\left(\sum_{\tilde{a}}\exp\left(\widehat{Q}^{soft}_{\widehat{r}_{\theta_k},\pi_{\theta_k}}(s,\tilde{a})\right)\right)\right| \\
&\quad + \left|\widehat{Q}^{soft}_{\widehat{r}_{\theta_k},\pi_k}(s,a) - \widehat{Q}^{soft}_{\widehat{r}_{\theta_k},\pi_{\theta_k}}(s,a)\right|,
\end{aligned}
\tag{37}
$$

where (i) is from $\pi(a \mid s) \propto \exp\left(Q^{soft}(s,a)\right)$ and (ii) is the result of triangle inequality.

We plug (36) into (37) and yield that

$$
\begin{aligned}
&\left|\log\left(\pi_{k+1}(a \mid s)\right) - \log\left(\pi_{\theta_k}(a \mid s)\right)\right| \\
&\leq \left|\widehat{Q}^{\text{soft}}_{\widehat{r}_{\theta_k},\pi_k}(s,a) - \widehat{Q}^{soft}_{\widehat{r}_{\theta_k},\pi_{\theta_k}}(s,a)\right| + \max_{\tilde{a}\in\mathcal{A}}\left|\widehat{Q}^{soft}_{\widehat{r}_{\theta_k},\pi_k}(s,\tilde{a}) - \widehat{Q}^{\text{soft}}_{\widehat{r}_{\theta_k},\pi_{\theta_k}}(s,\tilde{a})\right|.
\end{aligned}
\tag{38}
$$

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

$$
\left\|\log\pi_{k+1} - \log\pi_{\theta_k}\right\|_\infty \leq 2\left\|\widehat{Q}^{soft}_{\widehat{r}_{\theta_k},\pi_k} - \widehat{Q}^{soft}_{\widehat{r}_{\theta_k},\pi_{\theta_k}}\right\|_\infty,
\tag{39}
$$

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

By the definition of supremum norm and l_2 norm, we have the following relation:

$$
\left\|\widehat{Q}^{soft}_{\widehat{r}_{\theta_k},\pi_k} - \widehat{Q}^{soft}_{\widehat{r}_{\theta_k},\pi_{\theta_k}}\right\|_\infty \leq \left\|\widehat{Q}^{soft}_{\widehat{r}_{\theta_k},\pi_k} - \widehat{Q}^{soft}_{\widehat{r}_{\theta_k},\pi_{\theta_k}}\right\|_2.
\tag{40}
$$

Combining the inequality (39) and (40), we prove the lemma and now only need to analyze $\left\|\widehat{Q}^{soft}_{\widehat{r}_{\theta_k},\pi_k} - \widehat{Q}^{soft}_{\widehat{r}_{\theta_k},\pi_{\theta_k}}\right\|_2$ to show the convergence of the policy parameter.

Theorem 4 *(Norm of the Projection Operator). Let $v \in \mathbb{R}^n$ and $\Pi_{\mathcal{F}_{B,m}}$ projects v onto the subspace $\mathcal{F}_{B,m}$,*

$$
\left\|\Pi_{\mathcal{F}_{B,m}}v\right\|_2 \leq \|v\|_2.
$$

The theorem can be easily verified by decomposing v into $v = \Pi_{\mathcal{F}_{B,m}}v + r$, where r is orthogonal to $\Pi_{\mathcal{F}_{B,m}}v$.

D.1 Algorithm 2

We set $\eta_k = \min\{1/\sqrt{K}, \frac{1-\gamma}{8}\}$, Note that when $K \geq (8/(1-\gamma))^2$, we have $\eta_k = \frac{1}{\sqrt{K}}$. Therefore,

