In-Trajectory Inverse Reinforcement Learning: Learn Incrementally Before an Ongoing Trajectory Terminates

Shicheng Liu & Minghui Zhu
School of Electrical Engineering and Computer Science
Pennsylvania State University
University Park, PA 16802, USA
{sfl5539,muz16}@psu.edu

Abstract

Inverse reinforcement learning (IRL) aims to learn a reward function and a corresponding policy that best fit the demonstrated trajectories of an expert. However, current IRL works cannot learn incrementally from an ongoing trajectory because they have to wait to collect at least one complete trajectory to learn. To bridge the gap, this paper considers the problem of learning a reward function and a corresponding policy while observing the initial state-action pair of an ongoing trajectory and keeping updating the learned reward and policy when new state-action pairs of the ongoing trajectory are observed. We formulate this problem as an online bi-level optimization problem where the upper level dynamically adjusts the learned reward according to the newly observed state-action pairs with the help of a meta-regularization term, and the lower level learns the corresponding policy. We propose a novel algorithm to solve this problem and guarantee that the algorithm achieves sub-linear local regret $O(\sqrt{T} + \log T + \sqrt{T}\log T)$. If the reward function is linear, we prove that the proposed algorithm achieves sub-linear regret $O(\log T)$. Experiments are used to validate the proposed algorithm.

1 Introduction

Inverse reinforcement learning (IRL) aims to learn a reward function and a corresponding policy that are consistent with the demonstrated trajectories of an expert. In recent years, several IRL methods are proposed to help learn the reward and policy, including maximum margin methods [1, 2], maximum entropy methods [3, 4], maximum likelihood methods [5, 6], and Bayesian methods [7, 8].

The aforementioned IRL works learn from pre-collected demonstration sets and do not improve the learned model during deployment. Online IRL [9, 10, 11] instead can learn from sequentially arrived demonstrated trajectories and continuously improve the learned reward and policy from the newly observed complete trajectories. However, recent applications of IRL motivate the need to learn incrementally from an ongoing trajectory before it terminates. For example, inferring a moving shooter's intention from its ongoing movement in order to evacuate the hiding victims [12] before the shooter finds them. In this case, we need to quickly update the inference about the shooter's intention once a new movement of the shooter is observed, so that we can use the latest inference to plan a rescue strategy as soon as possible. We cannot wait until the shooter trajectory ends, in case the shooter has found the victims. Another example is learning a target customer's investment preference from its daily updated investment trajectory in a stock market [13] in order to recommend appropriate stocks [14, 15] before other competitors get this customer. However, current IRL works cannot learn from an ongoing trajectory because they have to wait to collect at least one complete trajectory to learn from. To bridge the gap, this paper proposes in-trajectory IRL, a new type of IRL that learns a

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

reward function and a corresponding policy at the initial state-action pair of an ongoing trajectory, and keeps updating the learned reward and policy once a new state-action pair of the ongoing trajectory is observed. We summarize our contributions as follows:

Contribution statement. This paper proposes the first in-trajectory IRL framework, termed **me**ta-**r**egularized **i**n-**t**rajectory inverse reinforcement learning" (MERIT-IRL), to learn a reward function and a corresponding policy from an ongoing trajectory. Our contributions are twofold.

First, we formulate this in-trajectory learning problem as an online bi-level optimization problem where the upper level continuously updates the learned reward according to the newly observed state-action pairs and the lower level computes the corresponding policy. We develop a novel online learning algorithm (MERIT-IRL) to solve this problem. The major novelty of the proposed algorithm is that we propose a novel reward update mechanism specially designed for the in-trajectory learning setting. This special reward update not only aims to explain the expert trajectory observed so far, but also aims to consider for the future. Moreover, since the data is lacking as there is only one ongoing trajectory, we introduce a meta-regularization term to embed prior knowledge and avoid overfitting.

Second, we theoretically guarantee that MERIT-IRL achieves sub-linear local regret $O(\sqrt{T} + \log T + \sqrt{T}\log T)$. If the reward function is linear, we prove that MERIT-IRL achieves sub-linear regret $O(\log T)$. The major novelty of the theoretical analysis is to address the difficulty that the input data is not identically independent distributed (i.i.d.) but temporally correlated, i.e., the data (s_t, a_t) at each time t is affected by the data (s_{t-1}, a_{t-1}) at last time.

2 Related works

Due to the space limit, we only discuss the related works on learning from incomplete trajectories here, and we include the discussion on more related works in Appendix A.

Papers [16, 17] use partial trajectories to update the learned reward function by comparing the expert trajectories after the expert states and the trajectories starting from those expert states rolled out by the learned policy. However, given the current expert state, they use expert trajectory suffix (i.e., future trajectory) to compare while we can only access expert trajectory prefix (i.e., previous trajectory) from an ongoing trajectory. Papers [18, 19, 20] use imitation learning to learn from incomplete demonstrations. In specific, paper [18] uses a discounted sum along the future trajectory as the weight for weighted behavior cloning and works effectively even if only portions of trajectories are observed. Papers [19, 20] extend GAIL [21] to solve for the case where the action sequences are not complete. However, these works all require a pre-collected set of demonstrations so that they are not in-trajectory learning since the trajectory in their cases is not ongoing.

3 Problem Formulation

In this section, we formulate the problem of in-trajectory IRL. In in-trajectory IRL, there is an expert whose decision-making is based on a Markov decision process (MDP). An MDP is a tuple $(\mathcal{S}, \mathcal{A}, \gamma, r_E, P_0, P)$ which consists of a state set $\mathcal{S}$, an action set $\mathcal{A}$, a discount factor $\gamma \in [0, 1)$, a reward function $r_E : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, and the initial state distribution $P_0(\cdot)$. The state transition probability (density) function is denoted by $P(\cdot|\cdot, \cdot)$ such that $P(s'|s, a)$ denotes the probability (density) of state transition to s' from s by taking action a. The expert is using its policy π_E to demonstrate an ongoing trajectory $\zeta^E = S_0^E, A_0^E, S_1^E, A_1^E, \cdots$ and at each time t, only the state-action pair (S_t^E, A_t^E) is observed. We want to learn a reward function and a corresponding policy from the ongoing trajectory and update the learned reward function and policy at each time t.

Many IRL algorithms [5, 6, 11, 22, 23] employ a bi-level learning structure. In this structure, the upper level learns a reward function while the lower level aims to find an associated policy by solving an RL problem under the current learned reward function. Inspired by their bi-level learning structure, we formulate the in-trajectory IRL problem as an online non-convex bi-level optimization problem. In specific, we aim to learn a reward function r_θ (parameterized by θ) in the upper level and a policy corresponding to r_θ in the lower level. The loss function at time t is defined as:

$$L_t(\theta; (S_t^E, A_t^E)) \triangleq -\gamma^t \log \pi_\theta(A_t^E|S_t^E) + \frac{\lambda\gamma^t}{2}||\theta - \bar{\theta}||^2, \quad \pi_\theta = \arg\max_\pi J_\theta(\pi) + H(\pi). \quad (1)$$

Note that $L_t(\theta; (S_t^E, A_t^E))$ is defined using π_θ and θ is the parameter of the reward function r_θ, here the policy π_θ is also parameterized by θ because it is computed by solving an RL problem (in the lower level) under the reward function r_θ, and thus is indirectly parameterized by θ. Maximum likelihood IRL (ML-IRL) [5] has a similar bi-level formulation with (1), however, ML-IRL only solves an offline optimization problem and its analysis does not hold for non i.i.d. input data and continuous state-action space. We discuss our distinctions from ML-IRL in Appendix A.1.

The upper-level loss function L_t has two terms. The first term $-\gamma^t \log \pi_\theta(A_t^E|S_t^E)$ is the discounted negative log-likelihood of the state-action pair (S_t^E, A_t^E) at time t and the second term $\frac{\lambda\gamma^t}{2}||\theta - \bar{\theta}||^2$ is the discounted meta-regularization term [24] where λ is a hyper-parameter. The likelihood function is commonly used in IRL [5, 6] to learn a reward function. Basically, the upper-level loss function at time t encourages to find a reward function r_θ that makes the observed state-action pair (S_t^E, A_t^E) most likely and meanwhile, the reward parameter θ should not be too far from the prior experience, i.e., the meta-prior $\bar{\theta}$. Note that $\bar{\theta}$ is a pre-trained meta-prior that embeds the information of "relevant experience". We will introduce the training of $\bar{\theta}$ in Subsection 4.3 and Appendix C.

The lower-level problem is used to compute π_θ using the current reward function r_θ. It proposes to find a policy π_θ that maximizes the entropy-regularized cumulative reward $J_\theta(\pi) + H(\pi)$. The cumulative reward of a policy π under the reward function r_θ is $J_\theta(\pi) \triangleq E_{S,A}^{\pi}[\sum_{t=0}^{\infty} \gamma^t r_\theta(S_t, A_t)]$ where the initial state is drawn from P_0. The causal entropy of a policy π is defined as $H(\pi) \triangleq E_{S,A}^{\pi}[-\sum_{t=0}^{\infty} \gamma^t \log \pi(A_t|S_t)]$.

Since the expert demonstrates $\{(S_t^E, A_t^E)\}_{t\geq 0}$ sequentially, we have a sequence of loss functions $\{L_t(\theta; (S_t^E, A_t^E))\}_{t\geq 0}$. We use this sequence of loss functions (1) to formulate an online learning problem. A typical online learning problem is to minimize the regret: $\sum_{t=0}^{T-1} L_t(\theta_t; (S_t^E, A_t^E)) - \min_\theta \sum_{t=0}^{T-1} L_t(\theta; (S_t^E, A_t^E))$. However, it is too challenging to minimize the regret in our case because the loss function L_t could be non-convex. Therefore, we aim to minimize the local regret which is widely adopted in online non-convex optimization [25, 26] and online IRL [11]. The local regret quantifies the general stationarity of a sequence of loss functions under the learned parameters. In specific, given a sequence of loss functions $\{f_t(x)\}_{t\geq 0}$, the local regret [11, 25, 26] at time t is defined as $||\frac{1}{t+1}\sum_{i=0}^{t} \nabla f_i(x_t)||^2$ which quantifies the gradient norms of the average of all the previous loss functions under the current learned parameter x_t. The total local regret is defined as the sum of the local regret at each time t, i.e., $\sum_{t=0}^{T-1} ||\frac{1}{t+1}\sum_{i=0}^{t} \nabla f_i(x_t)||^2$. In our case, we replace $\{f_t\}_{t\geq 0}$ with the loss function $\{L_t\}_{t\geq 0}$ defined in (1) and thus formulate the local regret (2)-(3) which has a bi-level formulation. We aim to minimize the following local regret:

$$E_{\{(S_t^E, A_t^E)\sim \mathbb{P}_t^{\pi_E}(\cdot,\cdot)\}_{t\geq 0}}\left[\sum_{t=0}^{T-1} ||\frac{1}{t+1}\sum_{i=0}^{t} \nabla L_i(\theta_t; (S_i^E, A_i^E))||^2\right], \tag{2}$$

$$\text{s.t.} \quad \pi_{\theta_t} = \arg\max_{\pi} J_{\theta_t}(\pi) + H(\pi), \tag{3}$$

where $(S_t^E, A_t^E) \sim \mathbb{P}_t^{\pi_E}(\cdot,\cdot)$ means that (S_t^E, A_t^E) is drawn from the state-action distribution $\mathbb{P}_t^{\pi_E}(\cdot,\cdot)$, and $\mathbb{P}_t^{\pi}(\cdot,\cdot)$ is the state-action distribution induced by π at time t in the MDP.

Difficulties of solving problem (2)-(3). We want to design a fast algorithm to solve problem (2)-(3) since we need to finish the update of θ and π before the next state-action pair is observed and the time between two consecutive state-action pairs can be short. However, designing and analyzing such a fast algorithm is difficult due to the following challenges:

(i) First and foremost, current state-of-the-arts [25, 26] on online non-convex optimization use follow-the-leader-based algorithms which solve $\min_\theta \sum_{i=0}^{t} L_i(\theta; (S_i^E, A_i^E))$ to near stationarity at each time t. This is time-consuming because they require multiple gradient descent updates of θ. One way to alleviate this problem is to use online gradient descent (OGD) which only updates θ by one gradient descent step at each time t. However, since OGD does not solve the problem to near stationarity at any time t, it is extremely difficult to quantify the overall stationarity after T iterations. While OGD has been well studied in online convex optimization, it is rarely studied in online non-convex optimization. The recent work [11] uses OGD to quantify the local regret, however, its analysis can only hold when the input data is i.i.d. In contrast, the input data in our problem is not i.i.d. In specific, the input data at time t (i.e., (S_t^E, A_t^E)) is actually affected by the input data at last step (i.e.,

(S_{t-1}^E, A_{t-1}^E)). This correlation between any two consecutive input data makes it difficult to analyze the growth rate of the local regret.

(ii) Second, it is time-consuming if we fully solve the lower-level problem (3) to get π_θ because this requires multiple policy updates to solve an RL problem. Therefore, we use a “single-loop" method which only requires one-step policy update for a given r_θ. However, since the policy is only updated once, the updated policy can be far from π_θ and thus making the analysis difficult. Single-loop methods are widely adopted to solve hierarchical problems, including bi-level optimization [27, 28], game theory [29, 30], min-max problems [31, 32], etc. Recently, single-loop methods are applied to IRL [5], however, the paper [5] only solves an offline optimization problem and its analysis does not hold for non i.i.d. input data and continuous state-action space. We include a section in Appendix A.1 to discuss our distinctions from [5].

4 Algorithm and Theoretical Analysis

This section has three parts. The first part presents a novel online learning algorithm that solves the problem (2)-(3) and tackles the aforementioned two difficulties. The second part proves that Algorithm 1 achieves sub-linear local regret. If the reward function is linear, we prove that Algorithm 1 achieves sub-linear regret. The third part introduces a meta-learning method to get the meta-prior $\bar{\theta}$.

4.1 The proposed algorithm

In practice, the expert will demonstrate a specific trajectory $s_0^E, a_0^E, s_1^E, a_1^E, \cdots$. For distinction, we use the capital letters (e.g., S) to represent random variables and the lower-case letters (e.g., s) to represent specific values. To design a fast algorithm, we propose an online-gradient-descent-based single-loop algorithm. In specific, at each time t, the algorithm updates both policy π and reward parameter θ only once. The policy update is to solve the lower-level problem (3) and the reward update is to solve the upper-level problem (2). In the following, we elaborate the procedure of policy update and reward update.

Algorithm 1 Meta-regularized In-trajectory Inverse Reinforcement Learning (MERIT-IRL)

```
Input: Initialized policy $\pi_0$, the streaming input data $\{(s_t^E, a_t^E)\}_{t\geq 0}$
Output: Learned reward parameter $\theta_T$ and policy $\pi_T$
1: Compute $\bar{\theta}$ using the meta-regularization in Section 4.3 and Appendix C, and set $\theta_0 = \bar{\theta}$
2: for $t = 0, 1, \cdots, T-1$ do
3:     Compute the soft Q-function $Q^{\text{soft}}_{\theta_t, \pi_t}$ (defined in Appendix B.1) under the current reward function $r_{\theta_t}$ and policy $\pi_t$
4:     Update $\pi_{t+1}(a|s) \propto \exp(Q^{\text{soft}}_{\theta_t, \pi_t}(s, a))$ for any $(s, a) \in \mathcal{S} \times \mathcal{A}$
5:     Roll out policy $\pi_{t+1}$ twice: one starting from $s_0^E$ to get $s_0^E, a'_0, s'_1, a'_1, \cdots$, and the other starting from $(s_t^E, a_t^E)$ to get $s''_{t+1}, a''_{t+1}, \cdots$
6:     Compute $g_t = \sum_{i=0}^{\infty} \gamma^i \nabla_\theta r_{\theta_t}(s'_i, a'_i) - \sum_{i=0}^{\infty} \gamma^i \nabla_\theta r_{\theta_t}(s''_i, a''_i) + \frac{\lambda(1-\gamma^{t+1})}{1-\gamma}(\theta_t - \bar{\theta})$ where $s'_0 = s_0^E$ and $(s''_i, a''_i) = (s_i^E, a_i^E)$ for $0 \leq i \leq t$
7:     Update $\theta_{t+1} = \theta_t - \alpha_t g_t$
8: end for
```

Policy update (lines 3-4 of Algorithm 1). At each time t, we only partially solve the lower-level problem (3) via one-step soft policy iteration [5, 33]. In specific, the soft policy iteration contains two steps: policy evaluation and policy improvement. Policy evaluation aims to compute the soft Q-function $Q^{\text{soft}}_{\theta_t, \pi_t}$ (see the expression in Appendix B.1) under the current learned reward function r_{θ_t} and learned policy π_t. Policy improvement aims to update policy according to $\pi_{t+1}(s, a) \propto \exp(Q^{\text{soft}}_{\theta_t, \pi_t}(s, a))$ for any $(s, a) \in \mathcal{S} \times \mathcal{A}$. In practical implementations, π_{t+1} can be obtained by one-step policy update in soft Q-learning [33] or one-step actor update in soft actor-critic [34].

Reward update (lines 5-7 of Algorithm 1). At each time t, the algorithm observes (s_t^E, a_t^E) and aims to leverage all the previously observed data to update the reward parameter. In specific, as $L_t(\theta; (s_t^E, a_t^E)) = -\gamma^t \log \pi_\theta(a_t^E|s_t^E) + \frac{\lambda\gamma^t}{2}||\theta - \bar{\theta}||^2$ is the meta-regularized negative log-likelihood of (s_t^E, a_t^E), the algorithm can formulate $\sum_{i=0}^{t} L_i(\theta; (s_i^E, a_i^E))$ at time t using all the previously collected data (i.e., $\{(s_i^E, a_i^E)\}_{i=0,\cdots,t}$). To update the reward parameter, the algorithm partially minimizes $\sum_{i=0}^{t} L_i(\theta; (s_i^E, a_i^E))$ via one-step gradient descent.

