FEDERATED MAXIMUM LIKELIHOOD INVERSE REINFORCEMENT LEARNING WITH CONVERGENCE GUARANTEE

Anonymous authors
Paper under double-blind review

ABSTRACT

Inverse Reinforcement Learning (IRL) aims to recover the latent reward function and corresponding optimal policy from observed demonstrations. Existing IRL research predominantly focuses on a centralized learning approach, not suitable for real-world problems with distributed data and privacy restrictions. To this end, this paper proposes a novel algorithm for federated maximum-likelihood IRL (F-ML-IRL) and provides a rigorous analysis of its convergence and time-complexity. The proposed F-ML-IRL leverages a dual-aggregation to update the shared global model and performs bi-level local updates – an upper-level learning task to optimize the parameterized reward function by maximizing the discounted likelihood of observing expert trajectories under the current policy and a low-level learning task to find the optimal policy concerning the entropy-regularized discounted cumulative reward under the current reward function. We analyze the convergence and time-complexity of the proposed F-ML-IRL algorithm and show that the global model in F-ML-IRL converges to a stationary point for both the reward and policy parameters within finite time, i.e., the log-distance between the recovered policy and the optimal policy, as well as the gradient of the likelihood objective, converge to zero. Finally, evaluating our F-ML-IRL algorithm on high-dimensional robotic control tasks in MuJoCo, we show that it ensures convergences of the recovered reward in decentralized learning and even outperforms centralized baselines due to its ability to utilize distributed data.

1 INTRODUCTION

Inverse learning is the problem of modeling the preferences and goals of an agent using its observed behavior (Arora & Doshi, 2020). When the behavior of a human expert is observed through demonstration trajectories containing state and action data, Inverse Reinforcement Learning (IRL) models the policy through a Markov Decision Process (MDP) to recover the latent reward function and potentially replicate the human expert's optimal policy (Russell, 1998). The learned reward function can support various downstream tasks such as agent modeling and transfer learning (Sutton & Barto, 2018; Arora & Doshi, 2020). Recent work has developed provably-efficient IRL algorithms, such as Generative Adversarial Inverse Learning (GAIL) (Ho & Ermon, 2016) and Maximum-likelihood IRL (ML-IRL)(Ratia et al., 2012; Zeng et al., 2022), all using a centralized learning approach. However, demonstration data in practice are often distributed across decentralized clients, e.g., devices, cars, and households. It is not realistic to assume that such sensitive data can always be shared or collected for centralized inverse learning, due to privacy restrictions.

To enable collaborative training of machine learning models among decentralized clients under the privacy restrictions, Federated Learning (FL) provides a promising solution by maintaining training data on local devices and aggregating local updates to build a global model. However, most existing work on FL consider only the forward learning problem, e.g., loss minimization (Li et al., 2019), policy improvement (Jin et al., 2022), learning with heterogeneous models (Zhou et al., 2024), and efficient optimization methods (Li et al., 2020; 2021b;a; Wang et al., 2020), and not the IRL problem. We note that IRL using decentralized clients and distributed data is an open problem. It often has a bi-level structure of maximizing the probability of observing expert trajectories under the current policy and optimizing discounted cumulative reward for the target reward function, which must be solved jointly during IRL. A naive integration of FL and IRL may not achieve convergence. A decentralized learning framework for IRL with theoretical analysis of the convergence and the time-complexity remains a significant challenge.

The goal of this paper is to develop a novel framework for federated maximum-likelihood IRL (F-ML-IRL) and provide a rigorous convergence/time-complexity analysis of the proposed algorithms. We adopt the Maximum Likelihood IRL (ML-IRL) approach in (Zeng et al., 2022) and consider the problem of decentralized IRL of a shared latent reward function, from distributed data and using decentralized client devices. Our solution attains the privacy-preserving benefits of FL in IRL. To address the bi-level nature of IRL, our proposed algorithm's local training round (McMahan et al., 2017) encompasses an upper-level learning task (on each client with local dataset) to optimize the parameterized reward function to maximize the discounted likelihood of observing expert trajectories under the current policy, as well as a low-level learning task to find the optimal policy concerning the entropy-regularized discounted cumulative reward for the current reward function. Then, we design a dual-aggregation method for aggregating both the action-value networks and reward function models every T local rounds, rather than just one set of model parameters in standard FL. Further, we leverage Soft Q-learning (Haarnoja et al., 2017) as the base RL algorithm. Instead of fully solving the forward RL problem before updating the reward parameter, we perform one-step updates for both the recovered policy and reward parameter alternately to improve the efficiency. To the best of our knowledge, this is the first proposal to formulate and solve this F-ML-IRL problem.

We conduct a rigorous convergence/time-complexity analysis of the proposed F-ML-IRL algorithm. Due to the tight coupling between the reward parameters and the recovered policy in IRL's bi-level optimization, the dual-aggregation method in our F-ML-IRL must be analyzed to understand its impact on convergence. By bounding the logarithmic distance between the estimated policy and the optimal policy by the distance between their corresponding Q-values, we control the variance introduced by local training by considering the time immediately after each global aggregation. Utilizing the γ-contraction property of soft Q-values, we establish the contraction property of the targeted distance, which allows us to provide a convergence proof for the policy estimate. Moreover, we leverage the Lipschitz continuity of the reward parameter and the convergence of the policy estimate to show that the gradient of the global reward parameter converges to zero as the number of communications increases. These techniques enable us to show that F-ML-IRL's policy estimation and reward optimization both converge in finite time. The change in convergence speed due to the use of only decentralized clients and distributed data (rather than centralized learning) is characterized.

Our F-ML-IRL is implemented and evaluated on high-dimensional robotic control tasks in MuJoCo (Todorov et al., 2012). We compared its performance with several centralized learning baseline including Behavior Cloning (BC) (Pomerleau, 1988), Generative Adversarial Imitation Learning (GAIL) (Ho & Ermon, 2016), and IRL methods like f-IRL (Ni et al., 2021) and ML-IRL (Zeng et al., 2022). We consider non-iid data distribution, where clients have different local demonstration data with varying performance levels. The baselines are evaluated using centralized data with two setups (i) a single client with medium-level demonstrations and (ii) a single client with a mixture of demonstrations of different levels. The results show that our F-ML-IRL could effectively leverage distributed data and client devices in learning, to achieve similar or better recovered reward than the baselines, while meeting decentralization and data privacy restrictions. Our evaluation code is available at `https://anonymous.4open.science/r/F-ML-IRL/`. The key contributions of this paper are summarized as follows:

- We propose a novel framework for federated maximum-likelihood IRL (F-ML-IRL). It enables decentralized IRL of a shared latent reward function, from distributed data and using decentralized client devices, under data privacy restrictions.
- To support the bi-level optimization structure in IRL – for jointly updating the optimal policy and the reward function estimate, the proposed F-ML-IRL algorithm leverages a dual-aggregation of the model parameters, which ensures convergence to optimal results.
- The convergence and time-complexity of the proposed F-ML-IRL algorithm is quantified, with respect to local rounds T and aggregation steps M. We show that F-ML-IRL achieves convergence in finite time and will have faster convergence with a smaller local rounds T.
- Our solution is evaluated on high-dimensional robotic control tasks in MuJoCo and is shown to achieve similar or higher recovered reward than a number of Imitation Learning and IRL baselines that employ centralized learning.

2 Related Work and Background

IRL aims to learn the reward function using expert demonstration data, which frees the forward RL problem from the requirement of specifying the reward function beforehand (Ng et al., 2000) and also facilitates imitation learning by using the recovered reward function to derive an effective policy

(Abbeel & Ng, 2004). Various formulations and solutions for the IRL problem have been explored. The Maximum Margin Planning algorithm frames the problem within a quadratic programming context Ratliff et al. (2006). Bayesian IRL models infer the posterior distribution of the reward function given a prior (Ramachandran & Amir, 2007). Probabilistic maximum entropy IRL methods favor stochastic policies using entropy regularization. In recent years, Generative Adversarial Imitation Learning (GAIL) (Ho & Ermon, 2016) has adopted a Generative Adversarial Networks (GANs) (Goodfellow et al., 2020) framework to recover the expert's policy. In this framework, a generator proposes new policies to confuse the discriminator, while the discriminator determines whether the state-action pair from the generator's policy originates from the expert. However, existing work has not considered the IRL problem with distributed data and decentralized clients, under data privacy.

ML-IRL models the policy through an MDP and recover the latent reward function based on maximum likelihood principle. The convergence of centralized ML-IRL with a single client has been analyzed (Ratia et al., 2012; Zeng et al., 2022) and is shown to outperform other IRL methods. ML-IRL considers a MDP defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \eta, r, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ represent the state space and action space, respectively. $\mathcal{P}(s'|s,a)$ denotes the state transition probability, $\eta(\cdot)$ is the initial state distribution, $r(s,a)$ is the reward function, and γ is the discount factor. Let θ denote the parameter vector for the reward function, making the reward function $r(s,a;\theta)$. The IRL problem states that the expert's behavior is characterized by a stochastic policy $\pi_{r_\theta}(\cdot|s)$. The dataset $\mathcal{D} := \{\tau_m\}_{m=1}^K$ contains trajectories $\tau_m = \{(s_t, a_t)\}_{t=0}^\infty$ from the expert policy $\pi_{r_\theta}(\cdot|s)$.

The discounted log-likelihood of observing all sample trajectories $\mathcal{D}$ from the expert is given by:

$$\mathbb{E}_{\tau\sim\mathcal{D}}\left[\sum_{t\geq 0}\gamma^t\left(\log\pi_{r_\theta}(a_t|s_t) + \log\mathcal{P}(s_{t+1} \mid s_t, a_t)\right)\right]. \tag{1}$$

Assume the state transition probabilities $\mathcal{P}(s_{t+1}|s_t, a_t)$ are known. Then, maximizing the discounted log-likelihood is equivalent to maximizing equation 2.

$$l(\theta) = \mathbb{E}_{\tau\sim\mathcal{D}}\left[\sum_{t\geq 0}\gamma^t\log\pi_{r_\theta}(a_t|s_t)\right]. \tag{2}$$

ML-IRL aims to maximizing $l(\theta)$ under the constraint that π_{r_θ} is the optimal policy targeting the discounted cumulative reward regularized by the entropy of the policy, i.e. $\pi_{r_\theta} := \arg\max_\pi \mathbb{E}_\pi[\sum_{t=0}^\infty \gamma^t(r(s_t,a_t;\theta) + \mathcal{H}(\pi(\cdot \mid s_t))]$, where the entropy of the policy is defined as $\mathcal{H}(\pi(\cdot|s)) := -\sum_{a\in\mathcal{A}}\pi(a|s)\log\pi(a|s)$. Incorporating the policy entropy term as a regularization makes the IRL problem well-defined. This adjustment encourages the agent to explore all possible trajectories in the environment, leading to a more stochastic policy with better generalization capabilities.

For decentralized learning, FL focuses on scenarios where multiple clients work together to train a model using distributed data. FL considers the objective of the form:

$$\min_{w\in\mathbb{R}^d} f(w) \quad \text{where} \quad f(w) = \frac{1}{n}\sum_{i=1}^n f_i(w) \tag{3}$$

We assume there are n clients over which the local data D_i is stored. Prior to federated averaging (FedAvg), most works in FL based on Stochastic Gradient Descent (FedSGD) (Shokri & Shmatikov, 2015) ignored the impact of data heterogeneity and imbalance. FedAvg derives from FedSGD but allows multiple rounds of local update $\omega^i \leftarrow \omega^i - \alpha\nabla f_i(\omega^i)$ by gradient descent before aggregating the model parameters at the central server, reducing the frequency and cost of communications. The convergence of FedAvg on non-i.i.d. data has been proved (Li et al., 2019). Since Fed-Avg was proposed as the vanilla FL algorithm, efficient federated optimization methods like FedProx (Li et al., 2020) FedBN (Li et al., 2021b), MOON (Li et al., 2021a), and FedNova (Wang et al., 2020) have been developed to address non-i.i.d. data and accelerate the model training process (Konečnỳ et al., 2016). Additionally, the convergence of model-heterogeneous FL, where reduced-size models are extracted from the global model and applied to low-end clients, was provided in (Zhou et al., 2024). However, existing FL methods could not be directly applied to the ML-IRL problem with decentralized clients, since ML-IRL requires a bi-level optimization involving both policy improvement and reward estimate using maximum likelihood. New algorithms needs to be developed for decentralized ML-IRL with rigorous convergence/time-complexity analysis.

3 FEDERATED MAXIMUM-LIKELIHOOD IRL

3.1 OUR PROBLEM STATEMENT

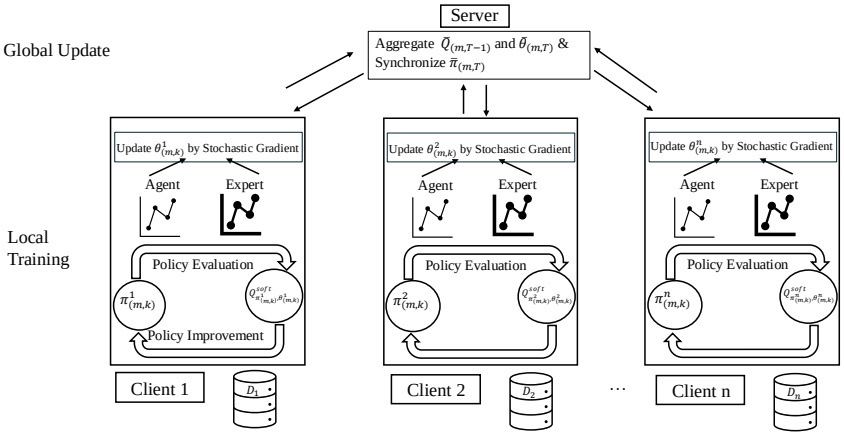


Figure 1: Our F-ML-IRL problem. It aims to recover reward function r_θ from sensitive data/demonstrations $\mathcal{D}_1, \ldots, \mathcal{D}_n$ that are distributed over n clients. This requires a novel decentralized algorithm to solve a bi-level optimization – optimizing the parameterized reward function with maximum likelihood and optimizing the corresponding policy concerning the entropy-regularized discounted cumulative reward. We prove the convergence and the time-complexity of F-ML-IRL.