$$
\begin{aligned}
\frac{1}{K}\sum_{k=0}^{K-1}\left\|\widehat{Q}^{\text{soft}}_{\widehat{r}_{\theta_k},\pi_k} - \widehat{Q}^{\text{soft}}_{\widehat{r}_{\theta_k},\pi_{\theta_k}}\right\|_2 &\stackrel{(i)}{\leq} \frac{B}{K}\sum_{k=0}^{K-1}(k+2)^{-1/4} + \mathcal{O}\left(B^{3/2}m^{-1/2} + B^{5/4}m^{-1/4}\right) \\
&\leq \frac{B}{K}\int_{k=0}^{K-1}(k+1)^{-\frac{1}{4}} + \mathcal{O}\left(B^{3/2}m^{-1/2} + B^{5/4}m^{-1/4}\right) \\
&\leq BK^{-\frac{1}{4}} + \mathcal{O}\left(B^{3/2}m^{-1/2} + B^{5/4}m^{-1/4}\right),
\end{aligned}
$$

where (i) comes from the dynamic truncation design that $T = k+2$ in Algorithm 1 and Lemma 12. Therefore, we have:

$$
\frac{1}{K}\sum_{k=0}^{K-1}\left\|\widehat{Q}^{\text{soft}}_{\widehat{r}_{\theta_k},\pi_k} - \widehat{Q}^{\text{soft}}_{\widehat{r}_{\theta_k},\pi_{\theta_k}}\right\|_2 \leq BK^{-\frac{1}{4}} + \mathcal{O}\left(B^{3/2}m^{-1/2} + B^{5/4}m^{-1/4}\right).
\tag{41}
$$

Following (39), we obtain that

$$
\frac{1}{K}\sum_{k=0}^{K-1}\left\|\log\pi_{k+1} - \log\pi_{\theta_k}\right\|_\infty = \mathcal{O}(K^{-\frac{1}{4}}) + \mathcal{O}\left(B^{3/2}m^{-1/2} + B^{5/4}m^{-1/4}\right).
\tag{42}
$$

D.2 Algorithm 3

Telescope (27) in Lemma 11 for k=0, . . . , K-1,

$$\frac{1}{K}\sum_{k=0}^{K-1}\left\|\widehat{Q}^{\text{soft}}_{\widehat{r}_{\theta_k},\pi_k}-\widehat{Q}^{\text{soft}}_{\widehat{r}_{\theta_k},\pi_{\theta_k}}\right\|_2^2 \le \frac{1}{K}\sum_{k=0}^{K-1}\frac{\left[\mathbb{E}\left[\left\|W_k-W^*_{\pi_k}\right\|_2^2\right]-\mathbb{E}\left[\left\|W_{k+1}-W^*_{\pi_k}\right\|_2^2\right]+\eta^2\sigma_z^2\right]}{2\eta(1-\gamma)-8\eta^2} + \mathcal{O}\left(B^3m^{-1}+B^{5/2}m^{-1/2}\right). \tag{43}$$

For each k, we apply triangle inequality on the right-hand side of (43) and yield the following inequality:

$$\frac{1}{K}\sum_{k=0}^{K-1}\left\|\widehat{Q}^{\text{soft}}_{\widehat{r}_{\theta_k},\pi_k}-\widehat{Q}^{\text{soft}}_{\widehat{r}_{\theta_k},\pi_{\theta_k}}\right\|_2^2 \le \frac{1}{K}\sum_{k=0}^{K-1}\frac{\mathbb{E}\left[\|W_k-W_{k+1}\|_2^2\right]+\eta^2\sigma_z^2}{2\eta(1-\gamma)-8\eta^2}+\mathcal{O}\left(B^3m^{-1}+B^{5/2}m^{-1/2}\right) \tag{44}$$

Recall that projection step (line 6) $W_{k+1}=\Pi_{\mathcal{F}_{B,m}}\widetilde{W}_{k+1}$ in Algorithm 3, we apply Theorem 4 to (44) and the following relation holds:

$$\frac{1}{K}\sum_{k=0}^{K-1}\left\|\widehat{Q}^{\text{soft}}_{\widehat{r}_{\theta_k},\pi_k}-\widehat{Q}^{\text{soft}}_{\widehat{r}_{\theta_k},\pi_{\theta_k}}\right\|_2^2 \le \frac{1}{K}\sum_{k=0}^{K-1}\frac{\mathbb{E}\left[\left\|W_k-\widetilde{W}_{k+1}\right\|_2^2\right]+\eta^2\sigma_z^2}{2\eta(1-\gamma)-8\eta^2}+\mathcal{O}\left(B^3m^{-1}+B^{5/2}m^{-1/2}\right). \tag{45}$$