Lemma 1. *The gradient of* $\sum_{i=0}^{t} L_i(\theta; (s_i^E, a_i^E))$ *can be calculated as follows:*

$$\nabla \sum_{i=0}^{t} L_i(\theta; (s_i^E, a_i^E)) = E_{S,A}^{\pi_\theta}\left[\sum_{i=0}^{\infty} \gamma^i \nabla_\theta r_\theta(S_i, A_i)\middle| S_0 = s_0^E\right] + \frac{\lambda(1-\gamma^{t+1})}{1-\gamma}(\theta - \bar{\theta})$$
$$- \sum_{i=0}^{t} \gamma^i \nabla_\theta r_\theta(s_i^E, a_i^E) - E_{S,A}^{\pi_\theta}\left[\sum_{i=t+1}^{\infty} \gamma^i \nabla_\theta r_\theta(S_i, A_i)\middle| S_t = s_t^E, A_t = a_t^E\right]. \tag{4}$$

Note that the gradient (4) holds for continuous state-action space. Since the gradient (4) has expectation terms under the policy π_θ, we can only approximate it. In specific, we roll out the policy π_{t+1} twice: one starting from s_0^E to get a trajectory $\{(s_i', a_i')\}_{i\geq 0}$ where $s_0' = s_0^E$, and the other one starting from (s_t^E, a_t^E) to get a trajectory $\{(s_i'', a_i'')\}_{i\geq 0}$ where $(s_i'', a_i'') = (s_i^E, a_i^E)$ for $0 \leq i \leq t$. Then we use the empirical estimate $g_t = \sum_{i=0}^{\infty} \gamma^i \nabla_\theta r_{\theta_t}(s_i', a_i') - \sum_{i=0}^{\infty} \gamma^i \nabla_\theta r_{\theta_t}(s_i'', a_i'') + \frac{\lambda(1-\gamma^{t+1})}{1-\gamma}(\theta_t - \bar{\theta})$ to approximate $\nabla \sum_{i=0}^{t} L_i(\theta_t; (s_i^E, a_i^E))$. With the gradient approximation g_t, we utilize stochastic online gradient descent $\theta_{t+1} = \theta_t - \alpha_t g_t$ to update the reward parameter.

Discussion on our special design of the reward update. The right subfigure in Figure 1 visualizes our reward update (modulo the meta-regularization term). The green trajectory (i.e., $\{(s_t^E, a_t^E)\}_{t\geq 0}$) is the expert trajectory, and the red trajectories (i.e., $\{(s_t', a_t')\}_{t\geq 0}$ and $\{(s_i'', a_i'')\}_{i>t}$) are the trajectories generated by the learned policy. Given the expert trajectory prefix (i.e., the incomplete trajectory $\{(s_i^E, a_i^E)\}_{i=0}^{t}$ observed so far), our method completes the expert trajectory by rolling out the learned policy starting from (s_t^E, a_t^E) and filling the trajectory suffix $\{(s_i'', a_i'')\}_{i>t}$. The combined complete trajectory includes the expert trajectory prefix $\{(s_i^E, a_i^E)\}_{i=0}^{t}$ and the learner-filled trajectory suffix $\{(s_i'', a_i'')\}_{i\geq t}$. We update the reward function by comparing this combined trajectory to a complete trajectory $\{(s_t', a_i')\}_{t\geq 0}$ generated by the learned policy starting from the expert's initial state s_0^E.

A more straightforward way for the reward update is to directly compare the trajectory prefixes (visualized in the middle of Figure 1) at each time t. However, this naive method can be problematic. We explain the issue of this naive method and the advantage of our method in the following context.

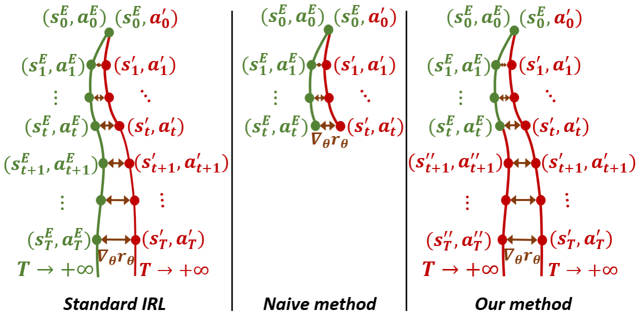


Figure 1: Standard IRL (left), the naive method for in-trajectory learning (middle), and our method (right).

Figure 1 visualizes the reward update for standard IRL (left), the naive method (i.e., directly run standard IRL algorithms on the expert trajectory prefix) (middle), and our method (right). The standard IRL (left) updates the reward function by comparing the complete expert trajectory and the complete trajectory generated by the learned policy. This case is ideal, however, it is infeasible when the trajectory is ongoing and we can only observe an incomplete expert trajectory $\{(s_i^E, a_i^E)\}_{i=0}^{t}$ at each time t. The naive method (middle) updates the reward function by simply comparing the expert trajectory prefix $\{(s_i^E, a_i^E)\}_{i=0}^{t}$ and the incomplete trajectory $\{(s_i', a_i')\}_{i=0}^{t}$ with the same length generated by the learned policy starting from the expert's initial state s_0^E. This kind of reward update is myopic as it does not consider for the future. Since it runs standard IRL on the trajectory prefix observed so far, it will always regard the current state-action pair as the terminal state-action pair and thus has no ability to consider the state-action pairs in the future. In contrast, our special design gives the algorithm the ability to consider for the future. Instead of only comparing incomplete trajectory prefixes, our method compares complete trajectories just as the standard IRL. In specific, we compare the combined complete trajectory $\{\{(s_i^E, a_i^E)\}_{i=0}^{t}, \{(s_i'', a_i'')\}_{i>t}\}$ and the complete trajectory $\{(s_i', a_i')\}_{i\geq 0}$ solely generated by the learned policy. The comparison between the learner prefix $\{s_i', a_i'\}_{i=0}^{t}$ and expert prefix $\{s_i^E, a_i^E\}_{i=0}^t$ encourages the learned reward function to explain the expert's demonstrated behaviors so far, and the comparison between the suffixes ($\{s_i', a_i'\}_{i>t}$ and $\{s_i'', a_i''\}_{i>t}$) encourages that we are learning a reward function that is useful for predicting the future. Note that as t increases, the expert trajectory prefix weights more and more in the combined complete trajectory, and eventually we will recover the standard IRL reward update when t goes to infinity. In Theorems 1 and 2, we theoretically guarantee that the proposed reward update can achieve sub-linear (local) regret. This shows the perfect consistency between the intuition and theory.

4.2 Theoretical analysis

To quantify the local regret of Algorithm 1, we have two challenges: (i) Since we only update π by one step at each time t, we have π_{t+1} instead of the optimal solution π_{θ_t} of the lower-level problem (3). The policy π_{t+1} can be far away from π_{θ_t} and thus the empirical gradient estimate g_t can be a bad approximation of the gradient (4). (ii) Since we only update θ once at each time t instead of finding a near-stationary point θ' such that $||\sum_{i=0}^t L_i(\theta'; (s_i^E, a_i^E))|| \leq \epsilon$ as in [25, 26], the gradient norm $||\sum_{i=0}^t \nabla L_i(\theta_t; (s_i^E, a_i^E))||$ is not stabilized under the threshold ϵ at every time t. Therefore the local regret (i.e., $\sum_{t=0}^{T-1} ||\frac{1}{t+1}\sum_{i=0}^t \nabla L_i(\theta_t; (s_i^E, a_i^E))||^2$) is hard to quantify and may not be sub-linear in T. What's worse, the input data is not i.i.d. but correlated, i.e., the input data (s_t^E, a_t^E) at time t is affected by the input data (s_{t-1}^E, a_{t-1}^E) at last step. This correlation makes it even more difficult to quantify the local regret.

To solve the first challenge, we adopt the idea of two-timescale stochastic approximation [27] where the lower level updates in a faster timescale and the upper level updates in a slower timescale. The policy update is faster because it converges linearly under a fixed reward function [35] while the reward update is slower given that we choose $\alpha_t \propto (t+1)^{-1/2}$. Intuitively, since the policy update is faster than the reward update, the reward parameter is "relatively fixed" compared to the policy. It is expected that π_{t+1} shall stay close to π_{θ_t} and at last converges to π_{θ_t} when t increases.

To solve the second challenge, we divide our analysis into two steps: (i) We quantify the difference of the gradient norms between the current correlated state-action distribution $\mathbb{P}_t^{\pi_E}(\cdot,\cdot)$ and a stationary state-action distribution for any loss function L_i, $i \geq 0$. Note that $\mathbb{P}_{t+1}^{\pi_E}(\cdot,\cdot)$ is affected by $\mathbb{P}_t^{\pi_E}(\cdot,\cdot)$. (ii) We quantify the local regret under the stationary distribution. The benefit of doing so is that the input data is i.i.d. under the stationary distribution, and thus we can cast the online gradient descent method as a stochastic gradient descent method and quantify its local regret. Finally, we can quantify the local regret under the current correlated distribution $\mathbb{P}_t^{\pi_E}(\cdot,\cdot)$ by combining (i) and (ii).

We start our analysis with the definitions of stationary state distribution and stationary state-action distribution. For a given policy π, the corresponding stationary state distribution is $\mu^\pi(s) \triangleq (1-\gamma)\sum_{t=0}^\infty \gamma^t \mathbb{P}_t^\pi(s)$ and the stationary state-action distribution is $\mu^\pi(s,a) \triangleq (1-\gamma)\sum_{t=0}^\infty \gamma^t \mathbb{P}_t^\pi(s,a)$.

Assumption 1. *The parameterized reward function r_θ satisfies $|r_{\theta_1}(s,a) - r_{\theta_2}(s,a)| \leq \bar{C}_r ||\theta_1 - \theta_2||$ and $||\nabla_\theta r_{\theta_1}(s,a) - \nabla_\theta r_{\theta_2}(s,a)|| \leq \tilde{C}_r ||\theta_1 - \theta_2||$ for any (θ_1, θ_2) and any $(s,a) \in \mathcal{S} \times \mathcal{A}$ where $\bar{C}_r$ and $\tilde{C}_r$ are positive constants.*

Assumption 2 (Ergodicity)**.** *There exist constants $C_M > 0$ and $\rho \in (0,1)$ such that for any policy π and any $t \geq 0$, the following holds for the Markov chain induced by the policy π and the state transition function P: $\sup_{S_0 \sim P_0} d_{TV}(\mathbb{P}_t^\pi(\cdot), \mu^\pi(\cdot)) \leq C_M \rho^t$ where $d_{TV}(\mathbb{P}_1(\cdot), \mathbb{P}_2(\cdot)) \triangleq \frac{1}{2}\int_{s \in \mathcal{S}} |\mathbb{P}_1(s) - \mathbb{P}_2(s)| ds$ is the total variation distance between the two state distributions $\mathbb{P}_1$ and $\mathbb{P}_2$, S_0 is the initial state, and $\mathbb{P}_t^\pi(\cdot)$ is the state distribution induced by the policy π at time t.*

Assumptions 1-2 are common in RL [36, 37, 38, 39]. Assumption 2 holds for any time-homogeneous Markov chain with finite state space or any uniformly ergodic Markov chain with general state space.

Proposition 1. *Suppose Assumptions 1-2 hold and $\alpha_t \in (0, \frac{1-\gamma}{\lambda})$, we have the following relation for any $i \geq 0$ and any $\theta_t, t \geq 0$:*

$$\left|E_{(S_i^E, A_i^E) \sim \mathbb{P}_i^{\pi_E}(\cdot,\cdot)}\left[||\nabla L_i(\theta_t; (S_i^E, A_i^E))||^2\right] - E_{(S_i^E, A_i^E) \sim \mu^{\pi_E}(\cdot,\cdot)}\left[||\nabla L_i(\theta_t; (S_i^E, A_i^E))||^2\right]\right|,$$
$$\leq 8 C_M \bar{C}_r^2 \Big(\frac{2-\gamma}{1-\gamma}\Big)^2 \rho^i \gamma^{2i},$$

where $(S_i^E, A_i^E) \sim \mathbb{P}_i^{\pi_E}(\cdot,\cdot)$ means that (S_i^E, A_i^E) is drawn from the correlated distribution $\mathbb{P}_i^{\pi_E}(\cdot,\cdot)$ and $(S_i^E, A_i^E) \sim \mu^{\pi_E}(\cdot,\cdot)$ means that (S_i^E, A_i^E) is drawn from the stationary distribution $\mu^{\pi_E}(\cdot,\cdot)$.

Proposition 1 quantifies the gap of gradient norms between the current correlated distribution $\mathbb{P}_i^{\pi_E}$ and the stationary distribution μ^{π_E}. We next quantify the local regret under the stationary distribution μ^{π_E} with the following lemma:

Lemma 2. *Suppose Assumptions 1-2 hold and choose $\alpha_t = \frac{(1-\gamma)(t+1)^{-1/2}}{\lambda}$, it holds that:*

$$E_{\{(S_t^E, A_t^E)\sim\mu^{\pi_E}(\cdot,\cdot)\}_{t\geq 0}}\left[\sum_{t=0}^{T-1}||\frac{1}{t+1}\sum_{i=0}^{t}\nabla L_i(\theta_t;(S_i^E,A_i^E))||^2\right]$$
$$\leq D_1(\log T+1)+D_2\sqrt{T}+D_3\sqrt{T}(\log T+1),$$

where D_1, D_2, and D_3 are positive constants whose expressions can be found in Appendix B.4.

Lemma 2 quantifies the local regret under the stationary distribution μ^{π_E}. With Proposition 1 and Lemma 2, we can quantify the local regret under the current correlated distribution.

Theorem 1. *Suppose Assumptions 1-2 hold and choose $\alpha_t = \frac{(1-\gamma)(t+1)^{-1/2}}{\lambda}$, we have that:*

$$E_{\{(S_t^E, A_t^E)\sim\mathbb{P}_t^{\pi_E}(\cdot,\cdot)\}_{t\geq 0}}\left[\sum_{t=0}^{T-1}||\frac{1}{t+1}\sum_{i=0}^{t}\nabla L_i(\theta_t;(S_i^E,A_i^E))||^2\right]$$
$$\leq\left(D_1+\frac{8C_M\bar{C}_r^2(2-\gamma)^2}{(1-\rho\gamma^2)(1-\gamma)^2}\right)(\log T+1)+D_2\sqrt{T}+D_3\sqrt{T}(\log T+1).$$

Theorem 1 is based on Proposition 1 and Lemma 2. It shows that Algorithm 1 achieves sub-linear local regret. Moreover, if the reward function is linear, Algorithm 1 achieves sub-linear regret:

Theorem 2. *Suppose the expert reward function r_E and the parameterized reward r_θ are linear, and Assumptions 1-2 hold. Choose $\alpha_t = \frac{1-\gamma}{\lambda(t+1)(1-\gamma^{t+1})}$ we have that:*

$$E_{\{(S_t^E, A_t^E)\sim\mathbb{P}_t^{\pi_E}(\cdot,\cdot)\}_{t\geq 0}}\left[\sum_{t=0}^{T-1}L_t(\theta_t;(S_t^E,A_t^E))\right]$$
$$-\min_\theta E_{\{(S_t^E, A_t^E)\sim\mathbb{P}_t^{\pi_E}(\cdot,\cdot)\}_{t\geq 0}}\left[\sum_{t=0}^{T-1}L_t(\theta;(S_t^E,A_t^E))\right]\leq D_4+D_5(\log T+1),$$

where D_4 and D_5 are positive constants whose expressions are in Appendix B.6.

4.3 Meta-Regularization

Since there is only one training trajectory and this trajectory is not complete during the learning process, we need to add a regularization term to avoid overfitting. Inspired by humans' using relevant experience to help do inference, we introduce the meta-regularization $\frac{\lambda}{2}||\theta-\bar{\theta}||^2$ where λ is the hyper-parameter and the meta-prior $\bar{\theta}$ is learned from "relevant experience". In specific, we introduce a set of relevant tasks $\{\mathcal{T}_j\}_{j\sim P_\mathcal{T}}$ where each task $\mathcal{T}_j$ is an IRL problem and $P_\mathcal{T}$ is the implicit task distribution. The tasks $\{\mathcal{T}_j\}_{j\sim P_\mathcal{T}}$ are relevant in the sense that they share the components $(\mathcal{S},\mathcal{A},\gamma,P_0,P)$ of the MDP with our in-trajectory learning problem. However, the expert's reward functions of different tasks are different and are drawn from an unknown reward function distribution. For example, in the stock market case mentioned in the introduction, the experts of different tasks invest in the same stock market but may have different preferences. As standard in meta-learning [40, 41, 42], we assume that the expert's reward function r_E of our in-trajectory learning problem is also drawn from the same unknown reward function distribution. Note that the reward functions of the relevant tasks $\{\mathcal{T}_j\}_{j\sim P_\mathcal{T}}$ are different from r_E even if they are drawn from the same unknown reward function distribution.