We consider a decentralized inverse learning problem to recover a common reward function r_θ from distributed datasets spread across n clients. Due to privacy requirements, the clients cannot directly share their data for learning. Specifically, we consider n clients, each with a dataset $\mathcal{D}_i := \{\tau_m^i\}_{m=1}^K$ containing trajectories $\tau_m^i = \{(s_t, a_t)\}_{t=0}^\infty$ from the i-th expert policy $\pi_{r_\theta}^i(\cdot|s)$. Different from centralized learning, the clients each have their local model trained on local data. By modeling the distributed expert trajectories as an MDP $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \eta, \gamma)$, our goal is to learn a common reward function r_θ – parameterized by θ – from distributed data and to recover the corresponding optimal policy π_{r_θ}. The F-ML-IRL in this paper is formulated as follows:

$$
\begin{aligned}
&\max_{\theta \in \mathbb{R}^d} \quad L(\theta) = \frac{1}{n} \sum_{i=1}^{n} l_i(\theta) \\
&\text{s.t.} \quad \pi_{r_\theta} = \arg\max_{\pi} \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \left(r_\theta(s_t, a_t) + \mathcal{H}(\pi(\cdot \mid s_t)) \right) \right] \\
&\text{where} \qquad l_i(\theta) = \mathbb{E}_{\tau \sim \mathcal{D}_i} \left[\sum_{t \geq 0} \gamma^t \log \pi_{r_\theta}(a_t|s_t) \right]
\end{aligned}
\tag{4}
$$

where $l_i(\theta)$ is the local likelihood calculated using client i's local data $\mathcal{D}_i$ and target policy π_{r_θ}, which further depends on the current reward function r_θ that is shared by all clients, making it a difficult bi-level optimization. We cannot directly apply FL to this problem, because the maximum likelihood problem on $L(\theta)$ depends on the recovered policy π_{r_θ}, while the policy search for an optimal π_{r_θ} futher relies on the estimation of the reward function parameter θ. Thus, the two-level optimization are entangled with each other and requires a new aggregation strategy in F-ML-IRL.

3.2 OUR PROPOSED F-ML-IRL ALGORITHM

We present F-ML-IRL algorithm to solve the decentralized inverse learning problem. Our proposed solution includes three modules - local policy improvement, local reward optimization, and global bi-level aggregation. Each round of F-ML-IRL algorithm consists of T local client steps running in parallel and a global server aggregation of selected model parameters at the end of each round. At each local step, each client i first executes (in parallel) a policy update (on local data D_i) through policy evaluation and improvement steps based on soft-Q learning to address the lower-level problem. Second, each client carries out a reward optimization, where the reward parameter gradient update is derived by contrasting sampled trajectories from both the expert policy and the current

policy estimate. Next, after every T local steps and at the end of round m, we perform a dual aggregation of both the action-value function and the reward parameters, i.e., to synchronize the local bi-level optimization of both policy and reward on decentralized clients.. While our solution is inspired by FL, F-ML-IRL performs a dual aggregation with respect to the bi-level optimization in ML-IRL. The algorithm details are presented below. Its convergence and time-complexity are rigorously analyzed in this paper.

Our F-ML-IRL is illustrated in Fig. 1. Different expert demonstration data D_i are stored at different client devices. We perform local training for policy evaluation and improvement based on soft Q-learning to improve the local policy $\pi^i_{(m,k)}$ under current reward parameter $\theta^i_{(m,k)}$. We then sample trajectories from the current local policy and the expert demonstration data D_i, to provide an update for the reward parameter $\theta^i_{(m,k)}$. At local step k of round m, we use $Q^{\text{soft}}_{\pi^i_{(m,k)},\theta^i_{(m,k)}}(s,a)$ to denote the action-value function (i.e., Q-value) for action a and state s, with respect to the current policy estimation $\pi^i_{(m,k)}$ under current reward parameter estimation $\theta^i_{(m,k)}$, on each client i. After every T steps of local training, we perform dual aggregation of the Q-values $\overline{Q}^{\text{soft}}_{(m,T-1)}$ and the reward parameter $\overline{\pi}_{(m,T)}$. To the best of our knowledge, this is the first paper considering an ML-IRL problem in this FL context.

Local training for policy improvement. Iterations of local training on each local client start with a shared model with parameters $\pi^i_{(m,0)}(\cdot|s)$ and $\theta^i_{(m,0)}$. During each local training round, we first evaluate the local policy $\pi^i_{(m,k)}(\cdot|s)$ by computing the Q-values $Q^i_{(m,k)}(\cdot,\cdot)$ under the fixed reward parameter θ^i for the i-th local client using the definitions of the soft value and soft Q functions in equation 5.

$$
\begin{aligned}
V^{\text{soft}}_{\pi^i_{(m,k)},\theta^i_{(m,k)}}(s) &= \mathbb{E}_{\pi^i_{(m,k)},s_0=s}\sum_{t=0}^{\infty}\gamma^t\left(r(s_t,a_t;\theta^i_{(m,k)})+\mathcal{H}(\pi_{\theta^i_{(m,k)}}(\cdot|s_t))\right)\\
Q^{\text{soft}}_{\pi^i_{(m,k)},\theta^i_{(m,k)}}(s,a) &= r(s,a;\theta^i_{(m,k)})+\gamma\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)}V^{\text{soft}}_{\pi^i_{(m,k)},\theta^i_{(m,k)}}(s')
\end{aligned}
\tag{5}
$$

Then, $\pi^i_{(m,k+1)}(\cdot|s)$ is updated according to the policy improvement step using soft Q-learning in equation 6. It does not assume an explicit policy function but uses the Boltzmann distribution of the Q function, making the probability of choosing an action at some state s proportional to the exponential of the Q-value of this action-state pair.

$$
\pi^i_{(m,k+1)}(a|s)\propto\exp(Q^{\text{soft}}_{\pi^i_{(m,k)},\theta^i_{(m,k)}}(s,a)),\forall s \tag{6}
$$

Local training for reward optimization. For the optimization towards the local reward parameter $\theta^i_{(m,k+1)}$, a stochastic gradient ascent method is proposed. The gradient of each local likelihood function $l_i(\theta)$ is given by equation 7, which derives from Lemma 1 in (Zeng et al., 2022).

$$
\nabla l_i(\theta)=\mathbb{E}_{\tau_i\sim\mathcal{D}_i}\sum_{t\geq 0}\gamma^t\nabla_\theta r(s_t,a_t;\theta)-\mathbb{E}_{\tau_i\sim\pi_\theta}\sum_{t\geq 0}\gamma^t\nabla_\theta r(s_t,a_t;\theta). \tag{7}
$$

We construct a stochastic estimator of the exact gradient $\nabla l_i(\theta^i_{(m,k)})$, approximating the optimal policy $\pi_{r_{\theta^i_{(m,k)}}}$ with the current policy $\pi^i_{(m,k+1)}$. Specifically, we sample one expert trajectory $\tau^{E_i}_{(m,k)}:=\{s_t,a_t\}_{t\geq 0}$ from the local dataset $\mathcal{D}_i$ and one agent trajectory $\tau^{A_i}_{(m,k)}:=\{s_t,a_t\}_{t\geq 0}$ from the current policy $\pi^i_{(m,k+1)}$. Then we use a stochastic estimate $g^i_{(m,k)}$ to approximate the exact gradient of the local likelihood objective function l_i for each local client in equation 8. The update of the reward relies on both the local softmax policy $\pi^i_{(m,k+1)}$ through $\tau^{A_i}_{(m,k)}$ and the local data $\mathcal{D}_i$ through $\tau^{E_i}_{(m,k)}$.

$$
g^i_{(m,k)}=h(\theta^i_{(m,k)};\tau^{E_i}_{(m,k)})-h(\theta_{(m,k)};\tau^{A_i}_{(m,k)}) \tag{8}
$$

where $h(\theta;\tau)=\sum_{t\geq 0}\gamma^t\nabla_\theta r(s_t,a_t;\theta)$. Finally, the local reward parameter $\theta^i_{(m,k)}$ is updated as:

$$
\theta^i_{(m,k+1)}=\theta^i_{(m,k)}+\alpha g^i_{(m,k)} \tag{9}
$$

where α is the learning rate of the reward parameter update.

Bi-level model aggregation. Every T local iterations, local Q-values and local reward parameters are communicated to the global server for aggregation, while the policy synchronization is performed based on the aggregated Q-values such that each local client has the same policy after the aggregation. We design the dual aggregation step after thorough thoughts. The reward update in equation 9 depends on how well the trajectories from policy $\pi^i_{(m,k)}$ approximates the optimal policy $\pi_{r_{\theta^i_{(m,k)}}}$, while the policy $\pi^i_{(m,k)}$ relies on the Q-value update from equation 5. Therefore, our FL algorithm aims to improve the Q-value estimates for local clients by aggregating their Q-values equation 10.

$$\overline{Q}^{\text{soft}}_{(m,T-1)}(\cdot,\cdot) := \sum_{j=1}^{N} Q^j_{(m,T-2)}(\cdot,\cdot)/N \tag{10}$$

We note that when the Q-values are represented by another network with parameter ψ, the aggregation of the Q-values will simply become aggregation of model parameters. The policy synchronization is automatically performed by policy improvement based on the aggregated Q-values and sent to each local client for update such that each local client has the same policy after the Q aggregation in equation 11:

$$\overline{\pi}_{(m,T)}(\cdot|s) \propto \exp(\overline{Q}^{\text{soft}}_{(m,T-1)}(s,\cdot)), \forall s \in S \tag{11}$$

Since ML-IRL requires a bi-level problem with respect to both the reward parameter and the recovered policy, we consider a dual aggregation that also applies to the reward parameter θ:

$$\overline{\theta}_{(m,T)} := \sum_{j=1}^{N} \theta^j_{(m,T-1)}/N \tag{12}$$

After each dual aggregation, the global policy and reward parameters are sent to each local clients as an initialization for future local training: $\pi^i_{(m,0)}(\cdot|s) = \overline{\pi}_{(m-1,T)}(\cdot|s)$ and $\theta^i_{(m,0)} = \overline{\theta}_{(m-1,T)}$ for all $i = 1, 2, \ldots, N$. The entire process of the F-ML-IRL algorithm is summarized in Algorithm 1.

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

Input: Initialize reward parameter $\theta^i_{(0,0)}$ and policy $\pi^i_{(0,0)}$. Set the aggregation period to be T, number of local server to be N, and reward parameter's local stepsize as α.
for $m = 0, 1, \ldots, M-1$ **do**
  if $m > 0$ **then**
    Inherit $\pi^i_{(m,0)}(\cdot|s)$ and $\theta^i_{(m,0)}$ from last aggregation
  end if
  for $k = 0, \ldots, T-2$ **do**
    for $i = 1, 2, \ldots, N$ **do**
      Compute $Q^{\text{soft}}_{r_{\theta^i_{(m,k)}}, \pi^i_{(m,k)}}(\cdot,\cdot)$ using equation 5
      Update $\pi^i_{(m,k+1)}(\cdot|s)$ based on equation 6
      Sample an expert trajectory $\tau^{E_i}_{(m,k)}$ from local dataset D_i
      Sample a trajectory $\tau^{A_i}_{(m,k)}$ from current policy $\pi^i_{(m,k+1)}$
      Estimating gradient $g^i_{(m,k)}$ following equation 8
      Update reward parameter $\theta^i_{(m,k+1)}$ using equation 9
    end for
  end for
  Set $k = T-1$
    Aggregate $\overline{Q}^{\text{soft}}_{(m,k)}(\cdot,\cdot)$ by equation 10
    Synchronize policies $\overline{\pi}_{(m,k+1)}(\cdot|s)$ using equation 11
    Aggregate reward parameters $\overline{\theta}_{(m,k+1)}$ by equation 12
end for

4 THEORETICAL ANALYSIS

4.1 ASSUMPTIONS

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

$$\sup_{s \in S} \|\mathbb{P}(s_t \in \cdot \mid s_0 = s, \pi) - \mu_\pi(\cdot)\|_{TV} \leq \kappa \rho^t, \quad \forall t \geq 0 \tag{13}$$

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

Equation 13 states that the Markov chain mixes at a geometric rate. This is a common assumption in the RL literature, which holds for any time-homogeneous Markov chain with a finite state space.

Bounded Gradient and Lipschitz Property. For any $s \in \mathcal{S}$, $a \in \mathcal{A}$, and any reward parameter θ, the following conditions hold, where L_r and L_g are positive constants:

$$\|\nabla_\theta r(s, a; \theta)\| \leq L_r, \text{ and } \|\nabla_\theta r(s, a; \theta_1) - \nabla_\theta r(s, a; \theta_2)\| \leq L_g \|\theta_1 - \theta_2\|, \tag{14}$$

Equation 14 posits that the parameterized reward function has a bounded gradient and is Lipschitz smooth, which is common in the literature.

4.2 IMPORTANT LEMMAS

We first introduce two important lemmas that are used repeatedly in the converge analysis. Due to space limitations, the proofs of these lemmas are included in the appendix.

Lemma 1. *Suppose the above assumptions hold. Given any reward parameters θ_1 and θ_2, the following results hold for any $s \in \mathcal{S}$ and $a \in \mathcal{A}$:*

$$\left|Q^{soft}_{r_{\theta_1}, \pi_{\theta_1}}(s, a) - Q^{soft}_{r_{\theta_2}, \pi_{\theta_2}}(s, a)\right| \leq L_q \|\theta_1 - \theta_2\|, \tag{15}$$

$$\|\nabla L(\theta_1) - \nabla L(\theta_2)\| \leq L_c \|\theta_1 - \theta_2\|, \tag{16}$$

where $Q^{soft}_{r_\theta, \pi_\theta}(\cdot, \cdot)$ denotes the soft Q-function under the reward function $r(\cdot, \cdot; \theta)$ and the policy π_θ.

Lemma 1 is directly derived from Lemma 2 in (Zeng et al., 2022), where the positive constants L_q and L_c are also defined. The Lipschitz properties of the Q-value function and the gradient of the log-likelihood are essential for convergence analysis, as they help control the distance between local and global models in the FL setting.

Lemma 2. *For any two policies $\pi(a|s)$ and $\pi'(a|s)$, the difference in their soft Q-values under some reward function r for a given state-action pair (s, a) is bounded as follows:*

$$\|Q^{soft}_{r_\theta, \pi} - Q^{soft}_{r_\theta, \pi'}\|_\infty \leq \frac{\gamma}{1 - \gamma} \|\log(\pi) - \log(\pi')\|_\infty \tag{17}$$

Controlling the distance between soft Q-values under different policies helps us analyze the optimality of the global policy with respect to the global reward parameter after aggregations.

4.3 MAIN CONVERGENCE RESULT

Theorem 1. *Under the above two assumptions, if we choose step size $\alpha_{(m,k)} = \alpha_0/(mT + k)^\sigma$ in F-ML-IRL (Algorithm 1), where $\alpha_0 > 0$ and $\sigma \in (0, 1)$ are constants and M is the total number of dual aggregations, the following convergence results hold for F-ML-IRL:*

$$\frac{1}{M} \sum_{m=0}^{M-1} \mathbb{E}\left[\left\|\log(\overline{\pi}_{(m,T)}) - \log(\pi_{\theta^i_{(m,T-1)}})\right\|_\infty\right] = \mathcal{O}\left(M^{-1}\gamma^{T-1}\right) + \mathcal{O}\left(M^{-\sigma}T^{1-\sigma}\right), \tag{18}$$

$$\frac{1}{M} \sum_{m=0}^{M-1} \mathbb{E}\left[\left\|\nabla L(\overline{\theta}_{(m,T)})\right\|^2\right] = \mathcal{O}(M^{-1}) + \mathcal{O}(M^{-\sigma}T^{-\sigma}) + \mathcal{O}(M^{-1-\sigma}T^{1-\sigma}). \tag{19}$$

Remarks: The time complexity of both policy estimate and reward parameter optimization depends on the number of global aggregation rounds M and the number of local training steps T. The policy and reward function parameters in F-ML-IRL converge at the rate of $M^{-\sigma}T^{1-\sigma}$ and $M^{-\sigma}T^{-\sigma}$, respectively, since we have $\sigma \in (0,1)$ and T is often fixed. We note that due to dual-aggregation and the variance caused by local training on distributed datasets across decentralized clients, F-ML-IRL exhibits a slightly slower convergence rate, compared with standard centralized ML-IRL with a single client (whose convergence rate is $M^{-\sigma}$). From Equations (18) and (19), there exists a sweet spot with respect to the number of local training steps T, since γ^{T-1} and $T^{-\sigma}$ both descreases with T, while $T^{1-\sigma}$ increases. Exploring this tradeoff will be considered in future work. Compared with Fed-Avg (whose convergence rate is $M^{-1}T^{-1}$), F-ML-IRL also has a slower convergence rate due to the complexity of the bi-level optimization problem.