By the TD update (line 5) of the Algorithm 3, we know that

$$\widetilde{W}_{k+1}=W_k-\eta\delta\cdot\nabla_W\widehat{Q}(s,a;W_k). \tag{46}$$

Plug (46) into (45), we get that:

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K-1}\left\|\widehat{Q}^{\text{soft}}_{\widehat{r}_{\theta_k},\pi_k}-\widehat{Q}^{\text{soft}}_{\widehat{r}_{\theta_k},\pi_{\theta_k}}\right\|_2^2 &\le \frac{1}{K}\sum_{k=0}^{K-1}\frac{\mathbb{E}\left[\eta^2\delta_k^2\left\|\nabla_W\widehat{Q}(s,a;W_k)\right\|_2^2\right]+\eta^2\sigma_z^2}{2\eta(1-\gamma)-8\eta^2}+\mathcal{O}\left(B^3m^{-1}+B^{5/2}m^{-1/2}\right)\\ &\overset{(i)}{\le}\frac{B^2+\sigma_z^2}{K}\sum_{k=0}^{K-1}\frac{\eta^2}{2\eta(1-\gamma)-8\eta^2}+\mathcal{O}\left(B^3m^{-1}+B^{5/2}m^{-1/2}\right),\end{aligned}$$

where (i) comes from Lemma 9.

If we set the stepsize $\eta\le\frac{1}{\sqrt{K}}$, we obtain the following result:

$$\frac{1}{K}\sum_{k=0}^{K-1}\left\|\widehat{Q}^{\text{soft}}_{\widehat{r}_{\theta_k},\pi_k}-\widehat{Q}^{\text{soft}}_{\widehat{r}_{\theta_k},\pi_{\theta_k}}\right\|_2^2 \le \frac{B^2+\sigma_z^2}{K}\frac{1}{2\eta(1-\gamma)-8\eta^2}+\mathcal{O}\left(B^3m^{-1}+B^{5/2}m^{-1/2}\right). \tag{47}$$

We select $\eta=\min\{K^{-\frac{3}{4}},(1-\gamma)/8\}$, which satisfies the condition $\eta\le\frac{1}{\sqrt{K}}$ required in (47). Note that when $K\ge(8/(1-\gamma))^{\frac{4}{3}}$, we have $\eta=K^{-\frac{3}{4}}$ and

$$K^{\frac{3}{4}}\cdot\left(2\eta(1-\gamma)-8\eta^2\right)=2(1-\gamma)-8K^{-\frac{3}{4}}\ge1-\gamma.$$

When $K<(8/(1-\gamma))^{\frac{4}{3}}$, we have $\eta=(1-\gamma)/8$ and

$$K^{\frac{3}{4}}\cdot\left(2\eta(1-\gamma)-8\eta^2\right)=K^{\frac{3}{4}}\cdot(1-\gamma)^2/8\ge(1-\gamma)^2/8.$$

Since $|1-\gamma|<1$, we obtain that for any $K\in\mathbb{N}$,

$$\frac{1}{K^{\frac{3}{4}}\cdot(2\eta(1-\gamma)-8\eta^2)}\le\frac{8}{(1-\gamma)^2}. \tag{48}$$

Then,

$$\frac{1}{K}\sum_{k=0}^{K-1}\left\|\widehat{Q}^{\text{soft}}_{\widehat{r}_{\theta_k},\pi_k}-\widehat{Q}^{\text{soft}}_{\widehat{r}_{\theta_k},\pi_{\theta_k}}\right\|_2^2 \le \frac{B^2+\sigma_z^2}{K^{\frac{1}{4}}}\frac{8}{(1-\gamma)^2}+\mathcal{O}\left(B^3m^{-1}+B^{5/2}m^{-1/2}\right). \tag{49}$$

Following Lemma 13, we have:

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K-1}\|\log\pi_{k+1}-\log\pi_{\theta_k}\|_\infty &\leq \frac{2}{K}\sum_{k=0}^{K-1}\left\|\widehat{Q}^{\text{soft}}_{\widehat{r}_{\theta_k},\pi_k}-\widehat{Q}^{\text{soft}}_{\widehat{r}_{\theta_k},\pi_{\theta_k}}\right\|_2\\ &\overset{(i)}{\leq} 2\sqrt{\frac{1}{K}\sum_{k=0}^{K-1}\left\|\widehat{Q}^{\text{soft}}_{\widehat{r}_{\theta_k},\pi_k}-\widehat{Q}^{\text{soft}}_{\widehat{r}_{\theta_k},\pi_{\theta_k}}\right\|_2^2}\\ &\overset{(ii)}{\leq} 2\sqrt{\frac{B^2+\sigma_z^2}{K}\frac{1}{2\eta(1-\gamma)-8\eta^2}+\mathcal{O}\left(B^3m^{-1}+B^{5/2}m^{-1/2}\right)}\\ &\overset{(iii)}{\leq} 2\sqrt{B^2+\sigma_z^2}\sqrt{\frac{1}{K^{\frac{1}{4}}}\frac{8}{(1-\gamma)^2}}+\mathcal{O}\left(B^{3/2}m^{-1/2}+B^{5/4}m^{-1/4}\right),\end{aligned} \tag{50}$$

where (i) comes from the Cauchy-Schwartz inequality, (ii) is due to (49), and (iii) is given by (48). Therefore, we obtain the convergence guarantee of policy parameter for Algorithm 3:

$$\frac{1}{K}\sum_{k=0}^{K-1}\|\log\pi_{k+1}-\log\pi_{\theta_k}\|_\infty=\mathcal{O}(K^{-\frac{1}{8}})+\mathcal{O}\left(B^{3/2}m^{-1/2}+B^{5/4}m^{-1/4}\right). \tag{51}$$

E Convergence of reward parameters

Since Proposition 1 establishes the approximate Lipschitz smooth property for the objective function, we have the following result of $\widehat{\mathcal{L}}(\theta)$:

$$\widehat{\mathcal{L}}(\theta_{k+1})+\mathcal{O}\left(m^{-1/2}\right)\geq\widehat{\mathcal{L}}(\theta_K)+\left\langle\nabla\widehat{\mathcal{L}}(\theta_K),\theta_{k+1}-\theta_k\right\rangle-\frac{L_c}{2}\|\theta_{k+1}-\theta_k\|^2. \tag{52}$$

Notice the reward update rule (line 13) of the Algorithm 3,

$$\theta_{k+1}:=\theta_k+\alpha g_k. \tag{53}$$

We plug (53) into (52) and get:

$$\begin{aligned}\widehat{\mathcal{L}}(\theta_{k+1})+\mathcal{O}\left(m^{-1/2}\right)&\geq\widehat{\mathcal{L}}(\theta_K)+\alpha\left(\nabla\widehat{\mathcal{L}}(\theta_K),g_k\right\rangle-\frac{L_c\alpha^2}{2}\|g_k\|^2\\ &=\widehat{\mathcal{L}}(\theta_K)+\alpha\left(\nabla\widehat{\mathcal{L}}(\theta_K),g_k-\nabla\widehat{\mathcal{L}}(\theta_K)\right\rangle+\alpha\left\|\nabla\widehat{\mathcal{L}}(\theta_K)\right\|^2-\frac{L_c\alpha^2}{2}\|g_k\|^2\\ &\overset{(i)}{\geq}\widehat{\mathcal{L}}(\theta_K)+\alpha\left(\nabla\widehat{\mathcal{L}}(\theta_K),g_k-\nabla\widehat{\mathcal{L}}(\theta_K)\right\rangle+\alpha\left\|\nabla\widehat{\mathcal{L}}(\theta_K)\right\|^2-2L_cL_q^2\alpha^2,\end{aligned} \tag{54}$$

where (i) is from the upper bound of stochastic reward updates in Lemma 10.