For each task $\mathcal{T}_j$, there is a batch of trajectories and we divide this batch into two sets, i.e., $\mathcal{D}_j^{\text{tr}}$ and $\mathcal{D}_j^{\text{eval}}$. The training set $\mathcal{D}_j^{\text{tr}}$ only has one trajectory, just as our in-trajectory learning problem, and the evaluation set $\mathcal{D}_j^{\text{eval}}$ has abundant trajectories. Define the loss function on a certain data set $\mathcal{D}\triangleq\{\zeta^v\}_{v=1}^m$ as $L(\theta,\mathcal{D})\triangleq-\sum_{v=1}^m\sum_{t=0}^\infty\gamma^t\log\pi_\theta(a_t^v|s_t^v)$. The goal of each task $\mathcal{T}_j$ is to learn a task-specific adaptation ϕ_j using the training set $\mathcal{D}_j^{\text{tr}}$, such that ϕ_j can minimize the test loss $L(\phi_j, \mathcal{D}_j^{\text{eval}})$ on the evaluation set $\mathcal{D}_j^{\text{eval}}$. The goal of meta-regularization is to find a meta-prior $\bar{\theta}$, from which such task-specific adaptations ϕ_j can be adapted to all tasks $\{\mathcal{T}_j\}_{j \sim P_{\mathcal{T}}}$. In specific, meta-regularization [24] proposes a bi-level optimization problem (5). The lower-level problem uses only one trajectory $\mathcal{D}_j^{\text{tr}}$ to find the task-specific adaptation ϕ_j such that the meta-regularized loss function $L(\phi, \mathcal{D}_j^{\text{tr}}) + \frac{\lambda}{2(1-\gamma)}||\phi - \bar{\theta}||^2$ is minimized. The upper-level problem is to find a meta-prior $\bar{\theta}$ such that the corresponding task-specific adaptations $\{\phi_j\}_{j \sim P_{\mathcal{T}}}$ can minimize the expected loss function $L(\phi_j, \mathcal{D}_j^{\text{eval}})$ over the evaluation sets of all tasks $\{\mathcal{T}_i\}_{j \sim P_{\mathcal{T}}}$.

$$\min_{\bar{\theta}} \ E_{j \sim P_{\mathcal{T}}}\big[L(\phi_j, \mathcal{D}_j^{\text{eval}})\big], \quad \text{s.t. } \phi_j = \arg\min_{\phi} L(\phi, \mathcal{D}_j^{\text{tr}}) + \frac{\lambda}{2(1-\gamma)}||\phi - \bar{\theta}||^2. \tag{5}$$

The lower-level loss function in (5) is the offline version of our in-trajectory loss function (1) (i.e., $L(\phi, \mathcal{D}_j^{\text{tr}}) + \frac{\lambda}{2(1-\gamma)}||\phi - \bar{\theta}||^2 = \sum_{t=0}^{\infty} L_t(\phi; (s_t^{\text{tr}}, a_t^{\text{tr}}))$) where $(s_t^{\text{tr}}, a_t^{\text{tr}}) \in \mathcal{D}_j^{\text{tr}}$. Our in-trajectory learning problem can also be regarded as to find a task-specific adaptation. Note that the in-trajectory learning problem is online while the lower-level problem in (5) is offline because the in-trajectory problem is ongoing where we keep observing new state-action pairs. In contrast, the lower-level problem in (5) is based on "experience" that has already happened. Due to the space limit, we include the algorithm and theoretical guarantees of solving the problem (5) in Appendix C.

5 Experiments

We present three experiments to show the effectiveness of MERIT-IRL. We use four baselines for comparisons. (i) **IT-IRL**: this method is MERIT-IRL without meta-regularization. (ii) **Naive MERIT-IRL**: this method has the meta-regularization term but uses the naive way (in the middle of Figure 1) to update reward. (iii) **Naive IT-IRL**: this method uses the naive way to update the learned reward and does not have the meta-regularization term. (iv) **Hindsight**: this method is meta-regularized ML-IRL [5] which can access the complete expert trajectory and uses the standard IRL (visualized in the left of Figure 1) with meta-regularization to update the learned reward. The experiment details are in Appendix D.

5.1 MuJoCo experiment

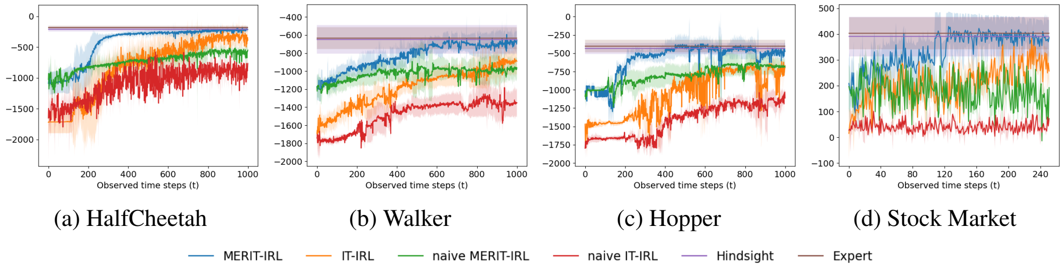
In this subsection, we consider the target velocity problem for three MuJoCo robots: HalfCheetah, Walker, and Hopper. The target velocity problem is widely used in meta-RL [43] and meta-IRL [44]. In specific, the robots aim to maintain a target velocity in each task and the target velocity of different tasks is different. To test the performance of MERIT-IRL, we use 10 test tasks whose target velocity is randomly between 1.5 and 2.0. In the test tasks, there is only one expert trajectory and the state-action pairs of this trajectory are sequentially revealed (to MERIT-IRL, IT-IRL, Naive MERIT-IRL, and Naive IT-IRL) in an online fashion. The baseline Hindsight uses the complete expert trajectory to learn a reward function. The ground truth reward is designed as $-|v - v_{\text{target}}|$ (as in [43]) where v is the current robot velocity and v_{target} is the target velocity. To learn the meta-prior $\bar{\theta}$, we use 50 relevant tasks whose target velocity is randomly between 0 and 3.

Figures 2a-2c show the in-trajectory learning performance where the x-axis is the time step t of the expert trajectory and the y-axis is the cumulative reward of the learned policy π_t when only the first t steps of the expert trajectory are observed. The x-limit is $1,000$ because the trajectory length in MuJoCo is $1,000$. Note that the baseline "Hindsight" is not in-trajectory learning since it learns from a complete expert trajectory. For comparison, we use two horizontal lines (close to each other) to show the performance of Hindsight and the expert in the figures. Figure 2a shows that MERIT-IRL achieves similar performance with the expert when only 40% of the complete expert trajectory ($t = 400$) is observed while IT-IRL can only achieve performance close to the expert after observing more than 90% of the complete expert trajectory ($t = 900$). This shows the effectiveness of the meta-regularization. Naive MERIT-IRL and Naive IT-IRL fail to imitate the expert even if the complete expert trajectory is observed ($t = 1,000$). This shows the effectiveness of our special design of the reward update. The discussions on Figures 2b and 2c are in Appendix D.2.

Table 1 shows the results after observing the complete expert trajectory. MERIT-IRL performs much better than IT-IRL, Naive MERIT-IRL, and Naive IT-IRL. MERIT-IRL achieves similar performance

Table 1: Experiment results. The mean and standard deviation are calculated from 10 test tasks.

	MERIT-IRL	IT-IRL	Naive MERIT-IRL	Naive IT-IRL	Hindsight	Expert
HalfCheetah	-214.63 ± 53.96	-386.78 ± 152.65	-548.22 ± 40.51	-765.27 ± 104.41	-208.74 ± 37.23	-181.51 ± 28.35
Walker	-654.77 ± 102.59	-891.79 ± 156.90	-962.42 ± 111.60	-1349.25 ± 158.88	-648.17 ± 157.92	-634.17 ± 120.57
Hopper	-476.72 ± 32.09	-669.88 ± 53.63	-691.03 ± 93.35	-1112.06 ± 74.33	-455.70 ± 74.93	-421.74 ± 84.30
Stock market	386.70 ± 62.95	256.81 ± 68.61	192.49 ± 75.34	72.33 ± 16.73	390.30 ± 77.37	403.15 ± 61.94



(a) HalfCheetah (b) Walker (c) Hopper (d) Stock Market

Figure 2: In-trajectory learning performance.

with Hindsight and expert. Note that it is not expected that MERIT-IRL outperforms Hindsight since Hindsight uses the complete expert trajectory to learn.

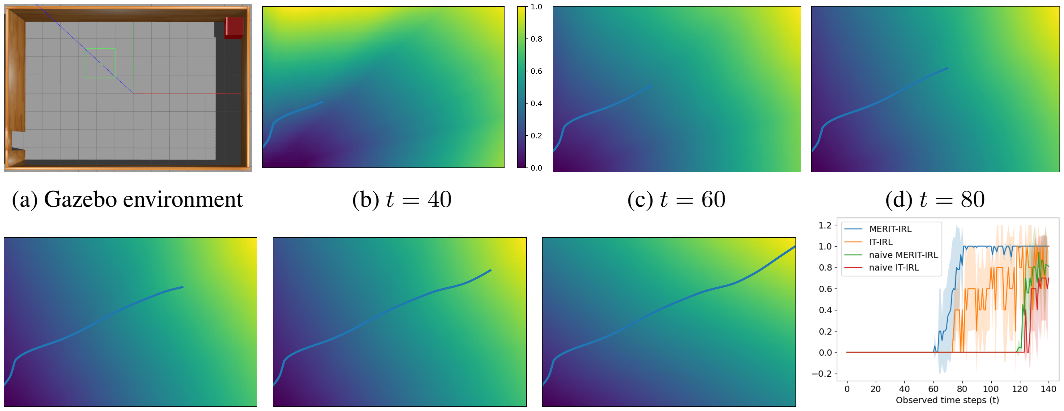
5.2 Stock market experiment

RL to train a stock trading agent has been widely studied in AI for finance [45, 46, 47]. In this experiment, we use IRL to learn the reward function (i.e., investing preference) of the target investor in a stock market scenario. In specific, we use the real-world data of 30 constituent stocks in Dow Jones Industrial Average from 2021-01-01 to 2022-01-01. We use a benchmark called “FinRL" [47] to configure the real-world stock data into an MDP environment. The target investor (i.e., expert) has an initial asset of $\$1,000$ and trades stocks on every stock market opening day. The stock market opens 252 days between 2021-01-01 and 2022-01-01, and thus the trajectory length is 252. The reward function of the target investor is defined as $p_1 - p_2$ where p_1 is the investor's profit which is the money earned from trading stocks subtracting the transaction cost, and p_2 models the investor's preference of whether willing to take risks. In specific, p_2 is positive if the investor buys stocks whose turbulence indices are larger than a certain turbulence threshold, and zero otherwise. The value of p_2 depends on the type and amount of the trading stocks. The turbulence thresholds of different investors are different. The turbulence index measures the price fluctuation of a stock. If the turbulence index is high, the corresponding stock has a high fluctuating price and thus is risky to buy [47]. Therefore, an investor unwilling to take risks has a relatively low turbulence threshold. We include experiment details in Appendix D.3. To test performance, we use 10 test tasks whose turbulence thresholds are randomly between 45 and 50. To learn $\bar{\theta}$, we use 50 relevant tasks whose turbulence thresholds are randomly between 30 and 60.

Figure 2d shows that MERIT-IRL achieves similar cumulative reward with the expert at $t = 140$ which is less than 60% of the whole trajectory, while the three in-trajectory baselines fail to imitate the expert before the ongoing trajectory terminates. The last row in Table 1 shows that MERIT-IRL achieves similar performance with Hindsight and the expert. More discussions on the results are in Appendix D.3.

5.3 Learning from a shooter's ongoing trajectory

This part presents the experiment of learning from an ongoing shooter trajectory. Following [12], we model the shooter's movement as a navigation problem. We build a simulator in Gazebo (Figure 3a) where the shooter moves from the door (lower left corner) to the red target (upper right corner). The learner observes the ongoing trajectory of the shooter and keeps updating the learned reward and policy. In our case, the complete shooter trajectory has the length of 140. Figures 3b-3g show our in-trajectory learning performance where the heat maps visualize the learned reward. We normalize the learned reward to $[0, 1]$. We can observe that as the ongoing trajectory is expanding, the learned reward function becomes more and more precise to locate the goal area. When $t = 40$, we cannot tell the goal area from the heat map (Figure 3b). However, as the time t grows, we can almost locate the goal area when $t = 60$ (Figure 3c) and precisely locate the goal area when $t = 80$ (Figure 3d).



(a) Gazebo environment (b) $t = 40$ (c) $t = 60$ (d) $t = 80$

(e) $t = 100$ (f) $t = 120$ (g) $t = 140$ (h) Success rate

Figure 3: Learning performance on the active shooting scenario.

Figure 3h shows the policy learning performance. Since there is no ground truth reward in this problem, we use “success rate" to quantify the performance of the learned policy. The success rate is the rate that the learned policy successfully reaches the goal. From 3h, we can see that MERIT-IRL outperforms the other baselines and can achieve 100% success rate when $t = 80$ (i.e., only observing 57% of the complete trajectory). Note that we do not include Hindsight and Expert in Figure 3 since they both achieve 100% success rate.

6 Conclusion

This paper proposes MERIT-IRL, the first in-trajectory inverse reinforcement learning theoretical framework that learns a reward function while observing an initial portion of a trajectory and keeps updating the learned reward function when extended portions (i.e., new state-action pairs) of the trajectory are observed. Experiments show that MERIT-IRL can imitate the expert from the ongoing expert trajectory before it terminates.

7 Acknowledgements

This work is partially supported by the National Science Foundation through grants ECCS 1846706 and ECCS 2140175. We would like to thank the reviewers for their insightful and constructive suggestions.

References

[1] P. Abbeel and A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," in *International Conference on Machine Learning*, pp. 1–8, 2004.

[2] N. D. Ratliff, J. A. Bagnell, and M. A. Zinkevich, "Maximum margin planning," in *International Conference on Machine Learning*, pp. 729–736, 2006.

[3] B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey, "Maximum entropy inverse reinforcement learning," in *National Conference on Artificial intelligence*, pp. 1433–1438, 2008.

[4] B. D. Ziebart, J. A. Bagnell, and A. K. Dey, "Modeling interaction via the principle of maximum causal entropy," in *International Conference on Machine Learning*, pp. 1255–1262, 2010.

[5] S. Zeng, C. Li, A. Garcia, and M. Hong, "Maximum-likelihood inverse reinforcement learning with finite-time guarantees," in *Advances in Neural Information Processing Systems*, 2022.

[6] S. Liu and M. Zhu, "Distributed inverse constrained reinforcement learning for multi-agent systems," *Advances in Neural Information Processing Systems*, vol. 35, pp. 33444–33456, 2022.

[7] D. Ramachandran and E. Amir, "Bayesian inverse reinforcement learning.," in *International Joint Conference on Artificial Intelligence*, pp. 2586–2591, 2007.

[8] A. J. Chan and M. van der Schaar, "Scalable bayesian inverse reinforcement learning," in *International Conference on Learning Representations*, 2021.

[9] N. Rhinehart and K. M. Kitani, "First-person activity forecasting with online inverse reinforcement learning," in *IEEE International Conference on Computer Vision*, pp. 3696–3705, 2017.

[10] S. Arora, P. Doshi, and B. Banerjee, "Online inverse reinforcement learning under occlusion," in *International Conference on Autonomous Agents and Multiagent Systems*, pp. 1170–1178, 2019.

[11] S. Liu and M. Zhu, "Learning multi-agent behaviors from distributed and streaming demonstrations," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[12] A. Aghalari, N. Morshedlou, M. Marufuzzaman, and D. Carruth, "Inverse reinforcement learning to assess safety of a workplace under an active shooter incident," *IISE Transactions*, vol. 53, no. 12, pp. 1337–1350, 2021.

[13] Q. Sun, X. Gong, and Y.-W. Si, "Transaction-aware inverse reinforcement learning for trading in stock markets," *Applied Intelligence*, vol. 53, no. 23, pp. 28186–28206, 2023.

[14] J. Chang and W. Tu, "A stock-movement aware approach for discovering investors' personalized preferences in stock markets," in *International Conference on Tools with Artificial Intelligence*, pp. 275–280, 2018.

[15] X. Hu, Y. Chen, L. Ren, and Z. Xu, "Investor preference analysis: An online optimization approach with missing information," *Information Sciences*, vol. 633, pp. 27–40, 2023.

[16] Y. Xu, W. Gao, and D. Hsu, "Receding horizon inverse reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27880–27892, 2022.

[17] G. Swamy, D. Wu, S. Choudhury, D. Bagnell, and S. Wu, "Inverse reinforcement learning without reinforcement learning," in *International Conference on Machine Learning*, pp. 33299–33318, 2023.

[18] K. Yan, A. Schwing, and Y.-X. Wang, "A simple solution for offline imitation from observations and examples with possibly incomplete trajectories," *Advances in Neural Information Processing Systems*, 2024.

[19] M. Sun and X. Ma, "Adversarial imitation learning from incomplete demonstrations," in *International Joint Conference on Artificial Intelligence*, pp. 3513–3519, 2019.

[20] D. Xu, F. Zhu, Q. Liu, and P. Zhao, “Arail: Learning to rank from incomplete demonstrations,” *Information Sciences*, vol. 565, pp. 422–437, 2021.

[21] J. Ho and S. Ermon, “Generative adversarial imitation learning,” in *Advances in Neural Information Processing Systems*, pp. 4572–4580, 2016.

[22] G. Qiao, G. Liu, P. Poupart, and Z. Xu, “Multi-modal inverse constrained reinforcement learning from a mixture of demonstrations,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[23] S. Liu and M. Zhu, “Meta inverse constrained reinforcement learning: Convergence guarantee and generalization analysis,” in *International Conference on Learning Representations*, 2023.

[24] A. Rajeswaran, C. Finn, S. M. Kakade, and S. Levine, “Meta-learning with implicit gradients,” in *Advances in Neural Information Processing Systems*, pp. 113–124, 2019.

[25] E. Hazan, K. Singh, and C. Zhang, “Efficient regret minimization in non-convex games,” in *International Conference on Machine Learning*, pp. 1433–1441, 2017.

[26] N. Hallak, P. Mertikopoulos, and V. Cevher, “Regret minimization in stochastic non-convex learning via a proximal-gradient approach,” in *International Conference on Machine Learning*, pp. 4008–4017, 2021.

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

[28] T. Chen, Y. Sun, Q. Xiao, and W. Yin, “A single-timescale method for stochastic bilevel optimization,” in *International Conference on Artificial Intelligence and Statistics*, pp. 2466–2488, 2022.

[29] F. Schäfer and A. Anandkumar, “Competitive gradient descent,” *Advances in Neural Information Processing Systems*, pp. 7625–7635, 2019.

[30] V. S. Varma, J. Veetaseveera, R. Postoyan, and I.-C. Morărescu, “Distributed gradient methods to reach a nash equilibrium in potential games,” in *IEEE Conference on Decision and Control*, pp. 3098–3103, 2021.

[31] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge university press, 2004.

[32] J. Zhang, P. Xiao, R. Sun, and Z. Luo, “A single-loop smoothed gradient descent-ascent algorithm for nonconvex-concave min-max problems,” *Advances in Neural Information Processing Systems*, pp. 7377–7389, 2020.