Proof Sketch: Due to space limitations, we outline the key steps of our convergence analysis and present the complete proof in the appendix. We first analyze the convergence of policy estimates and reduce it to the convergence of Q-values. We then analyze the distance between Q-values using the Lipschitz property, tracing back to the start of each dual aggregation around. In particular, we examine the extra distance between the estimated policy and the optimal policy caused by aggregation, seeking the contraction property of Q-value estimates between adjacent aggregation rounds. Next, for reward optimization, we leverage the Lipschitz smooth property of the likelihood and control the discrepancy between the stochastic gradient and the true gradient. This allows us to use the convergence of Q-values from the previous analysis to demonstrate the gradient convergence of the reward parameter. For simplicity of notations, we use $Q^{\text{soft}}_{i,(m,t)}$ to denote $Q^{\text{soft}}_{r_{\theta^i_{(m,t)}},\pi^i_{(m,t)}}$, the action-value function at a given state for the local policy and reward parameter estimations at round (m,t). Similarly, $Q^{\text{soft}}_{i,(m,t)}*$ denotes $Q^{\text{soft}}_{r_{\theta^i_{(m,t)}},\pi^i_{\theta_{(m,t)}}}$, which is the Q-function for the optimal policy under the reward parameter at round (m,t) and $Q^{\text{soft}}_{\overline{(m,t)}}$ denotes $Q^{\text{soft}}_{r_{\overline{\theta}^i_{(m,T)}},\pi_{\overline{\theta}^i_{(m,T)}}}$, which represents the Q-function for the aggregated policy and reward parameter at the m'th aggregation.

Convergence of Policy Estimate:

Step 1: We show the distance between the aggregated policy and the optimal policy under the pre-aggregation local reward parameter could be controlled using the distance between soft Q-functions:

$$\|\log(\overline{\pi}_{(m,T)}) - \log(\pi_{\theta^i_{(m,T-1)}})\|_\infty \leq 2\|\overline{Q}^{\text{soft}}_{(m,T-1)} - Q^{\text{soft}}_{i,(m,T-2)} * \|_\infty$$

This step relies on the policy update rule in equation 6.

Step 2: By introducing intermediary terms as bridges, specifically looking back to the time right after the last aggregation, where all local servers have the same reward parameter $\overline{\theta}_{(m-1,T)}$, we further bound the difference in **Step 1** by converting it to the difference of reward parameters using equation 9, 10, 15. Combining this with the γ-contraction property of the soft-Q update, we have:

$$\|\overline{Q}^{\text{soft}}_{(m,T-1)} - Q^{\text{soft}}_{i,(m,T-2)} * \|_\infty \leq \gamma^{T-2}\|Q^{\text{soft}}_{r_{\overline{\theta}_{(m-1,T)}},\overline{\pi}_{(m-1,T)}} - Q^{\text{soft}}_{r_{\overline{\theta}_{(m-1,T)}},\pi_{\overline{\theta}_{(m-1,T)}}}\|_\infty + E_1$$

where we use auxiliary variable $E_1 = 4\alpha\left(\frac{1-\gamma^{T-2}}{1-\gamma} + T - 2\right)L_q^2$.

Step 3: Using Lemma 2, we bound the difference in Q-values corresponding to different policies with the same reward during the aggregation step, and finally have:

$$\|\overline{Q}^{\text{soft}}_{(m,T-1)} - Q^{\text{soft}}_{i,(m,T-2)} * \|_\infty \leq (1-\gamma)\gamma^{T-}\|\overline{Q}^{\text{soft}}_{(m-1,T-1)} - Q^{\text{soft}}_{i,(m-1,T-2)} * \|_\infty + E_2$$

where we use auxiliary variable $E_2 = 2\frac{1-\gamma^{T-2}}{1-\gamma} + (1-\gamma)^2\gamma^{T-2} + \frac{1-\gamma}{\gamma}(2T-3) + 2(T-2)L_q^2$. Finally, we obtain the convergence rate of the policy estimate by the contraction of Q-difference.

Convergence of Reward Parameter Optimization:

Step 1: We first leverage the Lipschitz smooth property of $l(\theta)$ equation 16:

$$L(\overline{\theta}_{(m,T))}) \geq L(\overline{\theta}_{(m-1,T)}) + \left\langle \nabla L(\overline{\theta}_{(m,T)}), \overline{\theta}_{(m,T)} - \overline{\theta}_{(m-1,T)} \right\rangle - \frac{L_c}{2}\|\overline{\theta}_{(m,T)} - \overline{\theta}_{(m-1,T)}\|^2$$

Step 2: We show the bias between the stochastic gradient estimate $g^i_{(m,k)}$ and the true gradient $\nabla L(\theta_{(m-1,T)})$ could be controlled. In this process, we also compare the increments of local clients

to control the extra error terms introduced by the federated scheme leveraging equation 9, 12. We show that the gradient of the global reward parameter could be bounded by the distance between Q-values:

$$\alpha(T-1)\mathbb{E}[\|\nabla L(\overline{\theta}_{(m-1,T)})\|^2] \leq \alpha C_1 \mathbb{E}[\|\overline{Q}^{\text{soft}}_{(m-1,T-1)} - Q^{\text{soft}}_{i,(m-1,T-2)}\|_\infty] + \mathbb{E}[L(\overline{\theta}_{(m,T)}) - L(\overline{\theta}_{(m-1,T)})] + E_3 \tag{20}$$

where $C_1 = \frac{4(1-\gamma^{T-1})}{\gamma} L_q^2 C_d \sqrt{|\mathcal{S}| \cdot |\mathcal{A}|}$ and $E_3 = 8\alpha L_q^3 C_d \sqrt{|\mathcal{S}| \cdot |\mathcal{A}|} \cdot \frac{T-1-\frac{1-\gamma^{T-1}}{1-\gamma}}{1-\gamma} + \frac{(T-1)(3T-1)\alpha^2 L_c L_q^2}{2} + \frac{4(1-\gamma^{T-1})}{1-\gamma} \alpha L_q^2 C_d \sqrt{|\mathcal{S}| \cdot |\mathcal{A}|} \cdot [2(2T-3)\alpha L_q^2 + \frac{1-\gamma}{\gamma} \cdot 4\alpha L_q^2]$ are two auxiliary variables. By combining this with the convergence of the Q-value difference that was established in **Step 3** of the Policy Estimation proof, we obtain the desired convergence of the reward parameter.

5 EVALUATIONS

We evaluated the proposed F-ML-IRL method on five high-dimensional robotic control tasks in MuJoCo (Todorov et al., 2012). For comparison, we selected several state-of-the-art baselines, including imitation learning approaches that only learn the expert policy—specifically like BC (Pomerleau, 1988) and GAIL (Ho & Ermon, 2016), as well as IRL methods that simultaneously learn both a reward function and a policy, namely f-IRL (Ni et al., 2021) and ML-IRL (Zeng et al., 2022). To ensure fairness, we used Soft Actor-Critic (SAC) (Haarnoja et al., 2018) as the base RL algorithm for all methods since it incorporates elements of Soft Q-Learning and achieves strong performance using the actor-critic scheme. The experiments are conducted on a server with AMD EPYC 7513 32-Core Processors and NVIDIA RTX A6000 GPUs. We choose $M = 200$ rounds and $T = 5$ local steps and average the results over multiple runs. Our evaluation code is available at `https://anonymous.4open.science/r/F-ML-IRL/`.

Environment	Setting	F-ML-IRL	ML-IRL		BC		GAIL		f-IRL	
			Mixed	Medium	Mixed	Medium	Mixed	Medium	Mixed	Medium
Ant	(3, 200)	**6425.91**	6219.78	6161.65	983.99	984.04	989.30	988.73	5615.33	5930.89
	(3, 1000)	**6398.98**	5100.25	6402.87	5952.08	718.87	989.00	989.29	5370.28	5527.63
	(5, 200)	**6254.32**	5614.91	6161.65	983.51	984.04	988.67	988.73	5628.94	5930.89
	(5, 1000)	**6528.04**	6330.67	6402.87	411.83	718.87	989.77	989.29	5388.74	5527.63
HalfCheetah	(3, 200)	**13007.75**	8054.94	12581.28	-0.63	-0.73	7513.31	10288.42	10110.73	12962.52
	(3, 1000)	13228.98	**13642.82**	13124.24	-0.66	-11.74	12112.99	11506.59	13075.95	12871.64
	(5, 200)	11827.91	6406.45	12581.28	-0.57	-0.73	4910.45	10288.42	7132.01	**12962.52**
	(5, 1000)	12360.60	12750.04	**13124.24**	110.59	-11.74	11364.40	11506.59	12659.30	12871.64
Hopper	(3, 200)	3576.10	1871.83	**3623.07**	18.11	18.13	1022.90	1023.65	1297.25	3456.47
	(3, 1000)	**3674.64**	3518.04	3479.33	18.17	2290.58	1025.93	1032.07	3403.36	3390.08
	(5, 200)	3419.95	1484.52	**3623.07**	18.12	18.13	1020.19	1023.65	1313.12	3456.47
	(5, 1000)	**3618.44**	3601.40	3479.33	1016.31	2290.58	1111.08	1032.07	3468.72	3390.08
Humanoid	(3, 200)	5656.06	5484.99	5861.01	243.21	242.50	4666.38	3035.34	5510.06	**6004.58**
	(3, 1000)	5694.79	**5903.42**	5813.57	241.46	532.41	4627.15	4688.41	5708.21	5726.86
	(5, 200)	**6232.37**	5462.64	5861.01	242.69	242.50	4692.86	3035.34	5523.40	6004.58
	(5, 1000)	**6294.25**	5713.37	5813.57	545.01	532.41	4577.12	4688.41	5608.83	5726.86
Walker2d	(3, 200)	4057.25	3317.61	4400.43	8.38	8.27	353.66	18.74	1050.78	**5729.55**
	(3, 1000)	**5798.37**	5061.23	5673.49	8.27	507.69	344.55	19.27	4805.53	5255.57
	(5, 200)	4540.90	3024.14	4400.43	8.40	8.27	13.03	18.74	1115.37	**5729.55**
	(5, 1000)	**5853.42**	4669.73	5673.49	711.06	507.69	360.10	19.27	4704.77	5255.57
Average	–	**6712.60**	5661.64	6712.09	575.97	529.00	3183.64	3359.05	5424.58	6685.58

Table 1: Compare F-ML-IRL and baselines on MuJoCo tasks, with different number of clients and demonstration trajectory length. F-ML-IRL achieves similar or higher recovered reward in almost all scenarios and outperforms the baselines in more than half, as well as in terms of the average.

We evaluate different algorithms using the rewards associated with the recovered expert policies evaluated in the original environment (same as the method adopted in previous work). We compare F-ML-IRL with the baselines on five MuJoCo tasks under non-iid data distributions, where each

client contains different demonstration data corresponding to varying levels of expertise. For the baselines that rely on centralized learning, we consider two setups: (i) a single client with medium-level demonstrations, denoted as *medium* and (ii) a single client with a mixture of demonstrations of different levels, denoted as *mixed*. In either case, the total amount of local data per client remains the same in the experiments. More details on experiment set up is provided in the appendix.

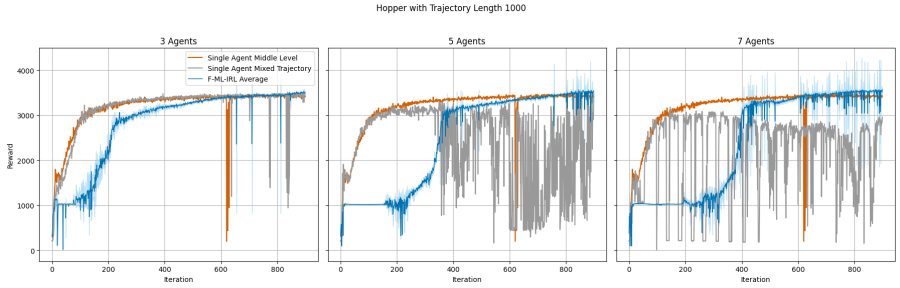


Figure 2: Convergence of F-ML-IRL in Hopper Environment compared with centralized ML-IRL with mixed and medium data. As the number of clients (and thus the non-iid datasets) increases from 3 (left) to 7 (right), F-ML-IRL takes longer to converge and nevertheless achieves more significant improvement by leveraging distributed demonstration data on the clients.

The evaluation results are summarized in Table 1. We have tested each algorithm and each MuJoCo task under 4 settings, i.e., with 3 or 5 clients and with demonstration trajectory length equals to 200 or 1000, respectively. As demonstrated in (Zeng et al., 2022), even a single expert trajectory of length 1000 can lead to a well-recovered policy using ML-IRL. To investigate the performance of our model under conditions of scarce and distributed data, we utilize a single expert trajectory of length 1000 and further reduce its length to 200 in the experiments. In Table 1, we also compute the average reward for each algorithm across all settings and tasks in our experiments.

We note that F-ML-IRL ensures convergences of the recovered reward in decentralized learning and achieves similar or higher recovered reward than the baselines in almost all settings and tasks. It even outperforms centralized baselines in more than half of the settings and tasks, due to its ability to utilize distributed data. The performance of F-ML-IRL is pretty robust as the number of clients increases to 5 and the expert trajectory length reduces to 200. Imitation learning baselines like BC and GAIL generally have lower performance and even fail in some settings. While ML-IRL performs generally well, it fails to recover a satisfactory policy when data is limited or in tasks involving mixed trajectories of different expertise. On the other hand, f-IRL performs relatively well when provided with longer expert trajectories but struggles when demonstration data is limited. In contrast, our F-ML-IRL consistently achieves similar or higher recovered rewards compared to all baselines, particularly maintaining robust performance even when data is limited and involves demonstrations of mixed expertise.