Taking an expectation over both sides of (54), it holds that

$$\begin{aligned}&\mathbb{E}\left[\widehat{\mathcal{L}}(\theta_{k+1})\right]+\mathcal{O}\left(m^{-1/2}\right)\\ &\geq\mathbb{E}\left[\widehat{\mathcal{L}}(\theta_K)\right]+\alpha\mathbb{E}\left[\left\langle\nabla\widehat{\mathcal{L}}(\theta_K),g_k-\nabla\widehat{\mathcal{L}}(\theta_K)\right\rangle\right]+\alpha\mathbb{E}\left[\left\|\nabla\widehat{\mathcal{L}}(\theta_K)\right\|^2\right]-2L_cL_q^2\alpha^2\\ &=\mathbb{E}\left[\widehat{\mathcal{L}}(\theta_K)\right]+\alpha\mathbb{E}\left[\left\langle\nabla\widehat{\mathcal{L}}(\theta_K),\mathbb{E}\left[g_k-\nabla\widehat{\mathcal{L}}(\theta_K)\mid\theta_k\right]\right\rangle\right]+\alpha\mathbb{E}\left[\left\|\nabla\widehat{\mathcal{L}}(\theta_K)\right\|^2\right]-2L_cL_q^2\alpha^2.\end{aligned} \tag{55}$$

Combining the definition of g_k in line 12 of Algorithm 3 and Lemma 4 into (55), we have:

$$
\begin{aligned}
& \mathbb{E}\left[\widehat{\mathcal{L}}(\theta_{k+1})\right] + \mathcal{O}\left(m^{-1/2}\right) \\
\geq\ & \mathbb{E}\left[\widehat{\mathcal{L}}(\theta_K)\right] + \alpha\mathbb{E}\left[\left\|\nabla\widehat{\mathcal{L}}(\theta_K)\right\|^2\right] - 2L_cL_q^2\alpha^2 \\
& + \alpha\mathbb{E}\left[\left\langle \nabla\widehat{\mathcal{L}}(\theta_K), \mathbb{E}_{\tau^E\sim\pi_{\theta_k}}\left[\sum_{t\geq 0}\gamma^t\nabla_\theta\widehat{r}(s_t,a_t;\theta_t)\right] - \mathbb{E}_{\tau^E\sim\pi_{k+1}}\left[\sum_{t\geq 0}\gamma^t\nabla_\theta\widehat{r}(s_t,a_t;\theta_t)\right]\right\rangle\right] \\
\overset{(i)}{\geq}\ & \mathbb{E}\left[\widehat{\mathcal{L}}(\theta_K)\right] - 2\alpha L_q\mathbb{E}\left[\left\|\mathbb{E}_{\tau^E\sim\pi_{\theta_k}}\left[\sum_{t\geq 0}\gamma^t\nabla_\theta\widehat{r}(s_t,a_t;\theta_k)\right] - \mathbb{E}_{\tau^E\sim\pi_{k+1}}\left[\sum_{t\geq 0}\gamma^t\nabla_\theta\widehat{r}(s_t,a_t;\theta_k)\right]\right\|\right] \\
& + \alpha\mathbb{E}\left[\left\|\nabla\widehat{\mathcal{L}}(\theta_K)\right\|^2\right] - 2L_cL_q^2\alpha^2,
\end{aligned}
$$

where (i) follows Lemma 5 and Cauchy-Schwartz inequality.

Therefore, we derive the following important relation by rearranging the inequality above:

$$
\begin{aligned}
\alpha\mathbb{E}\left[\left\|\nabla\widehat{\mathcal{L}}(\theta_K)\right\|^2\right] &\leq 2L_cL_q^2\alpha^2 + \mathbb{E}\left[\widehat{\mathcal{L}}(\theta_{k+1}) - \widehat{\mathcal{L}}(\theta_K)\right] + \mathcal{O}\left(m^{-1/2}\right) \\
&+ 2\alpha L_q\mathbb{E}\left[\left\|\mathbb{E}_{\tau^E\sim\pi_{\theta_k}}\left[\sum_{t\geq 0}\gamma^t\nabla_\theta\widehat{r}(s_t,a_t;\theta_k)\right] - \mathbb{E}_{\tau^E\sim\pi_{k+1}}\left[\sum_{t\geq 0}\gamma^t\nabla_\theta\widehat{r}(s_t,a_t;\theta_k)\right]\right\|\right].
\end{aligned} \tag{56}
$$

By invoking Lemma 7, the following relation holds:

$$
\alpha\mathbb{E}\left[\left\|\nabla\widehat{\mathcal{L}}(\theta_K)\right\|^2\right] \leq 2L_cL_q^2\alpha^2 + \alpha C_1\mathbb{E}\left[\left\|Q^{\rm soft}_{r_{\theta_k},\pi_{\theta_k}} - Q^{\rm sof}_{r_{\theta_k},\pi_k}\right\|_2\right] + \mathbb{E}\left[\widehat{\mathcal{L}}(\theta_{k+1}) - \widehat{\mathcal{L}}(\theta_K)\right] + \mathcal{O}\left(m^{-1/2}\right), \tag{57}
$$

where $C_1 = 4\alpha C_d L_q^2$.