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

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

[35] S. Cen, C. Cheng, Y. Chen, Y. Wei, and Y. Chi, “Fast global convergence of natural policy gradient methods with entropy regularization,” *Operations Research*, vol. 70, no. 4, pp. 2563–2578, 2022.

[36] Z. Zheng, F. Gao, L. Xue, and J. Yang, “Federated Q-learning: Linear regret speedup with low communication cost,” *arXiv preprint arXiv:2312.15023*, 2023.

[37] H. Gong and M. Wang, “A duality approach for regret minimization in average-award ergodic markov decision processes,” in *Conference on Learning for Dynamics and Control*, vol. 120, pp. 862–883, 2020.

[38] Z. Zheng, H. Zhang, and L. Xue, “Federated Q-learning with reference-advantage decomposition: Almost optimal regret and logarithmic communication cost,” *arXiv preprint arXiv:2405.18795*, 2024.

[39] Z. Zheng, H. Zhang, and L. Xue, "Gap-dependent bounds for Q-learning using reference-advantage decomposition," *arXiv preprint arXiv:2410.07574*, 2024.

[40] S. Xu and M. Zhu, "Efficient gradient approximation method for constrained bilevel optimization," in *AAAI Conference on Artificial Intelligence*, vol. 37, pp. 12509–12517, 2023.

[41] S. Xu and M. Zhu, "Meta value learning for fast policy-centric optimal motion planning," in *Robotics science and systems*, 2022.

[42] S. Xu and M. Zhu, "Online constrained meta-learning: provable guarantees for generalization," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[43] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International Conference on Machine Learning*, pp. 1126–1135, 2017.

[44] S. K. Seyed Ghasemipour, S. S. Gu, and R. Zemel, "Smile: Scalable meta inverse reinforcement learning through context-conditional policies," in *Advances in Neural Information Processing Systems*, pp. 7881–7891, 2019.

[45] Y. Deng, F. Bao, Y. Kong, Z. Ren, and Q. Dai, "Deep direct reinforcement learning for financial signal representation and trading," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 3, pp. 653–664, 2016.

[46] Z. Zhang, S. Zohren, and S. Roberts, "Deep reinforcement learning for trading," *The Journal of Financial Data Science*, vol. 2, no. 2, pp. 25–40, 2020.

[47] X.-Y. Liu, H. Yang, J. Gao, and C. D. Wang, "Finrl: Deep reinforcement learning framework to automate trading in quantitative finance," in *ACM International Conference on AI in Finance*, pp. 1–9, 2021.

[48] B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey, "Human behavior modeling with maximum entropy inverse optimal control.," in *AAAI Spring Symposium: Human Behavior Modeling*, vol. 92, 2009.

[49] M. Monfort, A. Liu, and B. D. Ziebart, "Intent prediction and trajectory forecasting via predictive inverse linear-quadratic regulation," in *AAAI Conference on Artificial Intelligence*, pp. 3672–3678, 2015.

[50] S. Gaurav and B. Ziebart, "Discriminatively learning inverse optimal control models for predicting human intentions," in *International Conference on Autonomous Agents and MultiAgent Systems*, pp. 1368–1376, 2019.

[51] W. Krichene, M. Balandat, C. Tomlin, and A. Bayen, "The hedge algorithm on a continuum," in *International Conference on Machine Learning*, pp. 824–832, 2015.

[52] N. Agarwal, A. Gonen, and E. Hazan, "Learning in non-convex games with an optimization oracle," in *Conference on Learning Theory*, pp. 18–29, 2019.

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

[54] M.-F. Balcan, M. Khodak, and A. Talwalkar, "Provable guarantees for gradient-based meta-learning," in *International Conference on Machine Learning*, pp. 424–433, 2019.

[55] K. Xu, E. Ratner, A. Dragan, S. Levine, and C. Finn, "Learning a prior over intent via meta-inverse reinforcement learning," in *International Conference on Machine Learning*, pp. 6952–6962, 2019.

[56] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, "Meta-learning in neural networks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 5149–5169, 2021.

[57] X.-Y. Liu, Z. Xia, J. Rui, J. Gao, H. Yang, M. Zhu, C. Wang, Z. Wang, and J. Guo, "FinRL-Meta: Market environments and benchmarks for data-driven financial reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 1835–1849, 2022.

This appendix consists of four parts: related works, proof, meta-regularization algorithm and convergence guarantee, and Experiment details.

A Related works

Applying IRL to predict from ongoing trajectories. Papers [48, 49, 50] use standard IRL to predict goals of incomplete (or ongoing) trajectories. In specific, they first learn the reward function corresponding to each potential goal candidate from complete trajectories in the training phase and then use Bayesian methods to pick the most likely goal candidate of incomplete trajectories in the testing phase. However, these works are not in-trajectory learning since they do not learn a reward function from an incomplete ongoing trajectory.

Online non-convex optimization. This paper casts the in-trajectory learning problem as an online non-convex bi-level optimization problem where at each online iteration, a new state-action pair is input. Current literature on online non-convex optimization has two major categories. The first one is to use the regret originally defined in online convex optimization [51, 52]. However, it assumes to find a global optimal solution of a non-convex optimization problem at each online iteration. Therefore, the second category studies “local regret" [25, 26] and uses follow-the-leader-based methods to minimize the local regret. However, the follow-the-leader-based methods need to solve a non-convex optimization problem to obtain a near-stationary point at each online iteration, which can be computationally expensive and time-consuming. If the streaming data arrives at a fast speed, the computation at each online iteration may not be finished before the next data arrives. The computational burden of each online iteration can be mitigated by online gradient descent (OGD) methods where we only partially solve the non-convex optimization problem by one-step gradient descent at each online iteration. While OGD is widely studied in online convex optimization, it is rarely studied in online non-convex optimization. [11] uses OGD to quantify the local regret, however, its analysis only holds when the input data is identically independent distributed (i.i.d.). In contrast, the input data in our problem is not i.i.d. In specific, the input data at time t (i.e., (S_t^E, A_t^E)) is affected by the input data at last step (i.e., (S_{t-1}^E, A_{t-1}^E)). This temporal correlation between any two consecutive input data makes it difficult to analyze the growth rate of the local regret.

Regularization and meta-learning in IRL. Moreover, the data of in-trajectory learning is extremely lacking since there is only one demonstrated trajectory and this trajectory is not complete during the learning process. The lack of data can easily lead to overfitting and a common way to alleviate this problem is to use regularizers [21, 53]. Inspired by humans' using relevant experience to help the inference, we introduce a novel regularization method called meta-regularization [24, 54]. Compared to the regularizers commonly used in IRL [21, 53], the meta-regularizer provides human-experience-like prior information which helps recover the reward function from few data. Similar to the meta-initialization method [54, 43] commonly used in IRL [55], meta-regularization provides an initialization that the algorithm starts at. However, more importantly, meta-regularization also provides a regularization term to avoid overfitting.

A.1 Distinction from Maximum-likelihood inverse reinforcement learning (ML-IRL) [5]

We discuss our distinctions from ML-IRL from the following three aspects: problem setting, algorithm design, and theoretical analysis.

Distinction in problem setting. We study in-trajectory IRL and formulate an online optimization problem, while ML-IRL studies standard IRL and formulates an offline optimization problem.

Distinctions in algorithm design. ML-IRL and our algorithm both update policy and reward in a single loop. However, we propose a novel reward update mechanism specially designed for the in-trajectory learning case. This special design requires to use the current learned policy to complete the expert trajectory, which gives the algorithm the ability to consider for the future. This special design of reward update is novel compared to ML-IRL.

Distinctions in theoretical analysis. The analysis in our paper is substantially different from that in ML-IRL due to three facts: (1) The input data in our paper is not i.i.d., while the input data in ML-IRL is i.i.d. (2) We solve an online optimization problem, while ML-IRL solves an offline optimization problem. (3) Our analysis holds for continuous state-action space, while the analysis of ML-IRL is limited to finite state-action space. We now discuss the distinctions in theoretical analysis caused by the three facts in detail.

In our case, the input data is not i.i.d. but temporally correlated, i.e., the input data (s_t, a_t) is affected by the input data (s_{t-1}, a_{t-1}) at last time step. In contrast, the input data in ML-IRL is i.i.d. sampled from a pre-collected data set. To solve this non i.i.d. issue of the input data, we propose a novel theoretical technique that has three steps (detailed in Subsection 4.2). Step 1: We propose the stationary distribution μ^{π_E} and quantify the gradient norm difference between the real distribution $\mathbb{P}_t^{\pi_E}$ and this stationary distribution μ^{π_E} in Proposition 1. Step 2: We quantify the local regret over the stationary distribution in Lemma 2. The benefit of doing this is that the input data can be regarded as i.i.d. sampled from this stationary distribution. Step 3: We combine step 1 and step 2, and quantify the (local) regret over the real distribution, where the data is not i.i.d., in Theorem 1 and Theorem 2. We can see that step 1 and step 3 are to solve the non i.i.d. issue of the input data, so that the corresponding theorem statements (Proposition 1, Theorem 1, and Theorem 2) are novel compared to ML-IRL because ML-IRL does not have this non i.i.d. issue. The only theorem statement relevant to ML-IRL is Lemma 2 in step 2 where we both analyze the algorithm over a stationary distribution, and the data is i.i.d. sampled from the stationary distribution.

However, Lemma 2 in step 2 still has significant distinctions from ML-IRL because Lemma 2 quantifies the local regret in the context of online optimization, while ML-IRL quantifies convergence in the context of offline optimization. First, the objective function is dynamically changing in the online setting because the learner observes a new state-action pair at each online iteration, while the objective function is fixed in ML-IRL. Second, the local regret contains the term $L_t(\theta_t; (s_t^E, a_t^E))$, however, θ_t is computed before the learner knows (s_t^E, a_t^E). This makes it more difficult to quantify the local regret because the learner does not know L_t when it computes θ_t. These two difficulties do not appear in the offline optimization in ML-IRL. To solve these two issues, we need to additionally construct a new time-invariant function $\bar{L}$ in Appendix B.4 and quantify the convergence of the new function $\bar{L}$. Then, in order to quantify the local regret of $\{L_t\}_{t\geq 0}$, we need to quantify the difference between the real loss function $\{L_t\}_{t\geq 0}$ and the constructed loss function $\bar{L}$.

Moreover, our theoretical analysis holds for continuous state-action space while the theoretical analysis in ML-IRL is limited to finite state-action space. The extension to continuous state-action space brings new difficulties and requires significant novel analysis. In general, the difficulties stem from two aspects: (1) The constants in ML-IRL, e.g., the smoothness constant of the loss function L and the coefficient of convergence rate, include the term $|\mathcal{S}| \times |\mathcal{A}|$. When the state-action space is continuous, those constants are not finite because $|\mathcal{S}| \times |\mathcal{A}|$ is now infinite. To address this issue, we propose new methods to bound those constants. For example, in order to show that the loss function L is smooth, rather than using $||\nabla L(\theta_1) - \nabla L(\theta_2)||$ to find the smoothness constant as in ML-IRL, we aim to show that $||\nabla^2 L(\theta)||$ is upper bounded by a constant C_L in Lemma A.2 and this constant C_L does not rely on $|\mathcal{S}| \times |A|$. Given that $||\nabla^2 L(\theta)|| \leq C_L$, the loss function L is C_L-smooth. (2) Since the action space $\mathcal{A}$ is finite in ML-IRL, their proved properties of the Q-function Q^{soft} (e.g., Lipschitz continuity, contraction property, monotonic improvement, and smoothness) can be easily extended to the value function V^{soft} by summing over different actions $a \in \mathcal{A}$. When the action space becomes continuous, summing over infinitely many different actions does not preserve those properties. Thus we have to propose new methods to prove those propoerties of the value function V^{soft}. In specific, we prove the Lipschitz continuity, contraction property, monotonic improvement, and smoothness of the value function V^{soft} in Claims 3-5 in Appendix B.4.

B Proof

This section provides the proof of all the proposition, lemmas, and theorems in the paper. To start with, we first introduce the expression of soft Q-function and soft Bellman policy.

B.1 Notions

The soft Q-function and soft value function are:

$$Q_{\theta,\pi}^{\text{soft}}(s, a) \triangleq r_\theta(s, a) + \gamma \int_{s' \in \mathcal{S}} P(s'|s, a) V_{\theta,\pi}^{\text{soft}}(s') ds',$$

$$V^{\text{soft}}_{\theta,\pi}(s) \triangleq E^{\pi}_{S,A}\left[\sum_{t=0}^{\infty}\gamma^t\big(r_\theta(S_t,A_t)-\log\pi(A_t|S_t)\big)\bigg|S_0=s^E_0\right].$$

The soft Bellman policy is as follows:

$$\begin{aligned}
\pi_\theta(a|s) &= \frac{\exp(Q^{\text{soft}}_\theta(s,a))}{\exp(V^{\text{soft}}_\theta(s))},\\
Q^{\text{soft}}_\theta(s,a) &= r_\theta(s,a)+\gamma\int_{s'\in\mathcal{S}}P(s'|s,a)V^{\text{soft}}_\theta(s')ds',\\
V^{\text{soft}}_\theta(s) &= \log\bigg(\int_{a\in\mathcal{A}}\exp(Q^{\text{soft}}_\theta(s,a))da\bigg).
\end{aligned}$$

It has been proved [33] that the soft Bellman policy π_θ is the optimal solution of the lower-level problem (4). We define $J_\theta(s) \triangleq E^{\pi_\theta}_{S,A}[\sum_{t=0}^{\infty}\gamma^t r_\theta(S_t,A_t)|S_0=s]$ as the expected cumulative reward of policy π_θ starting from state s and $J_\theta(s,a) \triangleq E^{\pi_\theta}_{S,A}[\sum_{t=0}^{\infty}\gamma^t r_\theta(S_t,A_t)|S_0=s,A_0=a]$.

Lemma 3. *We have the gradient* $\nabla_\theta\log\pi_\theta(a|s) = E^{\pi_\theta}_{S,A}[\sum_{t=0}^{\infty}\gamma^t\nabla_\theta r_\theta(S_t,A_t)|S_0=s,A_0=a] - E^{\pi_\theta}_{S,A}[\sum_{t=0}^{\infty}\gamma^t\nabla_\theta r_\theta(S_t,A_t)|S_0=s]$.

Proof. Define $Z_\theta(s,a)\triangleq\exp(Q^{\text{soft}}_\theta(s,a))$ and $Z_\theta(s)\triangleq\exp(V^{\text{soft}}_\theta(s))$, therefore Z_θ is smooth in θ given that it is a composition of logarithmic, exponential, and linear functions of r_θ and r_θ is smooth in θ (Assumption 1).

$$\begin{aligned}
\nabla_\theta\log Z_\theta(s) &= \frac{\int_{a\in\mathcal{A}}\nabla_\theta Z_\theta(s,a)da}{Z_\theta(s)},\\
&= \int_{a\in\mathcal{A}}\frac{Z_\theta(s,a)}{Z_\theta(s)}\nabla_\theta\log Z_\theta(s,a)da,\\
&= \int_{a\in\mathcal{A}}\pi_\theta(a|s)\bigg[\nabla_\theta r_\theta(s,a)+\gamma\int_{s'\in\mathcal{S}}P(s'|s,a)\nabla_\theta\log Z_\theta(s')ds'\bigg]da,\\
&= \int_{a\in\mathcal{A}}\pi_\theta(a|s)\bigg[\nabla_\theta r_\theta(s,a)+\gamma\int_{s'\in\mathcal{S}}P(s'|s,a)\int_{a'\in\mathcal{A}}\bigg(\nabla_\theta r_\theta(s',a')\\
&+\gamma\int_{s''\in\mathcal{S}}P(s''|s',a')\nabla_\theta\log Z_\theta(s'')ds''\bigg)da'ds'\bigg]da.
\end{aligned}$$

Keep the expansion, we can get $\nabla_\theta\log Z_\theta(s) = E^{\pi_\theta}_{S,A}[\sum_{t=0}^{\infty}\gamma^t\nabla_\theta r_\theta(S_t,A_t)|S_0=s]$ and similarly we can get $\nabla_\theta\log Z_\theta(s,a) = E^{\pi_\theta}_{S,A}[\sum_{t=0}^{\infty}\gamma^t\nabla_\theta r_\theta(S_t,A_t)|S_0=s,A_0=a]$. Thus we have the gradient $\nabla_\theta\log\pi_\theta(a|s) = \nabla_\theta\log Z_\theta(s,a)-\nabla_\theta\log Z_\theta(s)$. □

B.2 Proof of Lemma 1

Recall that $\sum_{i=0}^{t}L_i(\theta;(s^E_i,a^E_i)) = -\sum_{i=0}^{t}\gamma^i\log\pi_\theta(a^E_i|s^E_i)$. When the dynamics P is deterministic, we have that

$$\begin{aligned}
&\nabla\sum_{i=0}^{t}L_i(\theta;(s^E_i,a^E_i)) = -\sum_{i=0}^{t}\gamma^i\nabla_\theta\log\pi_\theta(a^E_i|s^E_i),\\
&= -\sum_{i=0}^{t}\gamma^i\bigg[\nabla_\theta Q^{\text{soft}}_\theta(s^E_i,a^E_i)-\nabla_\theta V^{\text{soft}}_\theta(s^E_i)\bigg],\\
&= -\sum_{i=0}^{t}\gamma^i\bigg[\nabla_\theta r_\theta(s^E_i,a^E_i)+\gamma\nabla_\theta V^{\text{soft}}_\theta(s^E_{i+1})-\nabla_\theta V^{\text{soft}}_\theta(s^E_i)\bigg],\\
&= -\sum_{i=0}^{t}\gamma^i\nabla_\theta r_\theta(s^E_i,a^E_i)-\sum_{i=1}^{t+1}\gamma^i\nabla_\theta V^{\text{soft}}_\theta(s^E_i)+\sum_{i=0}^{t}\gamma^i\nabla_\theta V^{\text{soft}}_\theta(s^E_i),
\end{aligned}$$

$$
\begin{aligned}
&= -\sum_{i=0}^{t} \gamma^i \nabla_\theta r_\theta(s_i^E, a_i^E) - \gamma^{t+1} \nabla_\theta V_\theta^{\text{soft}}(s_{t+1}^E) + \nabla_\theta V_\theta^{\text{soft}}(s_0^E), \\
&\overset{(a)}{=} -\sum_{i=0}^{t} \gamma^i \nabla_\theta r_\theta(s_i^E, a_i^E) - E_{S,A}^{\pi_\theta}\Bigg[\sum_{i=t+1}^{\infty} \gamma^i \nabla_\theta r_\theta(S_i, A_i) | S_i = s_{t+1}^E\Bigg] \\
&+ E_{S,A}^{\pi_\theta}\Bigg[\sum_{i=0}^{\infty} \gamma^i \nabla_\theta r_\theta(S_i, A_i) | S_0 = s_0^E\Bigg], \\
&= -\sum_{i=0}^{t} \gamma^i \nabla_\theta r_\theta(s_i^E, a_i^E) - E_{S,A}^{\pi_\theta}\Bigg[\sum_{i=t+1}^{\infty} \gamma^i \nabla_\theta r_\theta(S_i, A_i) | S_t = s_t^E, A_t = a_t^E\Bigg] \\
&+ E_{S,A}^{\pi_\theta}\Bigg[\sum_{i=0}^{\infty} \gamma^i \nabla_\theta r_\theta(S_i, A_i) | S_0 = s_0^E\Bigg],
\end{aligned}
$$

where equality (a) follows from the proof of Lemma 3.