We further illustrate the convergence of our F-ML-IRL algorithm compared with two different centralized learning baselines using ML-IRL (with medium and mixed-data, respectively) in the Hopper environment, as shown in Figure 2. As the number of clients (and thus the number of non-iid local datasets) increases (from 3 clients on the left to 7 clients on the right), it takes F-ML-IRL more rounds to converge, because of the increased variance introduced by local training on more participating clients and datasets. Nevertheless, F-ML-IRL is able to converge to higher recovered reward than both baselines. Centralized ML-IRL suffers with mixed demonstration data of varying expertise. In contrast, as the number of clients and demonstration dataset increases, F-ML-IRL shows more significant improvement by leverage distributed demonstration data on the clients.

6 Conclusions

This paper proposes F-ML-IRL for federated maximum-likelihood inverse reinforcement learning. It enables decentralized learning of a shared latent reward function from distributed datasets and using decentralized clients. F-ML-IRL algorithm leverages a dual-aggregation to update the shared global model and performs bi-level local updates for inverse learning. We analyze the convergence and time-complexity of F-ML-IRL. Evaluation results on MuJoCo tasks how that F-ML-IRL ensures convergences of the recovered reward and achieves similar or higher recovered reward, compared to state-of-the-art baselines using centralized inverse learning. For further work, we plan to investigate further communication reduction and the use of heterogeneous local models in F-ML-IRL.

REFERENCES

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

Saurabh Arora and Prashant Doshi. A survey of inverse reinforcement learning: Challenges, methods and progress, 2020. URL https://arxiv.org/abs/1806.06877.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

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

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. PMLR, 2018.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29, 2016.

Hao Jin, Yang Peng, Wenhao Yang, Shusen Wang, and Zhihua Zhang. Federated reinforcement learning with environment heterogeneity. In *International Conference on Artificial Intelligence and Statistics*, pp. 18–37. PMLR, 2022.

Jakub Konečnỳ, H Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*, 2016.

Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10713–10722, 2021a.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*, 2019.

Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fedbn: Federated learning on non-iid features via local batch normalization. *arXiv preprint arXiv:2102.07623*, 2021b.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

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

Tianwei Ni, Harshit Sikchi, Yufei Wang, Tejus Gupta, Lisa Lee, and Ben Eysenbach. f-irl: Inverse reinforcement learning via state marginal matching. In *Conference on Robot Learning*, pp. 529–551. PMLR, 2021.

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

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

Héctor Ratia, Luis Montesano, and Ruben Martinez-Cantin. On the performance of maximum likelihood inverse reinforcement learning. *arXiv preprint arXiv:1202.1558*, 2012.

Nathan D Ratliff, J Andrew Bagnell, and Martin A Zinkevich. Maximum margin planning. In *Proceedings of the 23rd international conference on Machine learning*, pp. 729–736, 2006.

Stuart Russell. Learning agents for uncertain environments. In *Proceedings of the eleventh annual conference on Computational learning theory*, pp. 101–103, 1998.

Reza Shokri and Vitaly Shmatikov. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pp. 1310–1321, 2015.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pp. 5026–5033. IEEE, 2012.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 33:7611–7623, 2020.

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

Hanhan Zhou, Tian Lan, Guru Prasadh Venkataramani, and Wenbo Ding. Every parameter matters: Ensuring the convergence of federated learning with dynamic heterogeneous models reduction. *Advances in Neural Information Processing Systems*, 36, 2024.

A PROOF OF LEMMA 2

Given the definition of soft-Q function following Bellman equation:

$$Q^{\text{soft}}_{r,\pi}(s,a) = \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[r(s,a) + \gamma \mathbb{E}_{a'\sim\pi(\cdot|s')}\left(Q^{\text{soft}}_{r,\pi}(s',a') - \log \pi(a'|s')\right)\right] \tag{21}$$

The difference between soft-Q values under policies π and π' is:

$$\begin{aligned}|Q^{\text{soft}}_{r,\pi}(s,a) - Q^{\text{soft}}_{r,\pi'}(s,a)| = \gamma \Bigg| &\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[\mathbb{E}_{a'\sim\pi(\cdot|s')}\left(Q^{\text{soft}}_{r,\pi}(s',a') - \log \pi(a'|s')\right)\right] \\ &- \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[\mathbb{E}_{a'\sim\pi'(\cdot|s')}\left(Q^{\text{soft}}_{r,\pi'}(s',a') - \log \pi'(a'|s')\right)\right]\Bigg|\end{aligned} \tag{22}$$

Using the triangle inequality, we separate the terms in equation 22:

$$\begin{aligned}|Q^{\text{soft}}_{r,\pi}(s,a) - Q^{\text{soft}}_{r,\pi'}(s,a)| \leq \gamma \Bigg(&\left|\mathbb{E}_{s'\sim P(\cdot|s,a)}\mathbb{E}_{a'\sim\pi(\cdot|s')}\left[Q^{\text{soft}}_{r,\pi}(s',a') - Q^{\text{soft}}_{r,\pi'}(s',a')\right]\right| \\ &+ \left|\mathbb{E}_{s'\sim P(\cdot|s,a)}\mathbb{E}_{a'\sim\pi(\cdot|s')}\left[\log \pi(a'|s') - \log \pi'(a'|s')\right]\right| \Bigg)\end{aligned} \tag{23}$$

For the first term in equation 22, we apply Jensen's inequality to the absolute value function:

$$\left|\mathbb{E}_{s'\sim P(\cdot|s,a)}\mathbb{E}_{a'\sim\pi(\cdot|s')}\left[Q^{\text{soft}}_{r,\pi}(s',a') - Q^{\text{soft}}_{r,\pi'}(s',a')\right]\right| \leq \sup_{s',a'} |Q^{\text{soft}}_{r,\pi}(s',a') - Q^{\text{soft}}_{r,\pi'}(s',a')| \tag{24}$$

Similarly, the second term in equation 22 involving the log policies is bounded as:

$$\left|\mathbb{E}_{s'\sim P(\cdot|s,a)}\mathbb{E}_{a'\sim\pi(\cdot|s')}\left[\log \pi(a'|s') - \log \pi'(a'|s')\right]\right| \leq \sup_{s',a'} |\log \pi(a'|s') - \log \pi'(a'|s')| \tag{25}$$

Thus, we combine the two terms in equation 22:

$$|Q^{\text{soft}}_{r,\pi}(s,a) - Q^{\text{soft}}_{r,\pi'}(s,a)| \leq \gamma \sup_{s',a'} |Q^{\text{soft}}_{r,\pi}(s',a') - Q^{\text{soft}}_{r,\pi'}(s',a')| + \gamma \sup_{s',a'} |\log \pi(a'|s') - \log \pi'(a'|s')| \tag{26}$$

Since soft-Q values depend recursively on future rewards, we apply this bound recursively over time. At $t = 1$, the same bound holds:

$$\begin{aligned}|Q^{\text{soft}}_{r,\pi}(s_1,a_1) - Q^{\text{soft}}_{r,\pi'}(s_1,a_1)| \leq &\gamma \sup_{s_2,a_2} |Q^{\text{soft}}_{r,\pi}(s_2,a_2) - Q^{\text{soft}}_{r,\pi'}(s_2,a_2)| \\ &+ \gamma \sup_{s',a'} |\log \pi(a'|s') - \log \pi'(a'|s')|\end{aligned} \tag{27}$$

Substituting this into the previous equation, we get:

$$\begin{aligned}&|Q^{\text{soft}}_{r,\pi}(s_0 = s, a_0 = a) - Q^{\text{soft}}_{r,\pi'}(s_0 = s, a_0 = a)| \\ &\leq \gamma^2 \sup_{s_2,a_2} |Q^{\text{soft}}_{r,\pi}(s_2,a_2) - Q^{\text{soft}}_{r,\pi'}(s_2,a_2)| + \gamma \sup_{s',a'} |\log \pi(a'|s') - \log \pi'(a'|s')|(1+\gamma)\end{aligned} \tag{28}$$

Applying this recursively over n steps:

$$\begin{aligned}
&|Q^{\text{soft}}_{r,\pi}(s_0 = s, a_0 = a) - Q^{\text{soft}}_{r,\pi'}(s_0 = s, a_0 = a)| \\
&\leq \gamma^n \sup_{s_n, a_n} |Q^{\text{soft}}_{r,\pi}(s_n, a_n) - Q^{\text{soft}}_{r,\pi'}(s_n, a_n)| + \sup_{s', a'} |\log \pi(a'|s') - \log \pi'(a'|s')| \sum_{k=0}^{n-1} \gamma^k
\end{aligned} \tag{29}$$

As $n \to \infty$, the term $\gamma^n \sup_{s_n, a_n} |Q^{\text{soft}}_{r,\pi}(s_n, a_n) - Q^{\text{soft}}_{r,\pi'}(s_n, a_n)|$ tends to zero. The geometric series sum is $\sum_{k=0}^{\infty} \gamma^k = \frac{1}{1-\gamma}$. Thus, taking the infinity norm with respect to all s and a, the final bound is:

$$\|Q^{\text{soft}}_{r,\pi} - Q^{\text{soft}}_{r,\pi'}\|_\infty \leq \frac{\gamma}{1-\gamma} \|\log(\pi) - \log(\pi')\|_\infty \tag{30}$$

B Proof of Theorem 1

In this section we show the complete proof for the convergence results for the recovered global policies $\overline{\pi}_{(m,T)}$ and the global reward parameter $\overline{\theta}_{(m,T)}$ after m communications with communication period T.

B.1 Convergence of Policy Estimate $\overline{\pi}_{(m,T)}$

We first analyze the approximation error between the logarithm of the synchronized policy $\log(\overline{\pi}_{(m,T)}(a|s))$ and the logarithm of the optimal policy corresponding to the previous local reward parameter $\log(\pi_{\theta^i_{(m,T-1)}}(a|s))$ for all i. Specifically, we aim to bound the difference:

$$\left|\log(\overline{\pi}_{(m,T)}(a|s)) - \log(\pi_{\theta^i_{(m,T-1)}}(a|s))\right|$$

This difference represents the discrepancy between the synchronized policy after the m-th global aggregation and the optimal policy corresponding to the previous local reward parameter $\theta^i_{(m,T-1)}$.

We aim to show that the distance between the logarithms of the synchronized policy and the optimal policy can be bounded by the difference between their corresponding soft-Q values. Specifically, we want to bound:

$$\left|\log(\overline{\pi}_{(m,T)}(a|s)) - \log(\pi_{\theta^i_{(m,T-1)}}(a|s))\right| \leq \Delta_Q,$$

where Δ_Q involves the difference between the soft-Q values $\overline{Q}^{\text{soft}}_{(m,T-1)}(s,a)$ and $Q^{\text{soft}}_{r_{\theta^i_{(m,T-2)}}, \pi_{\theta^i_{(m,T-2)}}}(s,a)$.

Recall that the policy is proportional to the exponential of the soft-Q value in equation 6. Thus, we can write:

$$\begin{aligned}
\log(\overline{\pi}_{(m,T)}(a|s)) &= \log\left(\frac{\exp\left(\overline{Q}^{\text{soft}}_{m,T-1}(s,a)\right)}{\sum_{\tilde{a}} \exp\left(\overline{Q}^{\text{soft}}_{(m,T-1)}(s,\tilde{a})\right)}\right) \\
&= \overline{Q}^{\text{soft}}_{(m,T-1)}(s,a) - \log\left(\sum_{\tilde{a}} \exp\left(\overline{Q}^{\text{soft}}_{(m,T-1)}(s,\tilde{a})\right)\right)
\end{aligned} \tag{31}$$

Since $\log(\pi_{\theta^i_{(m,T-1)}}(a|s))$ is the optimal policy under reward parameter $\theta^i_{(m,T-1)}$, according to (Haarnoja et al., 2017), it has the form

$$\pi_{\theta^i_{(m,T-1)}}(a|s) = \frac{\exp(Q^{\text{soft}}_{r_{\theta^i_{(m,T-2)}}, \pi_{\theta^i_{(m,T-2)}}}(a|s)}(s,a))}{\sum_{\tilde{a}} \exp(Q^{\text{soft}}_{r_{\theta^i_{(m,T-2)}}, \pi_{\theta^i_{(m,T-2)}}}(a|s)}(s,\tilde{a}))} \tag{32}$$

Similarly, we have:

$$\log(\pi_{\theta^i_{(m,T-1)}}(a|s)) = Q^{\text{soft}}_{r_{\theta^i_{(m,T-2)}},\pi_{\theta^i_{(m,T-2)}}}(s,a) - \log\left(\sum_{\tilde{a}} \exp\left(Q^{\text{soft}}_{r_{\theta^i_{(m,T-2)}},\pi_{\theta^i_{(m,T-2)}}}(s,\tilde{a})\right)\right) \tag{33}$$

Subtracting the two expressions and use the triangle inequality, we can bound the absolute value of the difference by the sum of the absolute values:

$$\begin{aligned}
&\left|\log(\overline{\pi}_{(m,T)}(a|s)) - \log(\pi_{\theta^i_{(m,T-1)}}(a|s))\right| \\
&= \left| \left[\overline{Q}^{\text{soft}}_{(m,T-1)}(s,a) - \log\left(\sum_{\tilde{a}} \exp\left(\overline{Q}^{\text{soft}}_{(m,T-1)}(s,\tilde{a})\right)\right)\right] \right. \\
&\quad \left. - \left[Q^{\text{soft}}_{r_{\theta^i_{(m,T-2)}},\pi_{\theta^i_{(m,T-2)}}}(s,a) - \log\left(\sum_{\tilde{a}} \exp\left(Q^{\text{soft}}_{r_{\theta^i_{(m,T-2)}},\pi_{\theta^i_{(m,T-2)}}}(s,\tilde{a})\right)\right)\right] \right| \\
&= \left| \overline{Q}^{\text{soft}}_{(m,T-1)}(s,a) - Q^{\text{soft}}_{r_{\theta^i_{(m,T-2)}},\pi_{\theta^i_{(m,T-2)}}}(s,a) \right. \\
&\quad \left. - \left[\log\left(\sum_{\tilde{a}} \exp\left(\overline{Q}^{\text{soft}}_{(m,T-1)}(s,\tilde{a})\right)\right) - \log\left(\sum_{\tilde{a}} \exp\left(Q^{\text{soft}}_{r_{\theta^i_{(m,T-2)}},\pi_{\theta^i_{(m,T-2)}}}(s,\tilde{a})\right)\right)\right] \right| \\
&\leq \left|\overline{Q}^{\text{soft}}_{(m,T-1)}(s,a) - Q^{\text{soft}}_{r_{\theta^i_{(m,T-2)}},\pi_{\theta^i_{(m,T-2)}}}(s,a)\right| \\
&\quad + \left| \log\left(\sum_{\tilde{a}} \exp\left(\overline{Q}^{\text{soft}}_{(m,T-1)}(s,\tilde{a})\right)\right) - \log\left(\sum_{\tilde{a}} \exp\left(Q^{\text{soft}}_{r_{\theta^i_{(m,T-2)}},\pi_{\theta^i_{(m,T-2)}}}(s,\tilde{a})\right)\right) \right|
\end{aligned} \tag{34}$$

The second term in equation 34 involves the difference of logarithms of sums. We can bound it using properties of logarithms and the maximum difference of the soft-Q values.