Summing the inequality (57) from $k=0$ to $K-1$ and dividing both sides by αK, it leads to

$$
\begin{aligned}
\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\widehat{\mathcal{L}}(\theta_K)\right\|^2\right] \leq 2L_cL_q^2\alpha + \frac{C_1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\widehat{Q}^{\rm soft}_{\widehat{r}_{\theta_k},\pi_{\theta_k}} - \widehat{Q}^{\rm soft}_{\widehat{r}_{\theta_k},\pi_k}\right\|_2\right] + \mathbb{E}\left[\frac{\widehat{\mathcal{L}}(\theta_K) - \widehat{\mathcal{L}}(\theta_0)}{K\alpha}\right] \\
+ \mathcal{O}\left(m^{-1/2}\right).
\end{aligned} \tag{58}
$$

E.1 Theorem 1

The stepsize of reward update is set as $\alpha = \frac{\alpha_0}{K^\sigma}$, where $\sigma > 0$. We plug (41) into (58) and get the convergence result of reward parameter:

$$
\begin{aligned}
\frac{1}{K}\sum_{K=0}^{K-1}\mathbb{E}\left[\left\|\nabla\widehat{\mathcal{L}}(\theta_K)\right\|^2\right] &= \mathcal{O}\left(K^{-\sigma}\right) + \mathcal{O}\left(K^{-1+\sigma}\right) + \mathcal{O}\left(K^{-\frac{1}{4}}\right) \\
&+ \mathcal{O}\left(m^{-1/2} + B^{3/2}m^{-1/2} + B^{5/4}m^{-1/4}\right).
\end{aligned} \tag{59}
$$

E.2 Theorem 2

The stepsize of reward update is set as $\alpha = \frac{\alpha_0}{K^\sigma}$, where $\sigma > 0$. We plug (49) into (58) and get the convergence result of reward parameter:

$$
\begin{aligned}
\frac{1}{K}\sum_{K=0}^{K-1}\mathbb{E}\left[\left\|\nabla\widehat{\mathcal{L}}(\theta_K)\right\|^2\right] &= \mathcal{O}\left(K^{-\sigma}\right) + \mathcal{O}\left(K^{-1+\sigma}\right) + \mathcal{O}\left(K^{-\frac{1}{8}}\right) + \mathcal{O}\left(K^{-1+\sigma}\right) + \mathcal{O}\left(K^{-\frac{1}{4}}\right) \\
&+ \mathcal{O}\left(m^{-1/2} + B^{3/2}m^{-1/4} + B^{5/4}m^{-1/8}\right).
\end{aligned} \tag{60}
$$

F Global Optimality of reward and policy parameters

Now we consider the general case that reward is parametrized by neural networks. Since $\widehat{\mathcal{L}}(\theta)$ is nonconcave in θ, showing the global optimality of problem (5) is non-trivial, it means that the stationary point of $\widehat{\mathcal{L}}(\theta)$ is not necessarily the global optimum. Therefore, we translate this problem to an equivalent saddle point problem:

$$\max_{\theta} \min_{\pi} \widehat{L}(\theta, \pi) := \mathbb{E}_{\tau^{\mathrm{E}} \sim \mathcal{D}}\left[\sum_{t=0}^{\infty} \gamma^t \left(\widehat{r}\left(s_t, a_t; \theta\right)\right)\right] - \mathbb{E}_{\tau^{\mathrm{A}} \sim \pi}\left[\sum_{t=0}^{\infty} \gamma^t \left(\widehat{r}\left(s_t, a_t; \theta\right) + \mathcal{H}\left(\pi\left(\cdot \mid s_t\right)\right)\right)\right].$$