When the dynamics P is stochastic, we can prove that the above gradient is an unbiased estimate of $\nabla E_{\{(S_i^E, A_i^E) \sim \mathbb{P}_i^{\pi_E}(\cdot,\cdot)(\cdot,\cdot)\}_{i\geq 0}}\Big[\sum_{i=0}^{t} L_i(\theta; (S_i^E, A_i^E))\Big]$:

$$
\begin{aligned}
&\nabla E_{\{(S_i^E, A_i^E) \sim \mathbb{P}_i^{\pi_E}(\cdot,\cdot)(\cdot,\cdot)\}_{i\geq 0}}\Bigg[\sum_{i=0}^{t} L_i(\theta; (S_i^E, A_i^E))\Bigg], \\
&= \nabla E_{\{(S_i^E, A_i^E) \sim \mathbb{P}_i^{\pi_E}(\cdot,\cdot)(\cdot,\cdot)\}_{i\geq 0}}\Bigg[-\sum_{i=0}^{t} \gamma^i \nabla_\theta \log \pi_\theta(A_i^E | S_i^E)\Bigg], \\
&= E_{\{(S_i^E, A_i^E) \sim \mathbb{P}_i^{\pi_E}(\cdot,\cdot)(\cdot,\cdot)\}_{i\geq 0}}\Bigg[-\sum_{i=0}^{t} \gamma^i [\nabla_\theta r_\theta(S_i^E, A_i^E) + \gamma E_{S_{i+1} \sim P(\cdot|S_i^E, A_i^E)}[\nabla_\theta V_\theta^{\text{soft}}(S_{i+1})] \\
&- \nabla_\theta V_\theta^{\text{soft}}(S_i^E)]\Bigg], \\
&= E_{\{(S_i^E, A_i^E) \sim \mathbb{P}_i^{\pi_E}(\cdot,\cdot)(\cdot,\cdot)\}_{i\geq 0}}\Bigg[-\sum_{i=0}^{t} \gamma^i [\nabla_\theta r_\theta(S_i^E, A_i^E) + \gamma \nabla_\theta V_\theta^{\text{soft}}(S_{i+1}^E) - \nabla_\theta V_\theta^{\text{soft}}(S_i^E)]\Bigg], \\
&= E_{\{(S_i^E, A_i^E) \sim \mathbb{P}_i^{\pi_E}(\cdot,\cdot)(\cdot,\cdot)\}_{i\geq 0}}\Bigg[-\sum_{i=0}^{t} \gamma^i \nabla_\theta r_\theta(S_i^E, A_i^E) - \gamma^{t+1} \nabla_\theta V_\theta^{\text{soft}}(S_{t+1}^E) \\
&+ E_{S,A}^{\pi_\theta}\Bigg[\sum_{i=0}^{\infty} \gamma^i \nabla_\theta r_\theta(S_i, A_i) | S_0 = S_0^E\Bigg]\Bigg], \\
&\overset{(b)}{=} E_{\{(S_i^E, A_i^E) \sim \mathbb{P}_i^{\pi_E}(\cdot,\cdot)(\cdot,\cdot)\}_{i\geq 0}}\Bigg[-\sum_{i=0}^{t} \gamma^i \nabla_\theta r_\theta(S_i^E, A_i^E) \\
&- E_{S,A}^{\pi_\theta}[\sum_{i=t+1}^{\infty} \gamma^i \nabla_\theta r_\theta(S_i, A_i) | S_t = S_t^E, A_t = A_t^E] + E_{S,A}^{\pi_\theta}[\sum_{i=0}^{\infty} \gamma^i \nabla_\theta r_\theta(S_i, A_i) | S_0 = S_0^E]\Bigg],
\end{aligned}
$$

where equality (b) follows from the fact that

$$
\begin{aligned}
&E_{\{(S_i^E, A_i^E) \sim \mathbb{P}_i^{\pi_E}(\cdot,\cdot)(\cdot,\cdot)\}_{i\geq 0}}\Bigg[E_{S,A}^{\pi_\theta}[\sum_{i=t+1}^{\infty} \gamma^i \nabla_\theta r_\theta(S_i, A_i) | S_t = S_t^E, A_t = A_t^E]\Bigg], \\
&= E_{\{(S_i^E, A_i^E) \sim \mathbb{P}_i^{\pi_E}(\cdot,\cdot)(\cdot,\cdot)\}_{i\geq 0}}\Bigg[E_{S,A}^{\pi_\theta}[\sum_{i=t+1}^{\infty} \gamma^i \nabla_\theta r_\theta(S_i, A_i) | S_{t+1} = S_{t+1}^E]\Bigg],
\end{aligned}
$$

because $\mathbb{P}_{t+1}^{\pi_E}(\cdot) = \mathbb{P}_t^{\pi_E}(S_t^E, A_t^E) P(\cdot | S_t^E, A_t^E)$ and $S_{t+1}^E \sim P(\cdot | S_t^E, A_t^E)$.

Since we quantify the local regret in expectation in Theorem 1, this unbiased estimate can be used when the dynamics is stochastic.

B.3 Proof of Proposition 1

From Assumption 2, we know that $d_{\text{TV}}(\mathbb{P}_t^{\pi_E}(\cdot), \mu^{\pi_E}(\cdot)) = \frac{1}{2}\int_{s\in\mathcal{S}} |\mathbb{P}_t^{\pi_E}(s) - \mu^{\pi_E}(s)|ds \leq C_M\rho^t$ where the initial state is s_0. For any state-action pair $(s,a) \in \mathcal{S}\times\mathcal{A}$, we know that $\mathbb{P}_t^{\pi_E}(s,a) = \mathbb{P}_t^{\pi_E}(s)\pi_E(a|s)$ and $\mu^{\pi_E}(s,a) = \mu^{\pi_E}(s)\pi_E(a|s)$. Therefore, we have that:

$$\begin{aligned}
&d_{\text{TV}}(\mathbb{P}^{\pi_E}(\cdot,\cdot), \mu^{\pi_E}(\cdot,\cdot)),\\
&= \frac{1}{2}\int_{s\in\mathcal{S}}\int_{a\in\mathcal{A}} |\mathbb{P}_t^{\pi_E}(s)\pi_E(a|s) - \mu^{\pi_E}(s)\pi_E(a|s)|dsda,\\
&= \frac{1}{2}\int_{s\in\mathcal{S}}\int_{a\in\mathcal{A}} |\mathbb{P}_t^{\pi_E}(s) - \mu^{\pi_E}(s)|\pi_E(a|s)dsda,\\
&= \frac{1}{2}\int_{s\in\mathcal{S}} |\mathbb{P}_t^{\pi_E}(s) - \mu^{\pi_E}(s)|ds,\\
&\leq C_M\rho^t.
\end{aligned}$$

Claim 1. *The trajectory of θ_t is bounded, i.e., $||\theta_t - \bar{\theta}|| \leq \frac{2\bar{C}_r}{\lambda}$ for any $t \geq 0$.*

Proof.

$$\begin{aligned}
&||\theta_{t+1} - \bar{\theta}|| = ||\theta_t - \alpha_t g_t - \bar{\theta}||,\\
&= \left|\left|\theta_t - \bar{\theta} - \alpha_t\left[\sum_{i=0}^{\infty}\gamma^i\nabla_\theta r_{\theta_t}(s_i', a_i') - \sum_{i=0}^{\infty}\gamma^i\nabla_\theta r_{\theta_t}(s_i'', a_i'') + \frac{\lambda(1-\gamma^{t+1})}{1-\gamma}(\theta_t - \bar{\theta})\right]\right|\right|,\\
&= \left|\left|(1 - \frac{\alpha_t\lambda(1-\gamma^{t+1})}{1-\gamma})(\theta_t - \bar{\theta}) - \alpha_t\left[\sum_{i=0}^{\infty}\gamma^i\nabla_\theta r_{\theta_t}(s_i', a_i') - \sum_{i=0}^{\infty}\gamma^i\nabla_\theta r_{\theta_t}(s_i'', a_i'')\right]\right|\right|,\\
&\overset{(a)}{\leq} (1 - \frac{\alpha_t\lambda(1-\gamma^{t+1})}{1-\gamma})||\theta_t - \bar{\theta}|| + \alpha_t\left|\left|\sum_{i=0}^{\infty}\gamma^i\nabla_\theta r_{\theta_t}(s_i', a_i') - \sum_{i=0}^{\infty}\gamma^i\nabla_\theta r_{\theta_t}(s_i'', a_i'')\right|\right|,\\
&\overset{(b)}{\leq} (1 - \frac{\alpha_t\lambda(1-\gamma^{t+1})}{1-\gamma})||\theta_t - \bar{\theta}|| + \frac{2\alpha_t\bar{C}_r}{1-\gamma},\\
&\leq (1 - \frac{\alpha_t\lambda}{1-\gamma})||\theta_t - \bar{\theta}|| + \frac{2\alpha_t\bar{C}_r}{1-\gamma},
\end{aligned}$$

where (a) follows triangle inequality and (b) uses the upper bound of $\nabla_\theta r_\theta$ in Assumption 1. Therefore, we have the following relation:

$$\begin{aligned}
&||\theta_{t+1} - \bar{\theta}|| - \frac{2\bar{C}_r}{\lambda} \leq (1 - \frac{\alpha_t\lambda}{1-\gamma})\bigg(||\theta_t - \bar{\theta}|| - \frac{2\bar{C}_r}{\lambda}\bigg),\\
&\Rightarrow ||\theta_t - \bar{\theta}|| \leq (1 - \frac{\alpha_t\lambda}{1-\gamma})^t\bigg(||\theta_0 - \bar{\theta}|| - \frac{2\bar{C}_r}{\lambda}\bigg) + \frac{2\bar{C}_r}{\lambda},\\
&\overset{(c)}{=} \frac{2\bar{C}_r}{\lambda}\left[1 - (1 - \frac{\alpha_t\lambda}{1-\gamma})^t\right] \overset{(d)}{\leq} \frac{2\bar{C}_r}{\lambda},
\end{aligned}$$

where (c) follows the fact that $\theta_0 = \bar{\theta}$ and (d) follows the fact that $\alpha_t \leq \frac{1-\gamma}{\lambda}$. □

Recall that the loss function $L_i(\theta; (s_i^E, a_i^E)) = -\gamma^i \log \pi_\theta(a_i^E|s_i^E) + \frac{\lambda\gamma^i}{2}||\theta - \bar{\theta}||^2$ and thus $\nabla L_i(\theta; (s_i^E, a_i^E)) = -\gamma^i[\nabla_\theta Q_\theta^{\text{soft}}(s_i^E, a_i^E) - \nabla_\theta V_\theta^{\text{soft}}(s_i^E)] + \lambda\gamma^i(\theta - \bar{\theta})$. From Lemma 3, we know that $\nabla_\theta V_\theta^{\text{soft}}(s_i^E) = E_{S,A}^{\pi_\theta}[\sum_{t=0}^{\infty}\gamma^t\nabla_\theta r_\theta(S_t, A_t)|S_0 = s_i^E]$ and $\nabla_\theta Q_\theta^{\text{soft}}(s_i^E, a_i^E) = E_{S,A}^{\pi_\theta}[\sum_{t=0}^{\infty}\gamma^t\nabla_\theta r_\theta(S_t, A_t)|S_0 = s_i^E, A_0 = a_i^E]$. Then, $||\nabla_\theta V_\theta^{\text{soft}}(s_i^E)|| \leq \frac{\bar{C}_r}{1-\gamma}$ and $||\nabla_\theta Q_\theta^{\text{soft}}(s_i^E, a_i^E)|| \leq \frac{\bar{C}_r}{1-\gamma}$.

Now we can see that

$$||\nabla L_i(\theta_t;; (s_i^E, a_i^E))|| = \gamma^i||\nabla_\theta Q_{\theta_t}^{\text{soft}}(s_i^E, a_i^E) - \nabla_\theta V_{\theta_t}^{\text{soft}}(s_i^E) + \lambda(\theta_t - \bar{\theta})||,$$

$$
\begin{aligned}
&\leq \gamma^i ||\nabla_\theta Q^{\text{soft}}_{\theta_t}(s^E_i, a^E_i) - \nabla_\theta V^{\text{soft}}_{\theta_t}(s^E_i) + \lambda(\theta_t - \bar{\theta})|| \leq \gamma^i \left(\frac{2\bar{C}_r}{1-\gamma} + 2\bar{C}_r \right), \\
&\Rightarrow ||\nabla L_i(\theta_t; (s^E_i, a^E_i))||^2 \leq 4\bar{C}_r^2 \gamma^{2i} \left(\frac{2-\gamma}{1-\gamma} \right)^2.
\end{aligned}
$$

Therefore, we have that

$$
\begin{aligned}
&\left| E_{(S^E_i, A^E_i) \sim \mathbb{P}^{\pi_E}_i(\cdot,\cdot)} \left[||\nabla L_i(\theta_t; (S^E_i, A^E_i))||^2 \right] - E_{(S^E_i, A^E_i) \sim \mu^{\pi_E}(\cdot,\cdot)} \left[||\nabla L_i(\theta_t; (S^E_i, A^E_i))||^2 \right] \right|, \\
&= \left| \int_{s \in \mathcal{S}} \int_{a \in \mathcal{A}} \mathbb{P}^{\pi_E}_i(s,a) ||\nabla L_i(\theta_t; (s,a))||^2 dads \right. \\
&\left. - \int_{s \in \mathcal{S}} \int_{a \in \mathcal{A}} \mu^{\pi_E}(s,a) ||\nabla L_i(\theta_t; (s,a))||^2 dads \right|, \\
&\leq \int_{s \in \mathcal{S}} \int_{a \in \mathcal{A}} |\mathbb{P}^{\pi_E}_i(s,a) - \mu^{\pi_E}(s,a)| \cdot ||\nabla L_i(\theta_t; (s,a))||^2 dads, \\
&\leq 2C_M \rho^i \cdot 4\bar{C}_r^2 \gamma^{2i} \left(\frac{2-\gamma}{1-\gamma} \right)^2 = 8C_M \bar{C}_r^2 \left(\frac{2-\gamma}{1-\gamma} \right)^2 \rho^i \gamma^{2i}.
\end{aligned}
$$

Lemma 4. *Suppose Assumptions 1-2 hold, the we have the following for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ and any θ_1, θ_2, t: $||\nabla L_t(\theta_1, (s,a)) - \nabla L_t(\theta_2, (s,a))|| \leq C_L ||\theta_1 - \theta_2||$ and $|Q^{soft}_{\theta_1, \pi_{\theta_1}}(s,a) - Q^{soft}_{\theta_2, \pi_{\theta_2}}(s,a)| \leq C_Q ||\theta_1 - \theta_2||$, where $C_L = \frac{2\tilde{C}_r}{1-\gamma} + \frac{4\bar{C}_r^3}{(1-\gamma)^4} + \lambda$ and $C_Q = \frac{\bar{C}_r}{1-\gamma}$.*

Proof. Note that $Q^{\text{soft}}_{\theta,\pi_\theta} = Q^{\text{soft}}_\theta$ and $\nabla_\theta Q^{\text{soft}}_\theta(s,a) = E^{\pi_\theta}_{S,A}[\sum_{t=0}^\infty \gamma^t \nabla_\theta r_\theta(S_t, A_t) | S_0 = s, A_0 = a]$ (proof of Lemma 3). Therefore, we have that

$$
||\nabla_\theta Q^{\text{soft}}_\theta(s,a)|| \leq \frac{\bar{C}_r}{1-\gamma} \triangleq C_Q
$$