We utilize the following inequality (as referenced in Equation 47 of (Zeng et al., 2022)):

$$\left|\log\left(\sum_{\tilde{a}} \exp(Q_1(s,\tilde{a}))\right) - \log\left(\sum_{\tilde{a}} \exp(Q_2(s,\tilde{a}))\right)\right| \leq \max_{\tilde{a}} |Q_1(s,\tilde{a}) - Q_2(s,\tilde{a})| \tag{35}$$

Applying equation 35, we get:

$$\begin{aligned}
&\left| \log\left(\sum_{\tilde{a}} \exp\left(\overline{Q}^{\text{soft}}_{(m,T-1)}(s,\tilde{a})\right)\right) - \log\left(\sum_{\tilde{a}} \exp\left(Q^{\text{soft}}_{r_{\theta^i_{(m,T-2)}},\pi_{\theta^i_{(m,T-2)}}}(s,\tilde{a})\right)\right) \right| \\
&\leq \max_{\tilde{a}} \left|\overline{Q}^{\text{soft}}_{(m,T-1)}(s,\tilde{a}) - Q^{\text{soft}}_{r_{\theta^i_{(m,T-2)}},\pi_{\theta^i_{(m,T-2)}}}(s,\tilde{a})\right|
\end{aligned} \tag{36}$$

Combining the results from equation 34 and equation 36:

$$\begin{aligned}
&\left|\log(\overline{\pi}_{(m,T)}(a|s)) - \log(\pi_{\theta^i_{(m,T-1)}}(a|s))\right| \\
&\leq \left|\overline{Q}^{\text{soft}}_{(m,T-1)}(s,a) - Q^{\text{soft}}_{r_{\theta^i_{(m,T-2)}},\pi_{\theta^i_{(m,T-2)}}}(s,a)\right| \\
&\quad + \max_{\tilde{a}} \left|\overline{Q}^{\text{soft}}_{(m,T-1)}(s,\tilde{a}) - Q^{\text{soft}}_{r_{\theta^i_{(m,T-2)}},\pi_{\theta^i_{(m,T-2)}}}(s,\tilde{a})\right|
\end{aligned} \tag{37}$$

Taking the infinity norm on equation 37 gives:

$$\| \log(\overline{\pi}_{(m,T)}) - \log(\pi_{\theta^i_{(m,T-1)}}) \|_\infty \leq 2 \| \overline{Q}^{\text{soft}}_{(m,T-1)} - Q^{\text{soft}}_{r_{\theta^i_{(m,T-2)}}, \pi_{\theta^i_{(m,T-2)}}} \|_\infty \tag{38}$$

By the definition of aggregation in equation 10, we have

$$\overline{Q}^{\text{soft}}_{(m,T-1)} = \frac{1}{N} \sum_{j=1}^{N} Q^{\text{soft}}_{r_{\theta^j_{(m,T-2)}}, \pi^j_{(m,T-2)}} \tag{39}$$

Plug above definition into equation 38 and by triangle inequality:

$$\begin{aligned} \| \overline{Q}^{\text{soft}}_{(m,T-1)} - Q^{\text{soft}}_{r_{\theta^i_{(m,T-2)}}, \pi_{\theta^i_{(m,T-2)}}} \|_\infty &= \| \sum_{j=1}^{N} \frac{1}{N} (Q^{\text{soft}}_{r_{\theta^j_{(m,T-2)}}, \pi^j_{(m,T-2)}} - Q^{\text{soft}}_{r_{\theta^i_{(m,T-2)}}, \pi_{\theta^i_{(m,T-2)}}}) \|_\infty \\ &\leq \frac{1}{N} \sum_{j=1}^{N} \| Q^{\text{soft}}_{r_{\theta^j_{(m,T-2)}}, \pi^j_{(m,T-2)}} - Q^{\text{soft}}_{r_{\theta^i_{(m,T-2)}}, \pi_{\theta^i_{(m,T-2)}}} \|_\infty \end{aligned} \tag{40}$$

Therefore, we move to analyse $\| Q^{\text{soft}}_{r_{\theta^j_{(m,T-2)}}, \pi^j_{(m,T-2)}} - Q^{\text{soft}}_{r_{\theta^i_{(m,T-2)}}, \pi_{\theta^i_{(m,T-2)}}} \|_\infty$, which is the difference of soft-Q values between two different local nodes, one under policy estimation, and the other under optimal policy. Looking back to the time right after last aggregation, where all local servers have the same reward parameter $\overline{\theta}_{(m-1,T)}$, we could further bound this difference using the difference of reward parameters, since the difference of local reward parameters are introduced by the local increment at each internal iteration except for the aggregation round.

We start by decomposing this difference into three terms and use the triangle inequality, we bound the sum :

$$\begin{aligned} &\left\| Q^{\text{soft}}_{r_{\theta^j_{(m,T-2)}}, \pi^j_{(m,T-2)}} - Q^{\text{soft}}_{r_{\theta^i_{(m,T-2)}}, \pi_{\theta^i_{(m,T-2)}}} \right\|_\infty \\ &= \Bigg\| \left(Q^{\text{soft}}_{r_{\theta^j_{(m,T-2)}}, \pi^j_{(m,T-2)}} - Q^{\text{soft}}_{r_{\theta^j_{(m,T-2)}}, \pi_{\theta^j_{(m,T-2)}}} \right) \\ &\quad + \left(Q^{\text{soft}}_{r_{\theta^j_{(m,T-2)}}, \pi_{\theta^j_{(m,T-2)}}} - Q^{\text{soft}}_{r_{\theta^j_{(m,0)}}, \pi_{\theta^j_{(m,0)}}} \right) \\ &\quad + \left(Q^{\text{soft}}_{r_{\theta^i_{(m,0)}}, \pi_{\theta^i_{(m,0)}}} - Q^{\text{soft}}_{r_{\theta^i_{(m,T-2)}}, \pi_{\theta^i_{(m,T-2)}}} \right) \Bigg\|_\infty \\ &\leq \left\| Q^{\text{soft}}_{r_{\theta^j_{(m,T-2)}}, \pi^j_{(m,T-2)}} - Q^{\text{soft}}_{r_{\theta^j_{(m,T-2)}}, \pi_{\theta^j_{(m,T-2)}}} \right\|_\infty \\ &\quad + \left\| Q^{\text{soft}}_{r_{\theta^j_{(m,T-2)}}, \pi_{\theta^j_{(m,T-2)}}} - Q^{\text{soft}}_{r_{\theta^j_{(m,0)}}, \pi_{\theta^j_{(m,0)}}} \right\|_\infty \\ &\quad + \left\| Q^{\text{soft}}_{r_{\theta^i_{(m,0)}}, \pi_{\theta^i_{(m,0)}}} - Q^{\text{soft}}_{r_{\theta^i_{(m,T-2)}}, \pi_{\theta^i_{(m,T-2)}}} \right\|_\infty \end{aligned} \tag{41}$$

The first term is the difference between the soft-Q values under the same reward parameter $\theta^j_{(m,T-2)}$ but different policies $\pi^j_{(m,T-2)}$ and $\pi_{\theta^j_{(m,T-2)}}$. The second term is the difference due to the change in reward parameters from $\theta^j_{(m,T-2)}$ to $\theta^j_{(m,0)}$, with corresponding optimal policies, and the third term is similar to the second term but for node i, comparing $\theta^i_{(m,0)}$ and $\theta^i_{(m,T-2)}$. We are utilizing the fact that $\theta^j_{(m,0)} = \theta^i_{(m,0)}$ since they are initialized after previous aggregation.

Applying equation 15 to equation 41, we have:

$$\begin{aligned}
&\|Q^{\text{soft}}_{r_{\theta^j_{(m,T-2)}},\pi^j_{(m,T-2)}} - Q^{\text{soft}}_{r_{\theta^i_{(m,T-2)}},\pi_{\theta^i_{(m,T-2)}}}\|_\infty \\
&\leq \left\| Q^{\text{soft}}_{r_{\theta^j_{(m,T-2)}},\pi^j_{(m,T-2)}} - Q^{\text{soft}}_{r_{\theta^j_{(m,T-2)}},\pi_{\theta^j_{(m,T-2)}}} \right\|_\infty \\
&\quad + L_q \left\| \theta^j_{(m,T-2)} - \theta^j_{(m,0)} \right\| + L_q \left\| \theta^i_{(m,0)} - \theta^i_{(m,T-2)} \right\|
\end{aligned} \tag{42}$$

Next, we express the differences in reward parameters in terms of gradient updates. According to equation 9:

$$\theta^j_{(m,T-2)} = \theta^j_{(m,0)} + \alpha \sum_{k=0}^{T-3} g^j_{(m,k)} \tag{43}$$

$$\theta^i_{(m,T-2)} = \theta^i_{(m,0)} + \alpha \sum_{k=0}^{T-3} g^i_{(m,k)} \tag{44}$$

Substituting back into our equation 42:

$$\begin{aligned}
&\|Q^{\text{soft}}_{r_{\theta^j_{(m,T-2)}},\pi^j_{(m,T-2)}} - Q^{\text{soft}}_{r_{\theta^i_{(m,T-2)}},\pi_{\theta^i_{(m,T-2)}}}\|_\infty \\
&\leq \left\| Q^{\text{soft}}_{r_{\theta^j_{(m,T-2)}},\pi^j_{(m,T-2)}} - Q^{\text{soft}}_{r_{\theta^j_{(m,T-2)}},\pi_{\theta^j_{(m,T-2)}}} \right\|_\infty + L_q\alpha \left\| \sum_{k=0}^{T-3} g^j_{(m,k)} \right\| + L_q\alpha \left\| \sum_{k=0}^{T-2} g^i_{(m,k)} \right\|
\end{aligned} \tag{45}$$

According to equation (56) in (Zeng et al., 2022)), the gradients are bounded as:

$$\|g^i_{(m,k)}\| \leq 2L_q \tag{46}$$

we can further bound the sums in equation 45:

$$\begin{aligned}
&\|Q^{\text{soft}}_{r_{\theta^j_{(m,T-2)}},\pi^j_{(m,T-2)}} - Q^{\text{soft}}_{r_{\theta^i_{(m,T-2)}},\pi_{\theta^i_{(m,T-2)}}}\|_\infty \\
&\leq \left\| Q^{\text{soft}}_{r_{\theta^j_{(m,T-2)}},\pi^j_{(m,T-2)}} - Q^{\text{soft}}_{r_{\theta^j_{(m,T-2)}},\pi_{\theta^j_{(m,T-2)}}} \right\|_\infty + 4(T-2)\alpha L_q^2
\end{aligned} \tag{47}$$

Now the difference is bounded by the difference between Q-values under reward parameter $\theta^j_{(m,T-2)}$ with respect to the previous approximated policy $\pi^j_{(m,T-2)}$ and the optimal policy $\pi_{\theta^j_{(m,T-2)}}$, plus some error terms. We come back from comparing Q-values across different nodes to evaluating the soft-Q value approximation within a single node.

By equation 57 in (Zeng et al., 2022):

$$\begin{aligned}
\|Q^{\text{soft}}_{r_{\theta^i_{(m,k)}},\pi^i_{(m,k)}} - Q^{\text{soft}}_{r_{\theta^i_{(m,k)}},\pi_{\theta^i_{(m,k)}}}\|_\infty \leq \gamma \|Q^{\text{soft}}_{r_{\theta^i_{(m,k-1)}},\pi^i_{(m,k-1)}} - Q^{\text{soft}}_{r_{\theta^i_{(m,k-1)}},\pi_{\theta^i_{(m,k-1)}}}\|_\infty \\
+ 4\alpha L_q^2, \quad 1 \leq k \leq T-1, \quad m \in \mathbb{N}, \quad \forall i
\end{aligned} \tag{48}$$

Above inequality provides a bounds of the local Q-value using the previous Q-value times a contraction factor γ plus some extra term. We could use it to compare the aggregated Q-value at m-th outer round with the aggregated Q-value at $m-1$ - th outer round:

$$
\begin{aligned}
&\|Q^{\text{soft}}_{r_{\theta^i_{(m,k)}},\pi^i_{(m,k)}} - Q^{\text{soft}}_{r_{\theta^i_{(m,k)}},\pi_{\theta^i_{(m,k)}}}\|_\infty \\
&\leq \gamma^k \|Q^{\text{soft}}_{r_{\theta^i_{(m,0)}},\pi^i_{(m,0)}} - Q^{\text{soft}}_{r_{\theta^i_{(m,0)}},\pi_{\theta^i_{(m,0)}}}\|_\infty + \sum_{i=0}^{k-1} \gamma^i \cdot 4\alpha L_q^2 \\
&= \gamma^k \|Q^{\text{soft}}_{r_{\theta^i_{(m,0)}},\pi^i_{(m,0)}} - Q^{\text{soft}}_{r_{\theta^i_{(m,0)}},\pi_{\theta^i_{(m,0)}}}\|_\infty + \frac{1-\gamma^k}{1-\gamma} \cdot 4\alpha L_q^2, \quad m \in \mathbb{N}, \quad 1 \leq k \leq T-1
\end{aligned} \tag{49}
$$

Apply equation 49 to equation 47:

$$
\begin{aligned}
&\|Q^{\text{soft}}_{r_{\theta^j_{(m,T-2)}},\pi^j_{(m,T-2)}} - Q^{\text{soft}}_{r_{\theta^i_{(m,T-2)}},\pi_{\theta^i_{(m,T-2)}}}\|_\infty \\
&\leq \|Q^{\text{soft}}_{r_{\theta^j_{(m,T-2)}},\pi^j_{(m,T-2)}} - Q^{\text{soft}}_{r_{\theta^j_{(m,T-2)}},\pi_{\theta^j_{(m,T-2)}}}\|_\infty + 4(T-2))\alpha L_q^2 \\
&\leq \gamma^{T-2} \|Q^{\text{soft}}_{r_{\theta^j_{(m,0)}},\pi^j_{(m,0)}} - Q^{\text{soft}}_{r_{\theta^j_{(m,0)}},\pi_{\theta^j_{(m,0)}}}\|_\infty + \frac{1-\gamma^{T-2}}{1-\gamma} \cdot 4\alpha L_q^2 + 4(T-2)\alpha L_q^2 \\
&= \gamma^{T-2} \|Q^{\text{soft}}_{r_{\overline{\theta}_{(m-1,T)}},\overline{\pi}_{(m-1,T)}} - Q^{\text{soft}}_{r_{\overline{\theta}_{(m-1,T)}},\pi_{\overline{\theta}_{(m-1,T)}}}\|_\infty + 4\alpha \left(\frac{1-\gamma^{T-2}}{1-\gamma} + T - 2 \right) L_q^2
\end{aligned} \tag{50}
$$