Based on the equivalence between (5) and (11) , to show the global optimality of reward and policy parameters, we first show that any stationary point $\tilde{\theta}$ in (5) together with its corresponding optimal policy $\pi_{\tilde{\theta}}$ consist of a saddle point $\left(\tilde{\theta}, \pi_{\tilde{\theta}}\right)$ to the problem (11) when the reward is overparameterized. Later, we illustrate that the reward parameter $\tilde{\theta}$ of the saddle point is the global maximizer of the empirical bi-level probelm (5). Recall that a tuple $\left(\tilde{\theta}, \pi_{\tilde{\theta}}\right)$ is called a saddle point of $\widehat{L}(\cdot, \cdot)$ if the following condition holds:

$$\widehat{L}\left(\theta, \pi_{\tilde{\theta}}\right) \leq \widehat{L}\left(\tilde{\theta}, \pi_{\tilde{\theta}}\right) \leq \widehat{L}(\tilde{\theta}, \pi) \tag{61}$$

for any other reward parameter θ and policy π. To show that $\left(\tilde{\theta}, \pi_{\tilde{\theta}}\right)$ satisfies the condition (61), we prove the following conditions respectively:

$$\tilde{\theta} \in \arg\max_{\theta} \widehat{L}\left(\theta, \pi_{\tilde{\theta}}\right), \tag{62}$$

$$\pi_{\tilde{\theta}} \in \arg\min_{\pi} \widehat{L}(\tilde{\theta}, \pi). \tag{63}$$

Here, we first show that any stationary point $\tilde{\theta}$ of the objective $\widehat{\mathcal{L}}(\cdot)$ in (5) satisfies the optimality condition (62). By the first-order condition, any stationary point $\tilde{\theta}$ of the objective $\widehat{\mathcal{L}}(\cdot)$ satisfies:

$$\nabla \widehat{\mathcal{L}}(\tilde{\theta}) = \mathbb{E}_{\tau^E \sim \mathcal{D}}\left[\sum_{t \geq 0} \gamma^t \nabla_\theta \widehat{r}\left(s_t, a_t; \tilde{\theta}\right)\right] - \mathbb{E}_{\tau^{\mathrm{A}} \sim \pi_\theta}\left[\sum_{t \geq 0} \gamma^t \nabla_\theta \widehat{r}\left(s_t, a_t; \tilde{\theta}\right)\right] = 0. \tag{64}$$

Then we can look back at the minimax formulation $\widehat{L}(\cdot, \cdot)$ in (10). Given any fixed policy π, we can write down the gradient of $\widehat{L}(\pi, \theta)$ w.r.t. the reward parameter θ explicitly as below:

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

From the formulation, we could notice that $\nabla_\theta \widehat{L}\left(\theta_k, \pi_{\theta_k}\right) = \nabla \widehat{L}(\theta_k)$. Therefore, Proposition 1 can be generalized to the Theorem 3.

Due to the first-order condition we show in (64), we obtain the following first-order optimality result for the minimax objective $\widehat{L}(\cdot, \cdot)$:

$$\nabla_\theta \widehat{L}\left(\theta = \tilde{\theta}, \pi = \pi_{\tilde{\theta}}\right) = 0. \tag{65}$$

From Theorem 3, we have shown $\widehat{L}(\cdot, \cdot)$ is approximately concave in terms of the reward parameter θ given any fixed policy π, if the neural network is sufficiently large. With the concavity property in (10) and the condition in (65), we have completed the proof of the first condition (62).

Next, we move to prove the second condition (63). Recall that $\pi_{\tilde{\theta}}$ is the optimal policy defined in (5) under the reward parameter θ. By observing the objective $\widehat{L}(\cdot, \cdot)$, we obtain the following result:

$$\widehat{L}(\theta, \pi) := \underbrace{\mathbb{E}_{\tau^{\mathrm{E}} \sim (\eta, \mathcal{D})}\left[\sum_{t=0}^{\infty} \gamma^t \left(\widehat{r}\left(s_t, a_t; \theta\right)\right)\right]}_{\text{Term } \mathrm{I}_1\text{: independent of } \pi} \underbrace{- \mathbb{E}_{\tau^{\mathrm{A}} \sim (\eta, \pi)}\left[\sum_{t=0}^{\infty} \gamma^t \left(\widehat{r}\left(s_t, a_t; \theta\right) + \mathcal{H}\left(\pi\left(\cdot \mid s_t\right)\right)\right)\right]}_{\text{Term } \mathrm{I}_2\text{ : a function of the policy } \pi}. \tag{66}$$