We know from Lemma 1 that $\nabla L_t(\theta; (s^E_t, a^E_t)) = -\gamma^t [\nabla_\theta Q^{\text{soft}}_\theta(s^E_t, a^E_t) - \nabla_\theta V^{\text{soft}}_\theta(s^E_t)] + \lambda \gamma^t (\theta - \bar{\theta})$. To find the smoothness constant of L_t, we need to compute the Hessian of L_t. First, we have that

$$
\begin{aligned}
&\nabla^2_{\theta\theta} Q^{\text{soft}}_\theta(s,a) = \nabla_\theta E^{\pi_\theta}_{S,A}[\sum_{t=0}^\infty \gamma^t \nabla_\theta r_\theta(S_t, A_t) | S_0 = s, A_0 = a], \\
&= \nabla^2_{\theta\theta} r_\theta(s,a) + \gamma \int_{s' \in \mathcal{S}} P(s'|s,a) \nabla_\theta E^{\pi_\theta}_{S,A}[\sum_{t=0}^\infty \gamma^t \nabla_\theta r_\theta(S_t, A_t) | S_0 = s'] ds', \\
&= \nabla^2_{\theta\theta} r_\theta(s,a) \\
&+ \gamma \int_{s' \in \mathcal{S}} P(s'|s,a) \nabla_\theta \int_{a' \in \mathcal{A}} \pi_\theta(a'|s') E^{\pi_\theta}_{S,A}[\sum_{t=0}^\infty \gamma^t \nabla_\theta r_\theta(S_t, A_t) | S_0 = s', A_0 = a'] da' ds', \\
&= \nabla^2_{\theta\theta} r_\theta(s,a) + \gamma \int_{s' \in \mathcal{S}} P(s'|s,a) \int_{a' \in \mathcal{A}} \Bigg[\nabla_\theta \pi_\theta(a'|s') \cdot E^{\pi_\theta}_{S,A}[\sum_{t=0}^\infty \gamma^t \nabla_\theta r_\theta(S_t, A_t) | S_0 = s', \\
&A_0 = a'] + \pi_\theta(a'|s') \cdot \nabla_\theta E^{\pi_\theta}_{S,A}[\sum_{t=0}^\infty \gamma^t \nabla_\theta r_\theta(S_t, A_t) | S_0 = s', A_0 = a'] \Bigg] da' ds'.
\end{aligned}
$$

Keep the expansion, we can get

$$
\begin{aligned}
&\nabla^2_{\theta\theta} Q^{\text{soft}}_\theta(s,a) = E^{\pi_\theta}_{S,A}[\sum_{t=0}^\infty \gamma^t \nabla^2_{\theta\theta} r_\theta(S_t, A_t) | S_0 = s_0, A_0 = a_0] \\
&+ E^{\pi_\theta}_{S,A} \left[\sum_{i=0}^\infty \gamma^i \nabla_\theta \pi_\theta(A_i|S_i) \cdot E^{\pi_\theta}_{S',A'}[\sum_{t=0}^\infty \gamma^t \nabla_\theta r_\theta(S'_t, A'_t) | S'_0 = S_i, A'_0 = A_i] \middle| S_0 = s_0, A_0 = a_0 \right].
\end{aligned}
$$

Now we take a look at the second term in the above equality:

$$
\begin{aligned}
&\nabla_\theta \pi_\theta(a|s) \cdot E^{\pi_\theta}_{S,A}[\sum_{t=0}^{\infty} \gamma^t \nabla_\theta r_\theta(S_t, A_t)|S_0 = s, A_0 = a], \\
&= \pi_\theta(a|s) \nabla_\theta \log \pi_\theta(a|s) \cdot E^{\pi_\theta}_{S,A}[\sum_{t=0}^{\infty} \gamma^t \nabla_\theta r_\theta(S_t, A_t)|S_0 = s, A_0 = a], \\
&= \pi_\theta(a|s) \Big[\nabla_\theta Q^{\text{soft}}_\theta(s,a) - \nabla_\theta V^{\text{soft}}_\theta(s)\Big] \cdot E^{\pi_\theta}_{S,A}[\sum_{t=0}^{\infty} \gamma^t \nabla_\theta r_\theta(S_t, A_t)|S_0 = s, A_0 = a], \\
&\Rightarrow \Big|\Big|\nabla_\theta \pi_\theta(a|s) \cdot E^{\pi_\theta}_{S,A}[\sum_{t=0}^{\infty} \gamma^t \nabla_\theta r_\theta(S_t, A_t)|S_0 = s, A_0 = a]\Big|\Big|, \\
&\leq (\frac{\bar{C}_r}{1-\gamma} + \frac{\bar{C}_r}{1-\gamma}) \cdot \frac{\bar{C}_r}{1-\gamma} = \frac{2\bar{C}_r^3}{(1-\gamma)^3}.
\end{aligned}
$$

Therefore, we have that

$$
\begin{aligned}
&||\nabla^2_{\theta\theta} Q^{\text{soft}}_\theta(s,a)|| \leq \Big|\Big|E^{\pi_\theta}_{S,A}[\sum_{t=0}^{\infty} \gamma^t \nabla^2_{\theta\theta} r_\theta(S_t, A_t)|S_0 = s_0, A_0 = a_0]\Big|\Big| \\
&+ \Big|\Big|E^{\pi_\theta}_{S',A'}\Big[\sum_{i=0}^{\infty} \gamma^i \nabla_\theta \pi_\theta(A'_i|S'_i) \cdot E^{\pi_\theta}_{S,A}[\sum_{t=0}^{\infty} \gamma^t \nabla_\theta r_\theta(S_t, A_t)|S_0 = S'_i, A_0 = A'_i]\Big]\Big|\Big|, \\
&\leq \frac{\tilde{C}_r}{1-\gamma} + \sum_{t=0}^{\infty} \gamma^t \frac{2\bar{C}_r^3}{(1-\gamma)^3} = \frac{\tilde{C}_r}{1-\gamma} + \frac{2\bar{C}_r^3}{(1-\gamma)^4}.
\end{aligned}
$$

Similarly, we can get $||\nabla^2_{\theta\theta} V^{\text{soft}}_\theta(s)|| \leq \frac{\tilde{C}_r}{1-\gamma} + \frac{2\bar{C}_r^3}{(1-\gamma)^4}$. Therefore, we have that

$$
\begin{aligned}
&||\nabla^2 L_t(\theta; (s^E_t, a^E_t))|| \leq \gamma^t \bigg(||\nabla^2_{\theta\theta} Q^{\text{soft}}_\theta(s^E_t, a^E_t)|| + ||\nabla^2_{\theta\theta} V^{\text{soft}}_\theta(s^E_t)|| + \lambda\bigg), \\
&\leq \gamma^t \bigg(\frac{2\tilde{C}_r}{1-\gamma} + \frac{4\bar{C}_r^3}{(1-\gamma)^4} + \lambda\bigg) \leq \frac{2\tilde{C}_r}{1-\gamma} + \frac{4\bar{C}_r^3}{(1-\gamma)^4} + \lambda \triangleq C_L. \qquad (6)
\end{aligned}
$$

□

B.4 Proof of Lemma 2

This proof is based on the proof in ML-IRL [5]. The differences are: (i) their proof only holds for finite state-action space while we extend to continuous state-action space; (ii) their analysis is for offline settings while we extend to online settings to quantify the local regret. We first introduce the following claims which serve as building blocks in this subsection.

Claim 2. *For any given policy π and state-action pair (s,a), it holds that $|Q^{\mathit{soft}}_{\theta_1,\pi}(s,a) - Q^{\mathit{soft}}_{\theta_2,\pi}(s,a)| \leq C_Q||\theta_1 - \theta_2||$ and $|V^{\mathit{soft}}_{\theta_1,\pi}(s) - V^{\mathit{soft}}_{\theta_2,\pi}(s)| \leq C_Q||\theta_1 - \theta_2||$.*

Proof.

$$
\begin{aligned}
&Q^{\text{soft}}_{\theta_1,\pi}(s,a) - Q^{\text{soft}}_{\theta_2,\pi}(s,a), \\
&= E^{\pi}_{S,A}\Big[\sum_{t=0}^{\infty} \gamma^t \big[r_{\theta_1}(S_t, A_t) - \log \pi(A_t|S_t)\big]\Big|S_0 = s, A_0 = a\Big] \\
&- E^{\pi}_{S,A}\Big[\sum_{t=0}^{\infty} \gamma^t \big[r_{\theta_2}(S_t, A_t) - \log \pi(A_t|S_t)\big]\Big|S_0 = s, A_0 = a\Big], \\
&= E^{\pi}_{S,A}\Big[\sum_{t=0}^{\infty} \gamma^t [r_{\theta_1}(S_t, A_t) - r_{\theta_2}(S_t, A_t)]\Big|S_0 = s, A_0 = a\Big],
\end{aligned}
$$

$$\Rightarrow |Q_{\theta_1,\pi}(s,a) - Q_{\theta_2,\pi}(s,a)| \leq \sum_{t=0}^{\infty} \gamma^t \bar{C}_r ||\theta_1 - \theta_2|| = \frac{\bar{C}_r}{1-\gamma} ||\theta_1 - \theta_2|| = C_Q ||\theta_1 - \theta_2||.$$

Similarly, we can get that

$$\begin{aligned} |V^{\text{soft}}_{\theta_1,\pi}(s) - V^{\text{soft}}_{\theta_2,\pi}(s)| &\leq E^{\pi}_{S,A}\left[\sum_{t=0}^{\infty} \gamma^t [r_{\theta_1}(S_t,A_t) - r_{\theta_2}(S_t,A_t)] \middle| S_0 = s\right], \\ &\leq \sum_{t=0}^{\infty} \gamma^t \bar{C}_r ||\theta_1 - \theta_2|| = \frac{\bar{C}_r}{1-\gamma} ||\theta_1 - \theta_2|| = C_Q ||\theta_1 - \theta_2||. \end{aligned}$$

□

Claim 3. *The soft Bellman operator $\mathcal{T}^{soft}_{\theta}$:*

$$\begin{aligned} (\mathcal{T}^{soft}_{\theta} Q)(s,a) &\triangleq r_\theta(s,a) + \gamma \int_{s' \in \mathcal{S}} P(s'|s,a) \log\left[\int_{a' \in \mathcal{A}} \exp(Q(s',a')) da'\right] ds', \\ (\mathcal{T}^{soft}_{\theta} V)(s) &\triangleq \log\left[\int_{a \in \mathcal{A}} \exp\left(r_\theta(s,a) + \gamma \int_{s' \in \mathcal{S}} P(s'|s,a) V(s') ds'\right) da\right], \end{aligned}$$

is a contraction map with constant γ.

Proof. It has been proved that $\mathcal{T}^{\text{soft}}_{\theta} Q$ is a contraction map with constant γ (Appendix A.2 in [33]). Here we show that $\mathcal{T}^{\text{soft}}_{\theta} V$ is a contraction map with constant γ. Define a norm of V as $||V_1 - V_2|| = \sup_{s \in \mathcal{S}} |V_1(s) - V_2(s)|$ and suppose $||V_1 - V_2|| = \epsilon$. Then we have that

$$\begin{aligned} \mathcal{T}^{\text{soft}}_{\theta} V_1(s) &= \log\left[\int_{a \in \mathcal{A}} \exp\left(r_\theta(s,a) + \gamma \int_{s' \in \mathcal{S}} P(s'|s,a) V_1(s') ds'\right) da\right], \\ &\leq \log\left[\int_{a \in \mathcal{A}} \exp\left(r_\theta(s,a) + \gamma \int_{s' \in \mathcal{S}} P(s'|s,a) [V_2(s') + \epsilon] ds'\right) da\right], \\ &= \log\left[\int_{a \in \mathcal{A}} \exp\left(r_\theta(s,a) + \gamma \int_{s' \in \mathcal{S}} P(s'|s,a) V_2(s') ds' + \gamma\epsilon\right) da\right], \\ &= \log\left[\int_{a \in \mathcal{A}} \exp(\gamma\epsilon) \exp\left(r_\theta(s,a) + \gamma \int_{s' \in \mathcal{S}} P(s'|s,a) V_2(s') ds'\right) da\right], \\ &= \mathcal{T}^{\text{soft}}_{\theta} V_2(s) + \gamma\epsilon. \end{aligned}$$

Similarly, we can get $\mathcal{T}^{\text{soft}}_{\theta} V_1(s) \geq \mathcal{T}^{\text{soft}}_{\theta} V_2(s) - \gamma\epsilon$. Therefore, $||\mathcal{T}^{\text{soft}}_{\theta} V_1 - \mathcal{T}^{\text{soft}}_{\theta} V_2|| \leq \gamma\epsilon = \gamma ||V_1 - V_2||$. □

Claim 4. *It holds that $Q^{soft}_{\theta_t,\pi_{t+1}}(s,a) \geq \mathcal{T}^{soft}_{\theta_t}(Q^{soft}_{\theta_t,\pi_t})(s,a)$ and $V^{soft}_{\theta_t,\pi_{t+1}}(s) \geq \mathcal{T}^{soft}_{\theta_t}(V^{soft}_{\theta_t,\pi_t})(s)$ for any (s,a).*

Proof.

$$\begin{aligned} Q^{\text{soft}}_{\theta_t,\pi_{t+1}}(s,a) &\overset{(i)}{=} r_{\theta_t}(s,a) + \gamma \int_{s' \in \mathcal{S}} P(s'|s,a) E_{a' \sim \pi_{t+1}}[Q^{\text{soft}}_{\theta_t,\pi_{t+1}}(s',a') - \log \pi_{t+1}(a'|s')] ds', \\ &\overset{(ii)}{\geq} r_{\theta_t}(s,a) + \gamma \int_{s' \in \mathcal{S}} P(s'|s,a) E_{A' \sim \pi_{t+1}(\cdot|s')}[Q^{\text{soft}}_{\theta_t,\pi_t}(s',A') - \log \pi_{t+1}(A'|s')] ds', \\ &= r_{\theta_t}(s,a) + \gamma \int_{s' \in \mathcal{S}} P(s'|s,a) \log\left[\int_{a' \in \mathcal{A}} \exp(Q^{\text{soft}}_{\theta_t,\pi_t}(s',a')) da'\right] ds', \\ &= \mathcal{T}^{\text{soft}}_{\theta}(Q^{\text{soft}}_{\theta_t,\pi_t})(s,a), \end{aligned}$$

where (i) follow equations (2)-(3) in [34] and (ii) follows policy improvement theorem (Theorem 4 in [33]). Similarly, we can get that

$$V^{\text{soft}}_{\theta_t,\pi_{t+1}}(s) = E_{A \sim \pi_{t+1}(\cdot|s)}[Q^{\text{soft}}_{\theta_t,\pi_{t+1}}(s,A) - \log \pi_{t+1}(A|s)],$$

$$\begin{aligned}
&\geq E_{A\sim\pi_{t+1}(\cdot|s)}[Q^{\text{soft}}_{\theta_t,\pi_t}(s,A) - \log\pi_{t+1}(A|s)],\\
&= \log\left[\int_{a\in\mathcal{A}} \exp(Q^{\text{soft}}_{\theta_t,\pi_t}(s,a))da\right],\\
&= \log\left[\int_{a\in\mathcal{A}} \exp\left(r_{\theta_t}(s,a) + \gamma\int_{s'\in\mathcal{S}} P(s'|s,a)V^{\text{soft}}_{\theta_t,\pi_t}(s')ds'\right)da\right],\\
&= \mathcal{T}^{\text{soft}}_{\theta_t}(V^{\text{soft}}_{\theta_t,\pi_t})(s).
\end{aligned}$$

□

Claim 5. *The following holds for any $(s,a)\in\mathcal{S}\times\mathcal{A}$ and θ_1, θ_2, t:*

$$|V^{\textit{soft}}_{\theta_1}(s,a) - V^{\textit{soft}}_{\theta_2}(s,a)| \leq C_Q||\theta_1-\theta_2||.$$

Proof. Note that $\nabla_\theta V^{\text{soft}}_\theta(s) = E^{\pi_\theta}_{S,A}[\sum_{t=0}^\infty \gamma^t\nabla_\theta r_\theta(S_t,A_t)|S_0=s]$ (proof of Lemma 3), therefore

$$||\nabla_\theta V^{\text{soft}}_\theta(s)|| \leq \frac{\bar{C}_r}{1-\gamma} = C_Q.$$

□

We first show the convergence of π_t in Algorithm 1. For any $(s,a)\in\mathcal{S}\times\mathcal{A}$, we know that

$$|\log\pi_{t+1}(a|s) - \log\pi_{\theta_t}(a|s)| \leq |Q^{\text{soft}}_{\theta_t,\pi_t}(s,a) - Q^{\text{soft}}_{\theta_t}(s,a)| + |V^{\text{soft}}_{\theta_t,\pi_t}(s) - V^{\text{soft}}_{\theta_t}(s)|.$$

Now we take a look at the term $|Q^{\text{soft}}_{\theta_t,\pi_t}(s,a) - Q^{\text{soft}}_{\theta_t}(s,a)|$:

$$\begin{aligned}
&|Q^{\text{soft}}_{\theta_t,\pi_t}(s,a) - Q^{\text{soft}}_{\theta_t}(s,a)|,\\
&\leq |Q^{\text{soft}}_{\theta_t,\pi_t}(s,a) - Q^{\text{soft}}_{\theta_{t-1},\pi_t}(s,a)| + |Q^{\text{soft}}_{\theta_{t-1},\pi_t}(s,a) - Q^{\text{soft}}_{\theta_{t-1}}(s,a)| + |Q^{\text{soft}}_{\theta_{t-1}}(s,a) - Q^{\text{soft}}_{\theta_t}(s,a)|,\\
&\overset{(a)}{\leq} C_Q||\theta_t-\theta_{t-1}|| + |Q^{\text{soft}}_{\theta_{t-1},\pi_t}(s,a) - Q^{\text{soft}}_{\theta_{t-1}}(s,a)| + C_Q||\theta_t-\theta_{t-1}||,\\
&= 2C_Q||\theta_t-\theta_{t-1}|| + |Q^{\text{soft}}_{\theta_{t-1},\pi_t}(s,a) - Q^{\text{soft}}_{\theta_{t-1}}(s,a)|,\\
&\overset{(b)}{=} 2C_Q||\theta_t-\theta_{t-1}|| + Q^{\text{soft}}_{\theta_{t-1}}(s,a) - Q^{\text{soft}}_{\theta_{t-1},\pi_t}(s,a),\\
&\overset{(c)}{\leq} 2C_Q||\theta_t-\theta_{t-1}|| + Q^{\text{soft}}_{\theta_{t-1}}(s,a) - \mathcal{T}^{\text{soft}}_{\theta_{t-1}}(Q^{\text{soft}}_{\theta_{t-1},\pi_{t-1}})(s,a),\\
&\overset{(d)}{=} 2C_Q||\theta_t-\theta_{t-1}|| + \mathcal{T}^{\text{soft}}_{\theta_{t-1}}(Q^{\text{soft}}_{\theta_{t-1}})(s,a) - \mathcal{T}^{\text{soft}}_{\theta_{t-1}}(Q^{\text{soft}}_{\theta_{t-1},\pi_{t-1}})(s,a),\\
&\overset{(e)}{\leq} 2C_Q||\theta_t-\theta_{t-1}|| + \gamma|Q^{\text{soft}}_{\theta_{t-1}}(s,a) - Q^{\text{soft}}_{\theta_{t-1},\pi_{t-1}}(s,a)|, \qquad (7)
\end{aligned}$$

where (a) follows Lemma 2 and Claim 2, (b) follows the fact that π_θ is the optimal solution, (c) follows Claim 4, (d) follows the fact that $Q^{\text{soft}}_{\theta_{t-1}}$ is a fixed point of $\mathcal{T}^{\text{soft}}_{\theta_{t-1}}$ (Theorem 2 in [33]), and (e) follows Claim 3.