Plug equation 50 into equation 40:

$$
\begin{aligned}
&\|\overline{Q}^{\text{soft}}_{(m,T-1)} - Q^{\text{soft}}_{r_{\theta^i_{(m,T-2)}},\pi_{\theta^i_{(m,T-2)}}}\|_\infty \\
&\leq \frac{1}{N} \sum_{j=1}^{N} \|Q^{\text{soft}}_{r_{\theta^j_{(m,T-2)}},\pi^j_{(m,T-2)}} - Q^{\text{soft}}_{r_{\theta^i_{(m,T-1)}},\pi_{\theta^i_{(m,T-1)}}}\|_\infty \\
&\leq \frac{1}{N} \sum_{j=1}^{N} \left[\gamma^{T-2} \|Q^{\text{soft}}_{r_{\overline{\theta}_{(m-1,T)}},\overline{\pi}_{(m-1,T)}} - Q^{\text{soft}}_{r_{\overline{\theta}_{(m-1,T)}},\pi_{\overline{\theta}_{(m-1,T)}}}\|_\infty + 4\alpha \left(2 \cdot \frac{1-\gamma^{T-2}}{1-\gamma} + T - 2 \right) L_q^2 \right] \\
&= \gamma^{T-2} \|Q^{\text{soft}}_{r_{\overline{\theta}_{(m-1,T)}},\overline{\pi}_{(m-1,T)}} - Q^{\text{soft}}_{r_{\overline{\theta}_{(m-1,T)}},\pi_{\overline{\theta}_{(m-1,T)}}}\|_\infty + 4\alpha \left(\frac{1-\gamma^{T-2}}{1-\gamma} + T - 2 \right) L_q^2
\end{aligned} \tag{51}
$$

We further analyze $\|Q^{\text{soft}}_{r_{\overline{\theta}_{(m-1,T)}},\overline{\pi}_{(m-1,T)}} - Q^{\text{soft}}_{r_{\overline{\theta}_{(m-1,T)}},\pi_{\overline{\theta}_{(m-1,T)}}}\|_\infty$ and use triangle inequality to decompose it into three parts to acquire the same form of Q-value difference in the previous outer round:

$$
\begin{aligned}
&\|Q^{\text{soft}}_{r_{\overline{\theta}_{(m-1,T)}},\overline{\pi}_{(m-1,T)}} - Q^{\text{soft}}_{r_{\overline{\theta}_{(m-1,T)}},\pi_{\overline{\theta}_{(m-1,T)}}}\|_\infty \\
&= \|(Q^{\text{soft}}_{r_{\overline{\theta}_{(m-1,T)}},\overline{\pi}_{(m-1,T)}} - \overline{Q}^{\text{soft}}_{(m-1,T-1)}) + (\overline{Q}^{\text{soft}}_{(m-1,T-1)} - Q^{\text{soft}}_{r_{\theta^i_{(m-1,T-1)}},\pi_{\theta^i_{(m-1,T-1)}}}) \\
&\quad + (Q^{\text{soft}}_{r_{\theta^i_{(m-1,T-1)}},\pi_{\theta^i_{(m-1,T-1)}}} - Q^{\text{soft}}_{r_{\overline{\theta}_{(m-1,T)}},\pi_{\overline{\theta}_{(m-1,T)}}})\|_\infty \\
&\leq \|(Q^{\text{soft}}_{r_{\overline{\theta}_{(m-1,T)}},\overline{\pi}_{(m-1,T)}} - \overline{Q}^{\text{soft}}_{(m-1,T-1)}\|_\infty + \|\overline{Q}^{\text{soft}}_{(m-1,T-1)} - Q^{\text{soft}}_{r_{\theta^i_{(m-1,T-1)}},\pi_{\theta^i_{(m-1,T-1)}}}\|_\infty \\
&\quad + \|Q^{\text{soft}}_{r_{\theta^i_{(m-1,T-1)}},\pi_{\theta^i_{(m-1,T-1)}}} - Q^{\text{soft}}_{r_{\overline{\theta}_{(m-1,T)}},\pi_{\overline{\theta}_{(m-1,T)}}}\|_\infty
\end{aligned} \tag{52}
$$

We first bound the first term in equation 52 by introducing an middle term and triangle inequality:

$$
\begin{aligned}
&\|Q^{\text{soft}}_{r_{\overline{\theta}_{(m-1,T)}},\overline{\pi}_{(m-1,T)}} - \overline{Q}^{\text{soft}}_{(m-1,T-1)}\|_\infty \\
&= \|\overline{Q}^{\text{soft}}_{(m-1,T-1)} - \frac{1}{N}\sum_{j=1}^{N} Q^{\text{soft}}_{r_{\theta^j_{(m-1,T-2)}},\pi^j_{(m-1,T-2)}}\|_\infty \\
&= \|(Q^{\text{soft}}_{r_{\overline{\theta}_{(m-1,T)}},\overline{\pi}_{(m-1,T)}} - Q^{\text{soft}}_{r_{\theta^j_{(m-1,T-2)}},\overline{\pi}_{(m-1,T)}}) \\
&\quad + (Q^{\text{soft}}_{r_{\theta^j_{(m-1,T-2)}},\overline{\pi}_{(m-1,T)}} - \frac{1}{N}\sum_{j=1}^{N} Q^{\text{soft}}_{r_{\theta^j_{(m-1,T-2)}},\pi^j_{(m-1,T-2)}})\| \\
&\leq \|Q^{\text{soft}}_{r_{\overline{\theta}_{(m-1,T)}},\overline{\pi}_{(m-1,T)}} - Q^{\text{soft}}_{r_{\theta^j_{(m-1,T-2)}},\overline{\pi}_{(m-1,T)}}\|_\infty \\
&\quad + \frac{1}{N}\sum_{j=1}^{N} \|Q^{\text{soft}}_{r_{\theta^j_{(m-1,T-2)}},\overline{\pi}_{(m-1,T)}} - Q^{\text{soft}}_{r_{\theta^j_{(m-1,T-2)}},\pi^j_{(m-1,T-2)}}\|_\infty
\end{aligned}
\tag{53}
$$

For the first term in equation 53, we leverage Lemma 7 in (Zeng et al., 2022), which states:

$$
|Q^{\text{soft}}_{r_{\theta_1},\pi} - Q^{\text{soft}}_{r_{\theta_2},\pi}| \leq L_q\|\theta_1 - \theta_2\|, \forall\pi, \forall\theta_1, \theta_2, \forall s \in \mathcal{S}, \forall a \in \mathcal{A} \tag{54}
$$

Then we could further bound the difference between θ's using equation 9 and equation 46:

$$
\begin{aligned}
&\|Q^{\text{soft}}_{r_{\overline{\theta}_{(m-1,T)}},\overline{\pi}_{(m-1,T)}} - Q^{\text{soft}}_{r_{\theta^j_{(m-1,T-2)}},\overline{\pi}_{(m-1,T)}}\|_\infty \\
&\leq L_q\|\overline{\theta}_{(m-1,T)} - \theta^j_{(m-1,T-2)}\|_\infty \\
&= L_q\|\frac{1}{N}\sum_{j=1}^{N}\theta^j_{(m-1,T-1)} - \theta^j_{(m-1,T-2)}\| \\
&\leq \frac{L_q}{N}\sum_{j=1}^{N}\|\theta^j_{(m-1,T-1)} - \theta^i_{(m-1,T-2)}\| \\
&= \frac{L_q}{N}\sum_{j=1}^{N}\|(\theta^j_{(m-1,T-1)} - \theta^j_{(m-1,0)}) + (\theta^i_{(m-1,0)} - \theta^i_{(m-1,T-2)})\| \\
&\leq \frac{L_q}{N}\sum_{j=1}^{N}(\|\theta^j_{(m-1,T-1)} - \theta^j_{(m-1,0)}\| + \|\theta^i_{(m-1,0)} - \theta^i_{(m-1,T-2)}\|) \\
&= \frac{L_q}{N}\sum_{j=1}^{N}\left(\alpha\left\|\sum_{k=0}^{T-2} g^j_{(m,k)}\right\| + \alpha\left\|\sum_{k=0}^{T-3} g^i_{(m,k)}\right\|\right) \\
&\leq 2(2T-3)\alpha L_q^2
\end{aligned}
$$

For the second term in equation 53, by Lemma 2:

$$
\begin{aligned}
&\|Q^{\text{soft}}_{r_{\theta^j_{(m-1,T-2)}},\overline{\pi}_{(m-1,T)}} - Q^{\text{soft}}_{r_{\theta^j_{(m-1,T-2)}},\pi^j_{(m-1,T-2)}}\|_\infty \\
&\leq \frac{1-\gamma}{\gamma}\|\log(\overline{\pi}_{(m-1,T)}) - \log(\pi^j_{(m-1,T-2)})\|_\infty
\end{aligned}
\tag{55}
$$

We get result similar to equation 38 using similar techniques and decompose equation 55 using triangle inequality:

$$
\begin{aligned}
&\| \log(\overline{\pi}_{(m-1,T)}) - \log(\pi^j_{(m-1,T-2)}) \|_\infty \\
&\leq 2\| \overline{Q}^{\text{soft}}_{(m-1,T-1)} - Q^{\text{soft}}_{r_{\theta^i_{(m-1,T-3)}},\pi_{\theta^i_{(m-1,T-3)}}} \|_\infty \\
&= 2\| (\overline{Q}^{\text{soft}}_{(m-1,T-1)} - Q^{\text{soft}}_{r_{\theta^i_{(m-1,T-2)}},\pi_{\theta^i_{(m-1,T-2)}}}) + (Q^{\text{soft}}_{r_{\theta^i_{(m-1,T-2)}},\pi_{\theta^i_{(m-1,T-2)}}} - Q^{\text{soft}}_{r_{\theta^i_{(m-1,T-3)}},\pi_{\theta^i_{(m-1,T-3)}}}) \|_\infty \\
&\leq 2(\| \overline{Q}^{\text{soft}}_{(m-1,T-1)} - Q^{\text{soft}}_{r_{\theta^i_{(m-1,T-2)}},\pi_{\theta^i_{(m-1,T-2)}}} \|_\infty + \| Q^{\text{soft}}_{r_{\theta^i_{(m-1,T-2)}},\pi_{\theta^i_{(m-1,T-2)}}} - Q^{\text{soft}}_{r_{\theta^i_{(m-1,T-3)}},\pi_{\theta^i_{(m-1,T-3)}}} \|_\infty)
\end{aligned} \tag{56}
$$

The last term could be controlled according to what we did for the last two terms in equation 41:

$$
\| Q^{\text{soft}}_{r_{\theta^i_{(m-1,T-2)}},\pi_{\theta^i_{(m-1,T-2)}}} - Q^{\text{soft}}_{r_{\theta^i_{(m-1,T-3)}},\pi_{\theta^i_{(m-1,T-3)}}} \|_\infty \leq 2\alpha L_q^2 \tag{57}
$$

Plugging above results into equation 52, we have:

$$
\begin{aligned}
&\| Q^{\text{soft}}_{r_{\overline{\theta}_{(m-1,T)}},\overline{\pi}_{(m-1,T)}} - Q^{\text{soft}}_{r_{\overline{\theta}_{(m-1,T)}},\pi_{\overline{\theta}_{(m-1,T)}}} \|_\infty \\
&\leq \frac{1-\gamma}{\gamma} \| \overline{Q}^{\text{soft}}_{(m-1,T-1)} - Q^{\text{soft}}_{r_{\theta^i_{(m-1,T-2)}},\pi_{\theta^i_{(m-1,T-2)}}} \|_\infty + 2(2T-3)\alpha L_q^2 + \frac{1-\gamma}{\gamma} \cdot 4\alpha L_q^2
\end{aligned} \tag{58}
$$

We further plug equation 58 to equation 51:

$$
\begin{aligned}
&\| \overline{Q}^{\text{soft}}_{(m,T-1)} - Q^{\text{soft}}_{r_{\theta^i_{(m,T-2)}},\pi_{\theta^i_{(m,T-2)}}} \|_\infty \\
&\leq \gamma^{T-2} \| Q^{\text{soft}}_{r_{\overline{\theta}_{(m-1,T)}},\overline{\pi}_{(m-1,T)}} - Q^{\text{soft}}_{r_{\overline{\theta}_{(m-1,T)}},\pi_{\overline{\theta}_{(m-1,T)}}} \|_\infty + 4\alpha \left(\frac{1-\gamma^{T-2}}{1-\gamma} + T - 2 \right) L_q^2 \\
&\leq (1-\gamma)\gamma^{T-1} \| \overline{Q}^{\text{soft}}_{(m-1,T-1)} - Q^{\text{soft}}_{r_{\theta^i_{(m-1,T-2)}},\pi_{\theta^i_{(m-1,T-2)}}} \|_\infty \\
&+ \left(2\frac{1-\gamma^{T-2}}{1-\gamma} + (1-\gamma)^2\gamma^{T-2} + \frac{1-\gamma}{\gamma}(2T-3) + 2(T-2) \right) L_q^2
\end{aligned} \tag{59}
$$

Summing the inequality from $m = 1$ to $m = M$ gives:

$$
\begin{aligned}
&\sum_{m=1}^{M} \| \overline{Q}^{\text{soft}}_{(m,T-1)} - Q^{\text{soft}}_{r_{\theta^i_{(m,T-2)}},\pi_{\theta^i_{(m,T-2)}}} \|_\infty \\
&\leq (1-\gamma)\gamma^{T-1} \sum_{m=1}^{M} \| \overline{Q}^{\text{soft}}_{(m-1,T-1)} - Q^{\text{soft}}_{r_{\theta^i_{(m-1,T-2)}},\pi_{\theta^i_{(m-1,T-2)}}} \|_\infty \\
&\quad + M \cdot [2\frac{1-\gamma^{T-2}}{1-\gamma} + (1-\gamma)^2\gamma^{T-2} + \frac{1-\gamma}{\gamma}(2T-3) + 2(T-2)] L_q^2
\end{aligned} \tag{60}
$$

Rearranging the inequality, it holds that:

$$\begin{aligned}
&\big(1-(1-\gamma)\gamma^{T-1}\big)\big)\sum_{m=1}^{M}\|\overline{Q}^{\text{soft}}_{(m,T-1)}-Q^{\text{soft}}_{r_{\theta^i_{(m,T-2)}},\pi_{\theta^i_{(m,T-2)}}}\|_\infty\\
&\leq(1-\gamma)\gamma^{T-1}\Bigg(\|\overline{Q}^{\text{soft}}_{(0,T-1)}-Q^{\text{soft}}_{r_{\theta^i_{(0,T-2)}},\pi_{\theta^i_{(0,T-2)}}}\|_\infty\\
&\quad-\|\overline{Q}^{\text{soft}}_{(M,T-1)}-Q^{\text{soft}}_{r_{\theta^i_{(M,T-2)}},\pi_{\theta^i_{(M,T-2)}}}\|_\infty\Bigg)\\
&\quad+M\cdot[2\frac{1-\gamma^{T-2}}{1-\gamma}+(1-\gamma)^2\gamma^{T-2}+\frac{1-\gamma}{\gamma}(2T-3)+2(T-2)]L_q^2\\
&\leq(1-\gamma)\gamma^{T-1}\Bigg(\|\overline{Q}^{\text{soft}}_{(0,T-1)}-Q^{\text{soft}}_{r_{\theta^i_{(0,T-2)}},\pi_{\theta^i_{(0,T-2)}}}\|_\infty\\
&\quad+M\cdot[2\frac{1-\gamma^{T-2}}{1-\gamma}+(1-\gamma)^2\gamma^{T-2}+\frac{1-\gamma}{\gamma}(2T-3)+2(T-2)]L_q^2
\end{aligned}\tag{61}$$

Dividing by $1-(1-\gamma)\gamma^{T-1}$ on both sides, we get

$$\begin{aligned}
&\sum_{m=1}^{M}\|\overline{Q}^{\text{soft}}_{(m,T-1)}-Q^{\text{soft}}_{r_{\theta^i_{(m,T-2)}},\pi_{\theta^i_{(m,T-2)}}}\|_\infty\\
&\leq\frac{(1-\gamma)\gamma^{T-1}}{1-(1-\gamma)\gamma^{T-1}}\|\overline{Q}^{\text{soft}}_{(0,T-1)}-Q^{\text{soft}}_{r_{\theta^i_{(0,T-2)}},\pi_{\theta^i_{(0,T-2)}}}\|_\infty\\
&\quad+M\cdot\frac{2\frac{1-\gamma^{T-2}}{1-\gamma}+(1-\gamma)^2\gamma^{T-2}+\frac{1-\gamma}{\gamma}(2T-3)+2(T-2)]L_q^2}{1-(1-\gamma)\gamma^{T-1}}
\end{aligned}$$

Denote $C_0=\|\overline{Q}^{\text{soft}}_{(0,T-1)}-Q^{\text{soft}}_{r_{\theta^i_{(0,T-2)}},\pi_{\theta^i_{(0,T-2)}}}\|_\infty$.