The second term in (66) coincides with the lower-level objective in (5). Moreover, since $\pi_{\tilde{\theta}}$ is the optimal policy defined in (5) under reward parameter $\tilde{\theta}$, the following statement naturally holds:

$$\pi_{\tilde{\theta}} \in \arg\min_{\pi} \widehat{L}(\tilde{\theta}, \pi).$$

We now complete the proof to show the second condition (63). Hence, we prove that any stationary point $\tilde{\theta}$ of the bi-level objective $\widehat{\mathcal{L}}(\cdot)$ together with its optimal policy $\pi_{\tilde{\theta}}$ is a saddle point of $\widehat{L}(\cdot,\cdot)$

Given a saddle point $\left(\tilde{\theta}, \pi_{\tilde{\theta}}\right)$ of $\widehat{L}(\cdot,\cdot)$, we have the following property that:

$$\min_{\pi}\max_{\theta}\widehat{L}(\theta,\pi) \le \max_{\theta}\widehat{L}\left(\theta,\pi_{\tilde{\theta}}\right) \overset{(i)}{=} \widehat{L}(\tilde{\theta},\pi_{\tilde{\theta}}) \overset{(ii)}{=} \min_{\pi}\widehat{L}(\tilde{\theta},\pi) \le \max_{\theta}\min_{\pi}\widehat{L}(\theta,\pi), \tag{67}$$

where (i) follows the optimality condition (62) and (ii) follows the optimality condition (63). According to the minimax inequality, we always have the following condition that

$$\max_{\theta}\min_{\pi} L(\theta,\pi) \le \min_{\pi}\max_{\theta} L(\theta,\pi). \tag{68}$$

Putting the saddle point inequality (67) and the minimax inequality (68) together, the following equality holds:

$$\min_{\pi}\max_{\theta}\widehat{L}(\theta,\pi) = \max_{\theta}\widehat{L}\left(\theta,\pi_{\tilde{\theta}}\right) = \widehat{L}\left(\tilde{\theta},\pi_{\tilde{\theta}}\right) = \min_{\pi}\widehat{L}(\tilde{\theta},\pi) = \max_{\theta}\min_{\pi}\widehat{L}(\theta,\pi).$$

Therefore, for any saddle point $\left(\tilde{\theta}, \pi_{\tilde{\theta}}\right)$, the reward parameter $\tilde{\theta}$ and the corresponding policy $\pi_{\tilde{\theta}}$ satisfy the following:

$$\tilde{\theta} \in \arg\max_{\theta}\min_{\pi}\widehat{L}(\theta,\pi), \tag{69}$$

$$\pi_{\tilde{\theta}} \in \arg\min_{\pi}\max_{\theta}\widehat{L}(\theta,\pi). \tag{70}$$

Due to the expression of the bi-level objective $\widehat{\mathcal{L}}(\cdot)$ in (5) and the objective $\widehat{L}(\cdot,\cdot)$ in (10), we have the following equality relationship for any reward parameter θ :

$$\widehat{\mathcal{L}}(\theta) = \min_{\pi}\widehat{L}(\theta,\pi). \tag{71}$$

Combining (69) and (71), we yield the following result:

$$\tilde{\theta} \in \arg\max_{\theta}\min_{\pi}\widehat{L}(\theta,\pi) = \arg\max_{\theta}\widehat{\mathcal{L}}(\theta).$$

Till now, we prove that for any saddle point $\left(\tilde{\theta}, \pi_{\tilde{\theta}}\right)$ of $\widehat{L}(\cdot,\cdot)$, the reward parameter $\tilde{\theta}$ constructs a globally optimal solution of the bi-level objective $\widehat{\mathcal{L}}(\cdot)$ in (5), when the neural network is overparameterized.