Similarly we can bound the term $|V^{\text{soft}}_{\theta_t,\pi_t}(s) - V^{\text{soft}}_{\theta_t}(s)|$:

$$\begin{aligned}
&|V^{\text{soft}}_{\theta_t,\pi_t}(s) - V^{\text{soft}}_{\theta_t}(s)|,\\
&\leq |V^{\text{soft}}_{\theta_t,\pi_t}(s) - V^{\text{soft}}_{\theta_{t-1},\pi_t}(s)| + |V^{\text{soft}}_{\theta_{t-1},\pi_t}(s) - V^{\text{soft}}_{\theta_{t-1}}(s)| + |V^{\text{soft}}_{\theta_{t-1}}(s) - V^{\text{soft}}_{\theta_{t-1},\pi_{t-1}}(s)|,\\
&\overset{(f)}{\leq} C_Q||\theta_t-\theta_{t-1}|| + |V^{\text{soft}}_{\theta_{t-1},\pi_t}(s) - V^{\text{soft}}_{\theta_{t-1}}(s)| + C_Q||\theta_t-\theta_{t-1}||,\\
&\overset{(g)}{\leq} 2C_Q||\theta_t-\theta_{t-1}|| + V^{\text{soft}}_{\theta_{t-1}}(s) - \mathcal{T}^{\text{soft}}_{\theta_{t-1}}(V^{\text{soft}}_{\theta_{t-1},\pi_t})(s),\\
&\leq 2C_Q||\theta_t-\theta_{t-1}|| + \gamma|V^{\text{soft}}_{\theta_{t-1}}(s) - V^{\text{soft}}_{\theta_{t-1},\pi_t}(s)|,
\end{aligned}$$

where (f) follows Claim 2 and claim 5 and (e) follows Claim 4.

Now we take a look at the term $||\theta_t-\theta_{t-1}||$:

$$||\theta_t-\theta_{t-1}|| = \alpha_{t-1}||g_{t-1}||,$$

$$\begin{aligned}
&\le \alpha_{t-1}\Bigg[\bigg|\bigg|\sum_{t=0}^{\infty}\gamma^t\nabla_\theta r_{\theta_{t-1}}(s'_t,a'_t)\bigg|\bigg| + \bigg|\bigg|\sum_{t=0}^{\infty}\gamma^t\nabla_\theta r_{\theta_{t-1}}(s''_t,a''_t)\bigg|\bigg| + \frac{\lambda(1-\gamma^{t-1})}{1-\gamma}\bigg|\bigg|\theta_{t-1}-\bar{\theta}\bigg|\bigg|\Bigg],\\
&\overset{(h)}{\le} \alpha_{t-1}\bigg(\frac{\bar{C}_r}{1-\gamma}+\frac{\bar{C}_r}{1-\gamma}+\frac{\lambda(1-\gamma^{t-1})}{1-\gamma}\cdot\frac{2\bar{C}_r}{\lambda}\bigg),\\
&\le \frac{4\alpha_{t-1}\bar{C}_r}{1-\gamma},
\end{aligned}\tag{8}$$

where (h) follows claim 1. Recall that

$$|\log\pi_{t+1}(a|s)-\log\pi_{\theta_t}(a|s)| \le |Q^{\text{soft}}_{\theta_t,\pi_t}(s,a)-Q^{\text{soft}}_{\theta_t}(s,a)| + |V^{\text{soft}}_{\theta_t,\pi_t}(s)-V^{\text{soft}}_{\theta_t}(s)|.$$

Summing from $i=0$ to t, we get

$$\begin{aligned}
&\sum_{i=1}^{t}\bigg[|Q^{\text{soft}}_{\theta_i,\pi_i}(s,a)-Q^{\text{soft}}_{\theta_i}(s,a)|+|V^{\text{soft}}_{\theta_i,\pi_i}(s)-V^{\text{soft}}_{\theta_i}(s)|\bigg],\\
&\le \sum_{i=1}^{t}\bigg[4C_Q||\theta_i-\theta_{i-1}||+\gamma\bigg(|Q^{\text{soft}}_{\theta_{i-1}}(s,a)-Q^{\text{soft}}_{\theta_{i-1},\pi_{i-1}}(s,a)|+|V^{\text{soft}}_{\theta_{i-1},\pi_{i-1}}(s)-V^{\text{soft}}_{\theta_{i-1}}(s)|\bigg)\bigg],\\
&\Rightarrow (1-\gamma)\sum_{i=0}^{t-1}\bigg[|Q^{\text{soft}}_{\theta_i,\pi_i}(s,a)-Q^{\text{soft}}_{\theta_i}(s,a)|+|V^{\text{soft}}_{\theta_i,\pi_i}(s)-V^{\text{soft}}_{\theta_i}(s)|\bigg],\\
&\overset{(i)}{\le} \frac{16C_Q\bar{C}_r}{1-\gamma}\sum_{i=1}^{t}\alpha_{i-1}+\bigg(|Q^{\text{soft}}_{\theta_0}(s,a)-Q^{\text{soft}}_{\theta_0,\pi_0}(s,a)|+|V^{\text{soft}}_{\theta_0,\pi_0}(s)-V^{\text{soft}}_{\theta_0}(s)|\\
&-|Q^{\text{soft}}_{\theta_t}(s,a)-Q^{\text{soft}}_{\theta_t,\pi_t}(s,a)|-|V^{\text{soft}}_{\theta_t,\pi_t}(s)-V^{\text{soft}}_{\theta_t}(s)|\bigg),\\
&\Rightarrow \frac{1}{t}\sum_{i=0}^{t-1}\bigg[|Q^{\text{soft}}_{\theta_i,\pi_i}(s,a)-Q^{\text{soft}}_{\theta_i}(s,a)|+|V^{\text{soft}}_{\theta_i,\pi_i}(s)-V^{\text{soft}}_{\theta_i}(s)|\bigg]\\
&\le \frac{16C_Q\bar{C}_r}{t(1-\gamma)^2}\sum_{i=1}^{t}\alpha_{i-1}+\frac{1}{t(1-\gamma)}\bigg(|Q^{\text{soft}}_{\theta_0}(s,a)-Q^{\text{soft}}_{\theta_0,\pi_0}(s,a)|+|V^{\text{soft}}_{\theta_0,\pi_0}(s)-V^{\text{soft}}_{\theta_0}(s)|\\
&-|Q^{\text{soft}}_{\theta_t}(s,a)-Q^{\text{soft}}_{\theta_t,\pi_t}(s,a)|-|V^{\text{soft}}_{\theta_t,\pi_t}(s)-V^{\text{soft}}_{\theta_t}(s)|\bigg),\\
&= \frac{\bar{D}_1}{\sqrt{t}}+\frac{\bar{D}_2}{t},
\end{aligned}$$

where (i) follows (8), $\bar{D}_1=\frac{16C_Q\bar{C}_r}{(1-\gamma)^2}$, and $\bar{D}_2=\frac{1}{1-\gamma}\bigg(|Q^{\text{soft}}_{\theta_0}(s,a)-Q^{\text{soft}}_{\theta_0,\pi_0}(s,a)|+|V^{\text{soft}}_{\theta_0,\pi_0}(s)-V^{\text{soft}}_{\theta_0}(s)|-|Q^{\text{soft}}_{\theta_t}(s,a)-Q^{\text{soft}}_{\theta_t,\pi_t}(s,a)|-|V^{\text{soft}}_{\theta_t,\pi_t}(s)-V^{\text{soft}}_{\theta_t}(s)|\bigg)$. Therefore, we can see that

$$\frac{1}{t}\sum_{i=0}^{t-1}|\log\pi_{i+1}(a|s)-\log\pi_{\theta_i}(a|s)| \le \frac{\bar{D}_1}{\sqrt{t}}+\frac{\bar{D}_2}{t}.\tag{9}$$

Define the loss function $\bar{L}(\theta)\triangleq E_{(S,A)\sim\mu^{\pi_E}(\cdot,\cdot)}\big[-\log\pi_\theta(A|S)+\frac{\lambda}{2}||\theta-\bar{\theta}||^2\big]$, then we can see that $E_{(S_i^E,A_i^E)\sim\mu^{\pi_E}(\cdot,\cdot)}[L_i(\theta;(S_i^E,A_i^E))]=\gamma^i\bar{L}(\theta)$. Moreover, we have that

$$\begin{aligned}
&||\nabla L_i(\theta_t;(S_i^E,A_i^E))|| \overset{(j)}{\le} \gamma^i\lambda||\theta_t-\bar{\theta}||\\
&+\gamma^i||E^{\pi_{\theta_t}}_{S,A}[\sum_{k=0}^{\infty}\gamma^k\nabla_\theta r_{\theta_t}(S_k,A_k)|S_0=S_i^E]-E^{\pi_{\theta_t}}_{S,A}[\sum_{k=0}^{\infty}\gamma^k\nabla_\theta r_{\theta_t}(S_k,A_k)|S_0=S_i^E,A_0=A_i^E]||,\\
&\overset{(k)}{\le} 2\gamma^i\bar{C}_r+\frac{2\gamma^i\bar{C}_r}{1-\gamma},
\end{aligned}\tag{10}$$

where (j) follows Lemma 3 and (k) follows Claim 1.

Therefore, we have that

$$\begin{aligned}
&E_{(S_i^E,A_i^E)\sim\mu^{\pi_E}(\cdot,\cdot)}[||\frac{\nabla L_i(\theta_t;(S_i^E,A_i^E))}{\gamma^i}-\nabla\bar{L}(\theta_t)||^2] \overset{(l')}{\leq} \left[2\bar{C}_r+\frac{2\bar{C}_r}{1-\gamma}\right]^2,\\
&\Rightarrow E_{(S_i^E,A_i^E)\sim\mu^{\pi_E}(\cdot,\cdot)}[||\frac{1}{t}\sum_{i=0}^{t-1}L_i(\theta_t;(S_i^E,A_i^E))-\gamma^i\bar{L}(\theta_t)||^2],\\
&=\left(\frac{1-\gamma^t}{1-\gamma}\right)^2 E_{(S_i^E,A_i^E)\sim\mu^{\pi_E}(\cdot,\cdot)}[||\frac{1}{t}\sum_{i=0}^{t-1}\frac{L_i(\theta_t;(S_i^E,A_i^E))}{\gamma^i}-\bar{L}(\theta_t)||^2],\\
&\leq\frac{1}{t}\cdot\frac{4\bar{C}_r^2(2-\gamma)^2}{(1-\gamma)^4}, \qquad (11)
\end{aligned}$$

where (l') follows the fact that a bounded variable $X\in[-a,a]$ has bounded variance at most a^2. Now we take a look at the term $g_t-\frac{1-\gamma^{t+1}}{1-\gamma}\nabla\bar{L}(\theta_t)$:

$$\begin{aligned}
&E_{(S,A)\sim\mu^{\pi_E}(\cdot,\cdot)}\left[g_t-\frac{1-\gamma^{t+1}}{1-\gamma}\nabla\bar{L}(\theta_t)\right],\\
&=E_{\{(S_i^E,A_i^E)\sim\mu^{\pi_E}(\cdot,\cdot)\}_{i\geq 0}}\left[g_t-\sum_{i=0}^{t}\nabla L_i(\theta_t;(S_i^E,A_i^E))\right]\\
&+E_{\{(S_i^E,A_i^E)\sim\mu^{\pi_E}(\cdot,\cdot)\}_{i\geq 0}}\left[\sum_{i=0}^{t}\nabla L_i(\theta_t;(S_i^E,A_i^E))-\frac{1-\gamma^{t+1}}{1-\gamma}\nabla\bar{L}(\theta_t)\right],\\
&=E_{\{(S_i^E,A_i^E)\sim\mu^{\pi_E}(\cdot,\cdot)\}_{i\geq 0}}\left[g_t-\sum_{i=0}^{t}\nabla L_i(\theta_t;(S_i^E,A_i^E))\right],\\
&=E_{\{(S_i^E,A_i^E)\sim\mu^{\pi_E}(\cdot,\cdot)\}_{i\geq 0}}\left[\sum_{i=0}^{\infty}\gamma^i\nabla_\theta r_{\theta_t}(s_i',a_i')-E_{S,A}^{\pi_{\theta_t}}[\sum_{i=0}^{\infty}\gamma^i\nabla_\theta r_{\theta_t}(S_i,A_i)|S_0=S_0^E]\right]\\
&+E_{\{(S_i^E,A_i^E)\sim\mu^{\pi_E}(\cdot,\cdot)\}_{i\geq 0}}\left[\sum_{i=t+1}^{\infty}\gamma^i\nabla_\theta r_{\theta_t}(s_i'',a_i'')\right.\\
&\left.-E_{S,A}^{\pi_{\theta_t}}[\sum_{i=t+1}^{\infty}\gamma^i\nabla_\theta r_{\theta_t}(S_i,A_i)|S_t=S_t^E,A_t=A_t^E]\right],\\
&=E_{\{(S_i^E,A_i^E)\sim\mu^{\pi_E}(\cdot,\cdot)\}_{i\geq 0}}\left[E_{S,A}^{\pi_{t+1}}[\sum_{i=0}^{\infty}\gamma^i\nabla_\theta r_{\theta_t}(S_i,A_i)|S_0=S_0^E]\right.\\
&-E_{S,A}^{\pi_{\theta_t}}[\sum_{i=0}^{\infty}\gamma^i\nabla_\theta r_{\theta_t}(S_i,A_i)|S_0=S_0^E]+E_{S,A}^{\pi_{t+1}}[\sum_{i=t+1}^{\infty}\gamma^i\nabla_\theta r_{\theta_t}(S_i,A_i)|S_t=S_t^E,A_t=A_t^E]\\
&\left.-E_{S,A}^{\pi_{\theta_t}}[\sum_{i=t+1}^{\infty}\gamma^i\nabla_\theta r_{\theta_t}(S_i,A_i)|S_t=S_t^E,A_t=A_t^E]\right].
\end{aligned}$$

From equation (64) in [5], we know that

$$\begin{aligned}
&\left|\left|E_{S,A}^{\pi_{t+1}}[\sum_{i=0}^{\infty}\gamma^i\nabla_\theta r_{\theta_t}(S_i,A_i)|S_0=s_0]-E_{S,A}^{\pi_{\theta_t}}[\sum_{i=0}^{\infty}\gamma^i\nabla_\theta r_{\theta_t}(S_i,A_i)|S_0=s_0]\right|\right|,\\
&\leq\frac{2\bar{C}_r}{1-\gamma}\int_{s\in\mathcal{S}}\int_{a\in\mathcal{A}}|Q_{\theta_t}^{\text{soft}}(s,a)-Q_{\theta_t,\pi_t}^{\text{soft}}(s,a)|dads,\\
&\leq\frac{2\bar{C}_rC_d}{1-\gamma}\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}}\{|Q_{\theta_t}^{\text{soft}}(s,a)-Q_{\theta_t,\pi_t}^{\text{soft}}(s,a)|\},
\end{aligned}$$

where C_d is the product of the area of $\mathcal{S}$ and the area of $\mathcal{A}$.