Dividing by M on both sides, we get

$$\begin{aligned}
&\frac{1}{M}\sum_{m=1}^{M}\|\overline{Q}^{\text{soft}}_{(m,T-1)}-Q^{\text{soft}}_{r_{\theta^i_{(m,T-2)}},\pi_{\theta^i_{(m,T-2)}}}\|_\infty\\
&\leq\frac{(1-\gamma)\gamma^{T-1}}{(1-(1-\gamma)\gamma^{T-1})M}C_0+\frac{2\frac{1-\gamma^{T-2}}{1-\gamma}+(1-\gamma)^2\gamma^{T-2}+\frac{1-\gamma}{\gamma}(2T-3)+2(T-2)]L_q^2}{1-(1-\gamma)\gamma^{T-1}}
\end{aligned}\tag{62}$$

Recall the step size is defined as $\alpha_{(m,t)}=\frac{\alpha_0}{(mT+t)^\sigma}$ where $\sigma>0$. Then we have the following result:

$$\begin{aligned}
&\frac{1}{M}\sum_{m=1}^{M}\|\overline{Q}^{\text{soft}}_{(m,T-1)}-Q^{\text{soft}}_{r_{\theta^i_{(m,T-2)}},\pi_{\theta^i_{(m,T-2)}}}\|_\infty\\
&=\mathcal{O}(M^{-1}\gamma^T)+\mathcal{O}(M^{-\sigma}T^{1-\sigma})
\end{aligned}\tag{63}$$

Going back to the convergence of policy approximation:

$$\begin{aligned}
&\frac{1}{M}\sum_{m=1}^{M}|\log(\overline{\pi}_{(m,T)}(a|s))-\log(\pi_{\theta^i_{(m,T-1)}}(a|s))|_\infty\\
&=\mathcal{O}(M^{-1}\gamma^T)+\mathcal{O}(M^{-\sigma}T^{1-\sigma})
\end{aligned}\tag{64}$$

B.2 CONVERGENCE OF THE GLOBAL REWARD PARAMETER $\overline{\theta}_{(m,T)}$

By the Lipschitz smooth property of the likelihood target equation 16, the definition of reward aggregation equation 12, and the reward parameter update rule equation 9:

$$
\begin{aligned}
& L(\overline{\theta}_{(m,T))}) \\
& \geq L(\overline{\theta}_{(m-1,T)}) + \left\langle \nabla L(\overline{\theta}_{(m,T)}), \overline{\theta}_{(m,T)} - \overline{\theta}_{(m-1,T)} \right\rangle - \frac{L_c}{2} \|\overline{\theta}_{(m,T)} - \overline{\theta}_{(m-1,T)}\|^2 \\
& = L(\overline{\theta}_{(m-1,T)}) + \left\langle \nabla L(\overline{\theta}_{(m,T)}, \frac{1}{N} \sum_{j=1}^{N} (\theta^j_{(m,T-1)} - \theta^j_{(m,0)}) \right\rangle - \frac{L_c}{2} \| \frac{1}{N} \sum_{j=1}^{N} (\theta^j_{(m,T-1)} - \theta^j_{(m,0)}) \|^2 \\
& = L(\overline{\theta}_{(m-1,T)}) + \alpha \left\langle \nabla L(\overline{\theta}_{(m,T)}), \frac{1}{N} \sum_{j=1}^{N} \sum_{k=0}^{T-2} g^j_{(m,k)} \right\rangle - \frac{L_c \alpha^2}{2} \| \frac{1}{N} \sum_{j=1}^{N} \sum_{k=0}^{T-2} g^j_{(m,k)} \|^2
\end{aligned}
\tag{65}
$$

We compare $g^i_{(m,k)}$ with the true gradient of $L(\theta^i_{(m,k)})$ and leverage the fact that $\|\nabla L(\theta)\|_\infty \leq 2L_q$:

$$
\begin{aligned}
& L(\overline{\theta}_{(m,T))}) \\
& \geq L(\overline{\theta}_{(m-1,T)}) + \alpha \left\langle \nabla L(\overline{\theta}_{(m,T)}), \frac{1}{N} \sum_{j=1}^{N} \sum_{k=0}^{T-2} g^j_{(m,k)} \right\rangle - \frac{L_c \alpha^2}{2} \| \frac{1}{N} \sum_{j=1}^{N} \sum_{k=0}^{T-2} g^j_{(m,k)} \|^2 \\
& \geq L(\overline{\theta}_{(m-1,T)}) + \alpha \left\langle \nabla L(\overline{\theta}_{(m,T)}), \frac{1}{N} \sum_{j=1}^{N} \sum_{k=0}^{T-2} g^j_{(m,k)} - (T-1) \nabla L(\overline{\theta}_{(m-1,T)}) \right\rangle \\
& \quad + \alpha(T-1) \|\nabla L(\overline{\theta}_{(m-1,T)})\|^2 - \frac{L_c \alpha^2}{2} \| \frac{1}{N} \sum_{j=1}^{N} \sum_{k=0}^{T-2} g^j_{(m,k)} \|^2 \\
& \geq L(\overline{\theta}_{(m-1,T)}) - 2\alpha L_q \cdot \| \frac{1}{N} \sum_{j=1}^{N} \sum_{k=0}^{T-2} [g^j_{(m,k)} - \nabla L(\overline{\theta}_{(m-1,T)})] \| \\
& \quad + \alpha(T-1) \|\nabla L(\overline{\theta}_{(m-1,T)})\|^2 - (T-1)^2 \cdot \frac{L_c L_q^2 \alpha^2}{2}
\end{aligned}
\tag{66}
$$

For $g^j(m,k)$ in equation 66, we evaluate its distance to $\nabla L(\theta^j_{(m,k)})$ and also consider the distance between $\nabla L(\theta^j_{(m,k)})$ and $\nabla L(\overline{\theta}_{(m,T)}$ with the help of triangle inequality:

$$
\begin{aligned}
& L(\overline{\theta}_{(m,T))}) \\
& \geq L(\overline{\theta}_{(m-1,T)}) - 2\alpha Lq \cdot \| \frac{1}{N} \sum_{j=1}^{N} \sum_{k=0}^{T-2} [g^j_{(m,k)} - \nabla L(\theta^j_{(m,k)}) + \nabla L(\theta^j_{(m,k)}) - \nabla L(\theta^j_{(m,0)})] \| \\
& \quad + \alpha(T-1) \|\nabla L(\overline{\theta}_{(m-1,T)})\|^2 - \frac{(T-1)^2 L_c L_q^2 \alpha^2}{2} \\
& \geq L(\overline{\theta}_{(m-1,T)}) - 2\alpha L_q \cdot [\frac{1}{N} \sum_{j=1}^{N} \sum_{k=0}^{T-2} (\|g^j_{(m,k)} - \nabla L(\theta^j_{(m,k)})\| + \|\nabla L(\theta^j_{(m,k)}) - \nabla L(\theta^j_{(m,0)})\|)] \\
& \quad + \alpha(T-1) \|\nabla L(\overline{\theta}_{(m-1,T)})\|^2 - \frac{(T-1)^2 L_c L_q^2 \alpha^2}{2}
\end{aligned}
\tag{67}
$$

Taking expectation over both sides:

$$
\begin{aligned}
&\mathbb{E}[L(\overline{\theta}_{(m,T)})] \\
&\geq \mathbb{E}[L(\overline{\theta}_{(m-1,T)})] - 2\alpha L_q \cdot [\frac{1}{N}\sum_{j=1}^{N}\sum_{k=0}^{T-2}(\mathbb{E}\|g^j_{(m,k)} - \nabla L(\theta^j_{(m,k)})\| + \|\nabla L(\theta^j_{(m,k)}) - \nabla L(\theta_{(m,0)})\|)] \\
&\quad + \alpha(T-1)\|\nabla L(\overline{\theta}_{(m-1,T)})\|^2 - \frac{(T-1)^2 L_c L_q^2 \alpha^2}{2}
\end{aligned} \tag{68}
$$

According to equation (62) and (63) in (Zeng et al., 2022), we have:

$$
\mathbb{E}\|g^j_{(m,k)} - \nabla L(\theta^j_{(m,k)})\| \leq 2L_q C_d \sqrt{|\mathcal{S}| \cdot |\mathcal{A}|}\mathbb{E}[\|Q^{\text{soft}}_{r_{\theta^j_{(m,k)}},\pi^j_{(m,k)}} - Q^{\text{soft}}_{r_{\theta^j_{(m,k)}},\pi_{\theta^j_{(m,k)}}}\|_\infty \tag{69}
$$

Then, using the Lipschitz property of L in equation 16:

$$
\begin{aligned}
&\mathbb{E}[L(\overline{\theta}_{(m,T)})] \\
&\geq \mathbb{E}[L(\overline{\theta}_{(m-1,T)})] - 2\alpha L_q \cdot \frac{1}{N}\sum_{j=1}^{N}\sum_{k=0}^{T-2}(2L_q C_d \sqrt{|\mathcal{S}| \cdot |\mathcal{A}|}\mathbb{E}[\|Q^{\text{soft}}_{r_{\theta^j_{(m,k)}},\pi^j_{(m,k)}} - Q^{\text{soft}}_{r_{\theta^j_{(m,k)}},\pi_{\theta^j_{(m,k)}}}\|_\infty \\
&\quad + \mathbb{E}[\|\nabla L(\theta^j_{(m,k)}) - \nabla L(\theta^j_{(m,0)})\|]) + \alpha(T-1)\mathbb{E}[\|\nabla L(\overline{\theta}_{(m-1,T)})\|^2] - \frac{(T-1)^2 L_c L_q^2 \alpha^2}{2} \\
&\geq \mathbb{E}[L(\theta^i_{(m-1,T)})] - 2\alpha L_q \cdot [\frac{1}{N}\sum_{j=1}^{N}\sum_{k=0}^{T-2}(2L_q C_d \sqrt{|\mathcal{S}| \cdot |\mathcal{A}|}\mathbb{E}[\|Q^{\text{soft}}_{r_{\theta^j_{(m,k)}},\pi^j_{(m,k)}} - Q^{\text{soft}}_{r_{\theta^j_{(m,k)}},\pi_{\theta^j_{(m,k)}}}\|_\infty] \\
&\quad + \mathbb{E}[L_c\|\theta^j_{(m,k)} - \theta^j_{(m,0)}\|])] + \alpha(T-1)\mathbb{E}[\|\nabla L(\overline{\theta}_{(m-1,T)})\|^2] - \frac{(T-1)^2 L_c L_q^2 \alpha^2}{2}
\end{aligned} \tag{70}
$$

Similar to equation 47, we have $\|\theta^j_{(m,k)} - \theta^j_{(m,0)}\| \leq 2k\alpha L_q$, applying equation 49 to equation 70, we have:

$$
\begin{aligned}
&\mathbb{E}[L(\overline{\theta}_{(m,T)})] \\
&\geq \mathbb{E}[L(\overline{\theta}_{(m-1,T)})] - 2\alpha L_q \cdot [\frac{1}{N}\sum_{j=1}^{N}\sum_{k=0}^{T-2}(2L_q C_d \sqrt{|\mathcal{S}| \cdot |\mathcal{A}|}\gamma^k\mathbb{E}[\|Q^{\text{soft}}_{r_{\theta^j_{(m,0)}},\pi^j_{(m,0)}} - Q^{\text{soft}}_{r_{\theta^j_{(m,0)}},\pi_{\theta^j_{(m,0)}}}\|_\infty]] \\
&\quad + 2L_q C_d \sqrt{|\mathcal{S}| \cdot |\mathcal{A}|}\frac{1-\gamma^k}{1-\gamma} \cdot 4\alpha L_q^2 + 2k\alpha L_c L_q) + \alpha(T-1)\mathbb{E}[\|\nabla L(\overline{\theta}_{(m-1,T)})\|^2] - \frac{(T-1)^2 L_c L_q^2 \alpha^2}{2} \\
&= \mathbb{E}[L(\overline{\theta}_{(m-1,T)})] - 2\alpha L_q \cdot \sum_{k=0}^{T-2}\gamma^k(2L_q C_d \sqrt{|\mathcal{S}| \cdot |\mathcal{A}|}\mathbb{E}[\|Q^{\text{soft}}_{r_{\overline{\theta}_{(m-1,T)}},\overline{\pi}_{(m-1,T)}} - Q^{\text{soft}}_{r_{\overline{\theta}_{(m-1,T)}},\pi_{\overline{\theta}_{(m-1,T)}}}\|_\infty] \\
&\quad + 2L_q C_d \sqrt{|\mathcal{S}| \cdot |\mathcal{A}|}\frac{1-\gamma^k}{1-\gamma} \cdot 4\alpha L_q^2 + k\alpha L_c L_q) + \alpha(T-1)\mathbb{E}[\|\nabla L(\overline{\theta}_{(m-1,T)})\|^2] - \frac{(T-1)^2 L_c L_q^2 \alpha^2}{2} \\
&= \mathbb{E}[L(\overline{\theta}_{(m-1,T)})] - \frac{4(1-\gamma^{T-1})}{1-\gamma}\alpha L_q^2 C_d \sqrt{|\mathcal{S}| \cdot |\mathcal{A}|}\mathbb{E}[\|Q^{\text{soft}}_{r_{\overline{\theta}_{(m-1,T)}},\overline{\pi}_{(m-1,T)}} - Q^{\text{soft}}_{r_{\overline{\theta}_{(m-1,T)}},\pi_{\overline{\theta}_{(m-1,T)}}}\|_\infty] \\
&\quad - 8\alpha L_q^3 C_d \sqrt{|\mathcal{S}| \cdot |\mathcal{A}|} \cdot \frac{T-1-\frac{1-\gamma^{T-1}}{1-\gamma}}{1-\gamma} - T(T-1)\alpha^2 L_c L_q^2 + \alpha(T-1)\mathbb{E}[\|\nabla L(\overline{\theta}_{(m-1,T)})\|^2] \\
&\quad - \frac{(T-1)^2 L_c L_q^2 \alpha^2}{2}
\end{aligned} \tag{71}
$$

Plugging the result in equation 58, we have:

$$
\begin{aligned}
&\mathbb{E}[L(\overline{\theta}_{(m,T)})] \\
&\geq \mathbb{E}[L(\overline{\theta}_{(m-1,T)})] - \frac{4(1-\gamma^{T-1})}{1-\gamma}\alpha L_q^2 C_d\sqrt{|\mathcal{S}|\cdot|\mathcal{A}|}\mathbb{E}[\|Q^{\text{soft}}_{r_{\overline{\theta}_{(m-1,T)}},\overline{\pi}_{(m-1,T)}} - Q^{\text{soft}}_{r_{\overline{\theta}_{(m-1,T)}},\pi_{\overline{\theta}_{(m-1,T)}}}\|_\infty] \\
&\quad - 8\alpha L_q^3 C_d\sqrt{|\mathcal{S}|\cdot|\mathcal{A}|}\cdot\frac{T-1-\frac{1-\gamma^{T-1}}{1-\gamma}}{1-\gamma} - T(T-1)\alpha^2 L_c L_q^2 + \alpha(T-1)\mathbb{E}[\|\nabla L(\overline{\theta}_{(m-1,T)})\|^2] \\
&\quad - \frac{(T-1)^2 L_c L_q^2\alpha^2}{2} \\
&\geq \mathbb{E}[L(\overline{\theta}_{(m-1,T)})] - \frac{4(1-\gamma^{T-1})}{\gamma}\alpha L_q^2 C_d\sqrt{|\mathcal{S}|\cdot|\mathcal{A}|}\mathbb{E}[\|\overline{Q}^{\text{soft}}_{(m-1,T-1)} - Q^{\text{soft}}_{r_{\theta^i_{(m-1,T-2)}},\pi_{\theta^i_{(m-1,T-2)}}}\|_\infty] \\
&\quad + \alpha(T-1)\mathbb{E}[\|\nabla L(\overline{\theta}_{(m-1,T)})\|^2] - 8\alpha L_q^3 C_d\sqrt{|\mathcal{S}|\cdot|\mathcal{A}|}\cdot\frac{T-1-\frac{1-\gamma^{T-1}}{1-\gamma}}{1-\gamma} - T(T-1)\alpha^2 L_c L_q^2 \\
&\quad - \frac{(T-1)^2 L_c L_q^2\alpha^2}{2} - \frac{4(1-\gamma^{T-1})}{1-\gamma}\alpha L_q^2 C_d\sqrt{|\mathcal{S}|\cdot|\mathcal{A}|}\cdot[2(2T-3)\alpha L_q^2 + \frac{1-\gamma}{\gamma}\cdot 4\alpha L_q^2]
\end{aligned}
\tag{72}
$$

Rearranging the inequality above and denote $C_1 = \frac{4(1-\gamma^{T-1})}{\gamma}L_q^2 C_d\sqrt{|\mathcal{S}|\cdot|\mathcal{A}|}$, we obtain:

$$
\begin{aligned}
&\alpha(T-1)\mathbb{E}[\|\nabla L(\overline{\theta}_{(m-1,T)})\|^2] \\
&\leq \alpha C_1\mathbb{E}[\|\overline{Q}^{\text{soft}}_{(m-1,T-1)} - Q^{\text{soft}}_{r_{\theta^i_{(m-1,T-2)}},\pi_{\theta^i_{(m-1,T-2)}}}\|_\infty] + \mathbb{E}[L(\overline{\theta}_{(m,T)}) - L(\overline{\theta}_{(m-1,T)})] \\
&\quad + 8\alpha L_q^3 C_d\sqrt{|\mathcal{S}|\cdot|\mathcal{A}|}\cdot\frac{T-1-\frac{1-\gamma^{T-1}}{1-\gamma}}{1-\gamma} + \frac{(T-1)(3T-1)\alpha^2 L_c L_q^2}{2} \\
&\quad + \frac{4(1-\gamma^{T-1})}{1-\gamma}\alpha L_q^2 C_d\sqrt{|\mathcal{S}|\cdot|\mathcal{A}|}\cdot[2(2T-3)\alpha L_q^2 + \frac{1-\gamma}{\gamma}\cdot 4\alpha L_q^2] \\
&\leq \alpha C_1\mathbb{E}[\|\overline{Q}^{\text{soft}}_{(m-1,T-1)} - Q^{\text{soft}}_{r_{\theta^i_{(m-1,T-2)}},\pi_{\theta^i_{(m-1,T-2)}}}\|_\infty] + \mathbb{E}[L(\overline{\theta}_{(m,T)}) - L(\overline{\theta}_{(m-1,T)})] \\
&\quad + 8\alpha L_q^3 C_d\sqrt{|\mathcal{S}|\cdot|\mathcal{A}|}\cdot\frac{T-1}{1-\gamma} + \frac{(T-1)(3T-1)\alpha^2 L_c L_q^2}{2} \\
&\quad + \frac{16(T-1+\frac{1-\gamma}{\gamma})}{1-\gamma}\alpha^2 L_q^4 C_d\sqrt{|\mathcal{S}|\cdot|\mathcal{A}|}
\end{aligned}
\tag{73}
$$

Summing the inequality above from $m=1$ to M and dividing both sides by $\alpha(T-1)M$, it holds that

$$
\begin{aligned}
&\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}[\|\nabla L(\overline{\theta}_{(m-1,T)})\|^2] \\
&\leq \frac{1}{M(T-1)}\sum_{m=1}^{M}\mathbb{E}[\|\overline{Q}^{\text{soft}}_{(m-1,T-1)} - Q^{\text{soft}}_{r_{\theta^i_{(m-1,T-2)}},\pi_{\theta^i_{(m-1,T-2)}}}\|_\infty] + \mathbb{E}[\frac{L(\overline{\theta}_{(m,T)}) - L(\overline{\theta}_{(0,T)})}{\alpha(T-1)M}] \\
&\quad + \frac{1}{M}\frac{8}{1-\gamma}L_q^3 C_d\sqrt{|\mathcal{S}|\cdot|\mathcal{A}|} + \frac{(3T-1)\alpha}{M}\frac{L_c L_q^2}{2} + \frac{16\alpha(1+\frac{1}{T-1}\frac{1-\gamma}{\gamma})}{M}\frac{L_q^4 C_d\sqrt{|\mathcal{S}|\cdot|\mathcal{A}|}}{1-\gamma}
\end{aligned}
\tag{74}
$$

Since $L(\overline{\theta}_{(m,T)})$ is negative and $L(\overline{\theta}_{(0,T)})$ is bounded constant, we plug equation 63 into equation 74 and get:

$$\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}[\|\nabla L(\overline{\theta}_{(m-1,T)})\|^2]] = \mathcal{O}(M^{-1}) + \mathcal{O}(M^{-\sigma}T^{-\sigma}) + \mathcal{O}(M^{-1-\sigma}T^{1-\sigma}) \tag{75}$$

C EVALUATION

We present the details of the experiment setup and show more convergence plots in MuJoCo tasks.

C.1 EXPERIMENT SETUP

For f-IRL, we utilize the official implementation available at `https://github.com/twni2016/f-IRL`, which also includes implementations for BC and GAIL. The official implementation of ML-IRL can be found at `https://github.com/Cloud0723/ML-IRL`.

To ensure a fair comparison, we use SAC as the base RL algorithm for our F-ML-IRL approach as well as for all baselines, and Adam as the optimizer. Both the Q-network and policy network are configured as 64×64 MLPs with ReLU activation functions, and the learning rate is set to 1×10^{-3}.

For the Ant and Humanoid environments, the reward function is parameterized by a 128×128 MLP with ReLU activation, while for HalfCheetah, Hopper, and Walker2d, a 64×64 MLP with ReLU activation is used. The learning rate for the reward parameter is 1×10^{-4} for Hopper and 1×10^{-3} for the other environments.

At each iteration, we sample 10 trajectories from the current local policy estimate and compare them with the expert demonstration to update the reward parameter.

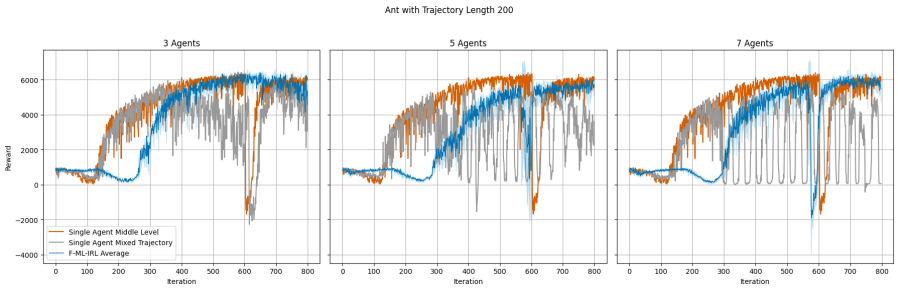
The reward levels of the expert demonstrations are shown in Table C.1. For 3 agents, we use Data 3, 4, and 5; for 5 agents, Data 2, 3, 4, 5, and 6 are used. For 7 agents, all 7 data sets are distributed across different local clients.

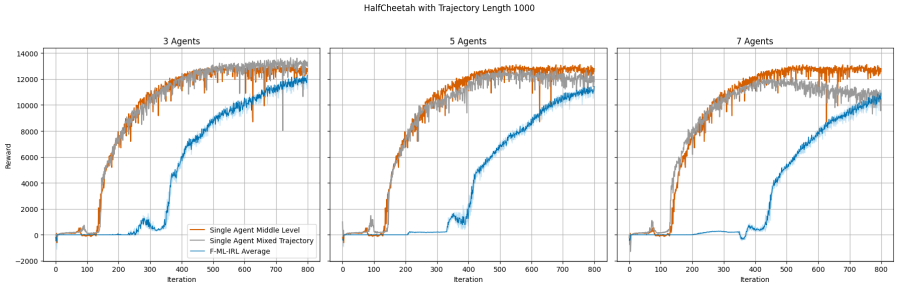
Environment	**Data 1**	**Data 2**	**Data 3**	**Data 4**	**Data 5**	**Data 6**	**Data 7**
Ant	5465.10	5544.83	5699.96	5758.39	5820.41	5927.86	6035.14
HalfCheetah	12831.84	12973.62	13045.36	13187.47	13236.31	13328.38	13434.40
Hopper	3122.05	3217.36	3305.78	3424.81	3553.03	3603.60	3709.89
Humanoid	4934.42	5074.53	5134.65	5297.35	5345.38	5420.32	5501.73
Walker2d	4801.82	4976.41	5081.29	5193.50	5220.25	5379.48	5440.2

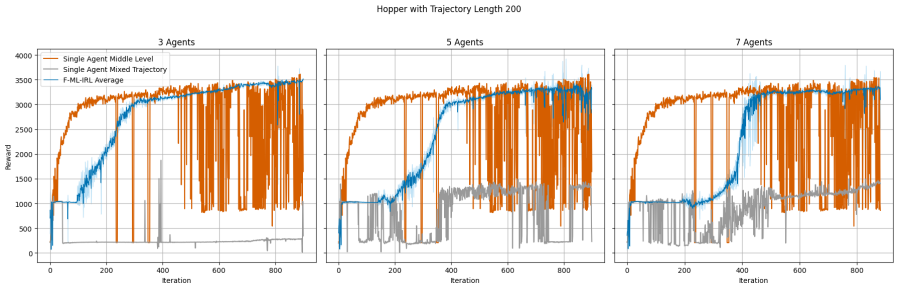
Table 2: The reward levels of 7 expert demonstration datasets that are used in our experiments across 5 MuJoCo tasks. We distribute these non-iid datasets to the clients in our experiments.

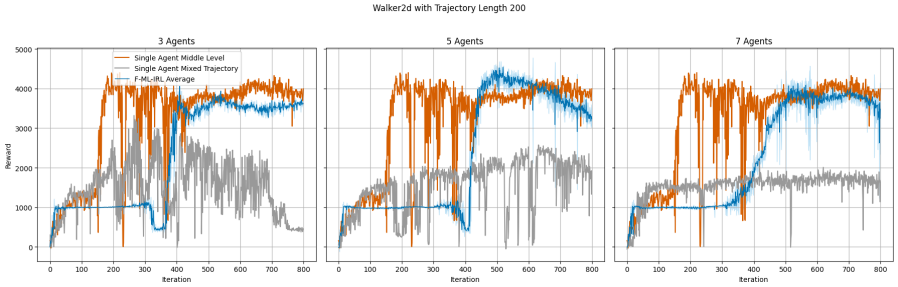
C.2 CONVERGENCE PLOTS

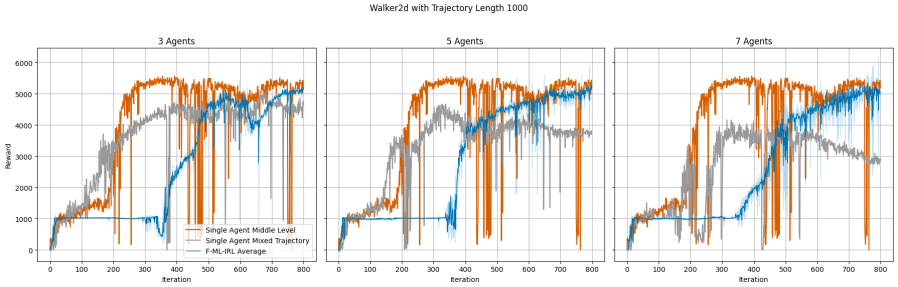
We provide supplementary plots in other settings (different environment and trajectory length) here to show the convergence of F-ML-IRL compared with ML-IRL in two centralized learning data cases:




HalfCheetah with Trajectory Length 1000
3 Agents
5 Agents
7 Agents
Reward
Iteration
Single Agent Middle Level
Single Agent Mixed Trajectory
F-ML-IRL Average


Hopper with Trajectory Length 200
3 Agents
5 Agents
7 Agents
Reward
Iteration
Single Agent Middle Level
Single Agent Mixed Trajectory
F-ML-IRL Average


Walker2d with Trajectory Length 200
3 Agents
5 Agents
7 Agents
Reward
Iteration
Single Agent Middle Level
Single Agent Mixed Trajectory
F-ML-IRL Average


Walker2d with Trajectory Length 1000
3 Agents
5 Agents
7 Agents
Reward
Iteration
Single Agent Middle Level
Single Agent Mixed Trajectory
F-ML-IRL Average