Therefore, we can get that

$$E_{(S,A)\sim\mu^{\pi_E}}\left[\left|\left|g_t-\frac{1-\gamma^{t+1}}{1-\gamma}\nabla\bar{L}(\theta_t)\right|\right|\right]\leq\frac{4\bar{C}_rC_d}{1-\gamma}\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}}\{|Q^{\text{soft}}_{\theta_t}(s,a)-Q^{\text{soft}}_{\theta_t,\pi_t}(s,a)|\}. \quad (12)$$

From (7) and (8), we know that

$$\begin{aligned}
&|Q^{\text{soft}}_{\theta_i,\pi_i}(s,a)-Q^{\text{soft}}_{\theta_i}(s,a)|\leq\gamma|Q^{\text{soft}}_{\theta_{i-1},\pi_{i-1}}(s,a)-Q^{\text{soft}}_{\theta_{i-1}}(s,a)|+\frac{8\alpha_{i-1}C_Q\bar{C}_r}{1-\gamma},\\
&\Rightarrow\alpha_i|Q^{\text{soft}}_{\theta_i,\pi_i}(s,a)-Q^{\text{soft}}_{\theta_i}(s,a)|\leq\alpha_{i-1}\gamma|Q^{\text{soft}}_{\theta_{i-1},\pi_{i-1}}(s,a)-Q^{\text{soft}}_{\theta_{i-1}}(s,a)|+\frac{8\alpha^2_{i-1}C_Q\bar{C}_r}{1-\gamma},\\
&\Rightarrow\sum_{i=1}^{t}\alpha_i|Q^{\text{soft}}_{\theta_i,\pi_i}(s,a)-Q^{\text{soft}}_{\theta_i}(s,a)|\leq\sum_{i=0}^{t-1}\alpha_i\gamma|Q^{\text{soft}}_{\theta_i,\pi_i}(s,a)-Q^{\text{soft}}_{\theta_i}(s,a)|+\sum_{i=0}^{t-1}\frac{8\alpha^2_iC_Q\bar{C}_r}{1-\gamma},\\
&\Rightarrow(1-\gamma)\sum_{i=0}^{t-1}\alpha_i|Q^{\text{soft}}_{\theta_i,\pi_i}(s,a)-Q^{\text{soft}}_{\theta_i}(s,a)|,\\
&\leq\alpha_0|Q^{\text{soft}}_{\theta_0,\pi_0}(s,a)-Q^{\text{soft}}_{\theta_0}(s,a)|-\alpha_t|Q^{\text{soft}}_{\theta_t,\pi_t}(s,a)-Q^{\text{soft}}_{\theta_t}(s,a)|+\sum_{i=0}^{t-1}\frac{8\alpha^2_iC_Q\bar{C}_r}{1-\gamma}, \qquad (13)
\end{aligned}$$

and similarly we can see that

$$\begin{aligned}
&\sum_{i=1}^{t}|Q^{\text{soft}}_{\theta_i,\pi_i}(s,a)-Q^{\text{soft}}_{\theta_i}(s,a)|\leq\sum_{i=0}^{t-1}\gamma|Q^{\text{soft}}_{\theta_i,\pi_i}(s,a)-Q^{\text{soft}}_{\theta_i}(s,a)|+\frac{8\alpha_iC_Q\bar{C}_r}{1-\gamma},\\
&\Rightarrow(1-\gamma)\sum_{i=1}^{t-1}|Q^{\text{soft}}_{\theta_i,\pi_i}(s,a)-Q^{\text{soft}}_{\theta_i}(s,a)|,\\
&\leq|Q^{\text{soft}}_{\theta_0,\pi_0}(s,a)-Q^{\text{soft}}_{\theta_0}(s,a)|-|Q^{\text{soft}}_{\theta_t,\pi_t}(s,a)-Q^{\text{soft}}_{\theta_t}(s,a)|+\sum_{i=0}^{t-1}\frac{8\alpha_iC_Q\bar{C}_r}{1-\gamma}, \qquad (14)
\end{aligned}$$

Telescoping from $i=0$ to $t-1$, we get

$$\begin{aligned}
&\sum_{i=0}^{t-1}\alpha_iE_{(S,A)\sim\mu^{\pi_E}}\left[\left|\left|g_i-\frac{1-\gamma^{i+1}}{1-\gamma}\nabla\bar{L}(\theta_i)\right|\right|\right],\\
&\overset{(l)}{\leq}\frac{4\bar{C}_rC_d}{1-\gamma}\sum_{i=0}^{t-1}\alpha_i|Q^{\text{soft}}_{\theta_i,\pi_i}(S,A)-Q^{\text{soft}}_{\theta_i}(S,A)|,\\
&\overset{(m)}{\leq}\frac{4\bar{C}_rC_d}{(1-\gamma)^2}\left[\alpha_0|Q^{\text{soft}}_{\theta_0,\pi_0}(S,A)-Q^{\text{soft}}_{\theta_0}(S,A)|-\alpha_t|Q^{\text{soft}}_{\theta_t,\pi_t}(S,A)-Q^{\text{soft}}_{\theta_t}(S,A)|+\sum_{i=0}^{t-1}\frac{8\alpha_i^2C_Q\bar{C}_r}{1-\gamma}\right], \qquad (15)
\end{aligned}$$

where (l) follows (12) and (m) follows (13). Similarly, we can see that

$$\begin{aligned}
&\sum_{i=0}^{t-1}E_{(S,A)\sim\mu^{\pi_E}}\left[\left|\left|g_i-\frac{1-\gamma^{i+1}}{1-\gamma}\nabla\bar{L}(\theta_i)\right|\right|\right],\\
&\leq\frac{4\bar{C}_rC_d}{1-\gamma}\sum_{i=0}^{t-1}|Q^{\text{soft}}_{\theta_i,\pi_i}(S,A)-Q^{\text{soft}}_{\theta_i}(S,A)|,\\
&\leq\frac{4\bar{C}_rC_d}{(1-\gamma)^2}\left[|Q^{\text{soft}}_{\theta_0,\pi_0}(S,A)-Q^{\text{soft}}_{\theta_0}(S,A)|-|Q^{\text{soft}}_{\theta_t,\pi_t}(S,A)-Q^{\text{soft}}_{\theta_t}(S,A)|+\sum_{i=0}^{t-1}\frac{8\alpha_iC_Q\bar{C}_r}{1-\gamma}\right], \qquad (16)
\end{aligned}$$

Now, we start to quantify the local regret:

$$
\begin{aligned}
&\bar{L}(\theta_{i+1}) \geq \bar{L}(\theta_i) + [\nabla \bar{L}(\theta_i)]^\top (\theta_{i+1} - \theta_i) - \frac{C_L}{2}||\theta_{i+1} - \theta_i||^2, \\
&\overset{(n)}{\geq} \bar{L}(\theta_i) + \alpha_t [\nabla \bar{L}(\theta_i)]^\top g_i - \frac{8\alpha_i^2 \bar{C}_r^2 C_L}{(1-\gamma)^2}, \\
&= \bar{L}(\theta_i) + \alpha_i \frac{1-\gamma^{i+1}}{1-\gamma}||\nabla \bar{L}(\theta_i)||^2 + \alpha_i [\nabla \bar{L}(\theta_i)]^\top \Big(g_i - \frac{1-\gamma^{i+1}}{1-\gamma}\nabla \bar{L}(\theta_i)\Big) - \frac{8\alpha_i^2 \bar{C}_r^2 C_L}{(1-\gamma)^2}, \\
&\geq \bar{L}(\theta_i) + \alpha_i \frac{1-\gamma^{i+1}}{1-\gamma}||\nabla \bar{L}(\theta_i)||^2 - \alpha_i ||\nabla \bar{L}(\theta_i)|| \cdot ||g_i - \frac{1-\gamma^{i+1}}{1-\gamma}\nabla \bar{L}(\theta_i)|| - \frac{8\alpha_i^2 \bar{C}_r^2 C_L}{(1-\gamma)^2}, \\
&\overset{(o)}{\geq} \bar{L}(\theta_i) + \alpha_i \frac{1-\gamma^{i+1}}{1-\gamma}||\nabla \bar{L}(\theta_i)||^2 - \alpha_i \frac{2\bar{C}_r(2-\gamma)}{1-\gamma} \cdot ||g_i - \frac{1-\gamma^{i+1}}{1-\gamma}\nabla \bar{L}(\theta_i)|| - \frac{8\alpha_i^2 \bar{C}_r^2 C_L}{(1-\gamma)^2}, \\
&\Rightarrow E_{\{(S_i^E, A_i^E) \sim \mu^{\pi_E}(\cdot,\cdot)\}_{i \geq 0}} \Bigg[\sum_{i=0}^{t-1} \alpha_i \frac{1-\gamma^{i+1}}{1-\gamma}||\nabla \bar{L}(\theta_i)||^2\Bigg], \\
&\leq \bar{L}(\theta_t) - \bar{L}(\theta_0) + \frac{2\bar{C}_r(2-\gamma)}{1-\gamma}\sum_{i=0}^{t-1} \alpha_i E_{\{(S_i^E, A_i^E) \sim \mu^{\pi_E}(\cdot,\cdot)\}_{i \geq 0}} \Bigg[||g_i - \frac{1-\gamma^{i+1}}{1-\gamma}\nabla \bar{L}(\theta_i)||\Bigg] \\
&+ \sum_{i=0}^{t-1} \frac{8\alpha_i^2 \bar{C}_r^2 C_L}{(1-\gamma)^2}, \\
&\overset{(p)}{\leq} \bar{L}(\theta_t) - \bar{L}(\theta_0) + \frac{8\bar{C}_r^2 C_d (2-\gamma)}{(1-\gamma)^3} E_{\{(S_i^E, A_i^E) \sim \mu^{\pi_E}(\cdot,\cdot)\}_{i \geq 0}} \Bigg[\alpha_0 |Q^{\text{soft}}_{\theta_0,\pi_0}(S_0^E, A_0^E) - Q^{\text{soft}}_{\theta_0}(S_0^E, A_0^E)| \\
&- \alpha_t |Q^{\text{soft}}_{\theta_t,\pi_t}(S_t^E, A_t^E) - Q^{\text{soft}}_{\theta_t}(S_t^E, A_t^E)|\Bigg] + \Bigg(\frac{32 C_Q \bar{C}_r^2 C_d}{(1-\gamma)^3} + \frac{8\bar{C}_r^2 C_L}{(1-\gamma)^2}\Bigg)\sum_{i=0}^{t-1}\alpha_i^2,
\end{aligned}
$$

where (n) follows (8), (o) follows (10), (p) follows (15). Therefore, we have that

$$
\begin{aligned}
&E_{\{(S_i^E, A_i^E) \sim \mu^{\pi_E}(\cdot,\cdot)\}_{i \geq 0}} \Bigg[\sum_{t=0}^{T-1} \alpha_{T-1}||\nabla \bar{L}(\theta_t)||^2\Bigg], \qquad (17)\\
&\leq E_{\{(S_i^E, A_i^E) \sim \mu^{\pi_E}(\cdot,\cdot)\}_{i \geq 0}} \Bigg[\sum_{t=0}^{T-1} \alpha_t \frac{1-\gamma^t}{1-\gamma}||\nabla \bar{L}(\theta_t)||^2\Bigg], \\
&\leq \bar{L}(\theta_T) - \bar{L}(\theta_0) + \frac{8\bar{C}_r^2 C_d (2-\gamma)}{(1-\gamma)^3} E_{\{(S_i^E, A_i^E) \sim \mu^{\pi_E}(\cdot,\cdot)\}_{i \geq 0}} \Bigg[\alpha_0 |Q^{\text{soft}}_{\theta_0,\pi_0}(S_0^E, A_0^E) - Q^{\text{soft}}_{\theta_0}(S_0^E, A_0^E)| \\
&- \alpha_T |Q^{\text{soft}}_{\theta_T,\pi_T}(S_T^E, A_T^E) - Q^{\text{soft}}_{\theta_T}(S_T^E, A_T^E)|\Bigg] + \Bigg(\frac{32 C_Q \bar{C}_r^2 C_d}{(1-\gamma)^3} + \frac{8\bar{C}_r^2 C_L}{(1-\gamma)^2}\Bigg)\sum_{i=0}^{T-1}\alpha_i^2, \\
&\Rightarrow E_{\{(S_i^E, A_i^E) \sim \mu^{\pi_E}(\cdot,\cdot)\}_{i \geq 0}} \Bigg[\sum_{t=0}^{T-1} ||\nabla \bar{L}(\theta_t)||^2\Bigg] \leq D_2\sqrt{T} + D_3\sqrt{T}(\log T + 1), \qquad (18)
\end{aligned}
$$

where $D_2 = \bar{L}(\theta_T) - \bar{L}(\theta_0) + \frac{8\bar{C}_r^2 C_d(2-\gamma)}{(1-\gamma)^3} E_{\{(S_i^E, A_i^E) \sim \mu^{\pi_E}(\cdot,\cdot)\}_{i \geq 0}} \Big[\alpha_0 |Q^{\text{soft}}_{\theta_0,\pi_0}(S_0^E, A_0^E) - Q^{\text{soft}}_{\theta_0}(S_0^E, A_0^E)| - \alpha_T |Q^{\text{soft}}_{\theta_T,\pi_T}(S_T^E, A_T^E) - Q^{\text{soft}}_{\theta_T}(S_T^E, A_T^E)|\Big]$ and $D_3 = \frac{2(1-\gamma)}{\lambda}\Big(\frac{32 C_Q \bar{C}_r^2 C_d}{(1-\gamma)^3} + \frac{8\bar{C}_r^2 C_L}{(1-\gamma)^2}\Big)$. Then we can see that

$$
\begin{aligned}
&E_{\{(S_i^E, A_i^E) \sim \mu^{\pi_E}(\cdot,\cdot)\}_{i \geq 0}} \Bigg[\sum_{t=0}^{T-1} ||\frac{1}{t+1}\sum_{i=0}^{t} \nabla L_i(\theta_t; (S_i^E, A_i^E))||^2\Bigg], \\
&\leq 2E_{\{(S_i^E, A_i^E) \sim \mu^{\pi_E}(\cdot,\cdot)\}_{i \geq 0}} \Bigg[\sum_{t=0}^{T-1} ||\frac{1}{t+1}\sum_{i=0}^{t} \nabla L_i(\theta_t; (S_i^E, A_i^E)) - \gamma^i \nabla \bar{L}(\theta_t)||^2\Bigg]
\end{aligned}
$$

$$
\begin{aligned}
&+ 2E_{\{(S_i^E,A_i^E)\sim\mu^{\pi_E}(\cdot,\cdot)\}_{i\geq 0}}\left[\sum_{t=0}^{T-1} ||\frac{1-\gamma^t}{(t+1)(1-\gamma)}\nabla\bar{L}(\theta_t)||^2\right],\\
&\leq 2E_{\{(S_i^E,A_i^E)\sim\mu^{\pi_E}(\cdot,\cdot)\}_{i\geq 0}}\left[\sum_{t=0}^{T-1} ||\frac{1}{t+1}\sum_{i=0}^{t}\nabla L_i(\theta_t;(S_i^E,A_i^E)) - \gamma^i\nabla\bar{L}(\theta_t)||^2\right]\\
&+ 2E_{\{(S_i^E,A_i^E)\sim\mu^{\pi_E}(\cdot,\cdot)\}_{i\geq 0}}\left[\sum_{t=0}^{T-1} ||\nabla\bar{L}(\theta_t)||^2\right],\\
&\overset{(q)}{\leq} \sum_{t=0}^{T-1}\frac{8\bar{C}_r^2(2-\gamma)^2}{(t+1)(1-\gamma)^4} + 2E_{\{(S_i^E,A_i^E)\sim\mu^{\pi_E}(\cdot,\cdot)\}_{i\geq 0}}\left[\sum_{t=0}^{T-1} ||\nabla\bar{L}(\theta_t)||^2\right],\\
&\leq D_1(\log T+1) + D_2\sqrt{T} + D_3\sqrt{T}(\log T+1),
\end{aligned}
$$

where $D_1 = \frac{8\bar{C}_r^2(2-\gamma)^2}{(1-\gamma)^4}$. The (q) follows (11). Note that to achieve this local regret rate, we actually need to take an extra expectation over the dynamics P because we need to roll out π_t to formulate g_t. Here, we omit the expectation over the dynamics.

B.5 Proof of Theorem 1

We know that

$$
\begin{aligned}
&E_{\{(S_i^E,A_i^E)\sim\mu^{\pi_E}(\cdot,\cdot)\}_{i\geq 0}}\left[\sum_{i=0}^{t} ||\nabla L_i(\theta_t;(S_i^E,A_i^E))||^2\right],\\
&\leq E_{\{(S_i^E,A_i^E)\sim\mu^{\pi_E}(\cdot,\cdot)\}_{i\geq 0}}\left[\sum_{i=0}^{t} ||\gamma^i\nabla\bar{L}(\theta_t)||^2\right]\\
&+ E_{\{(S_i^E,A_i^E)\sim\mu^{\pi_E}(\cdot,\cdot)\}_{i\geq 0}}\left[\sum_{i=0}^{t} ||\nabla L_i(\theta_t;(S_i^E,A_i^E)) - \gamma^i\nabla\bar{L}(\theta_t)||^2\right],\\
&\leq E_{\{(S_i^E,A_i^E)\sim\mu^{\pi_E}(\cdot,\cdot)\}_{i\geq 0}}\left[\sum_{i=0}^{t} ||\nabla\bar{L}(\theta_t)||^2\right]\\
&+ E_{\{(S_i^E,A_i^E)\sim\mu^{\pi_E}(\cdot,\cdot)\}_{i\geq 0}}\left[\sum_{i=0}^{t} ||\nabla L_i(\theta_t;(S_i^E,A_i^E)) - \gamma^i\nabla\bar{L}(\theta_t)||^2\right],\\
&\overset{(a)}{\leq} O(\sqrt{t+1}+\sqrt{t+1}\log(t+1)) + \sum_{i=0}^{t}\gamma^{2i}\cdot 4\bar{C}_r^2\left(\frac{2-\gamma}{1-\gamma}\right)^2,\\
&\Rightarrow E_{\{(S_i^E,A_i^E)\sim\mu^{\pi_E}(\cdot,\cdot)\}_{i\geq 0}}\left[\frac{1}{t+1}\sum_{i=0}^{t} ||\nabla L_i(\theta_t;(S_i^E,A_i^E))||^2\right] \leq O(\frac{1}{\sqrt{t+1}} + \frac{\log(t+1)}{\sqrt{t+1}} + \frac{1}{t+1}),
\end{aligned}
\tag{19}
$$

where (a) follows (18) and (11).

Therefore,

$$
\begin{aligned}
&\sum_{t=0}^{T-1} E_{\{(S_i^E,A_i^E)\sim\mathbb{P}_i^{\pi_E}(\cdot,\cdot)\}_{i\geq 0}}\left[||\frac{1}{t+1}\sum_{i=0}^{t}\nabla L_i(\theta_t;(S_i^E,A_i^E))||^2\right],\\
&\leq \sum_{t=0}^{T-1} E_{\{(S_i^E,A_i^E)\sim\mathbb{P}_i^{\pi_E}(\cdot,\cdot)\}_{i\geq 0}}\left[\frac{1}{t+1}\sum_{i=0}^{t} ||\nabla L_i(\theta_t;(S_i^E,A_i^E))||^2\right],\\
&\leq \sum_{t=0}^{T-1} E_{\{(S_i^E,A_i^E)\sim\mu^{\pi_E}(\cdot,\cdot)\}_{i\geq 0}}\left[\frac{1}{t+1}\sum_{i=0}^{t} ||\nabla L_i(\theta_t;(S_i^E,A_i^E))||^2\right]
\end{aligned}
$$

$$
\begin{aligned}
&+\sum_{t=0}^{T-1}\frac{1}{t+1}\sum_{i=0}^{t}\Big[E_{\{(S_i^E,A_i^E)\sim\mathbb{P}_i^{\pi_E}(\cdot,\cdot)\}_{i\geq 0}}[||\nabla L_i(\theta_t;(S_i^E,A_i^E))||^2\\
&- E_{\{(S_i^E,A_i^E)\sim\mu^{\pi_E}(\cdot,\cdot)\}_{i\geq 0}}[||\nabla L_i(\theta_t;(S_i^E,A_i^E))||^2]\Big],\\
&\overset{(s)}{\leq} D_1\log T + D_2\sqrt{T} + D_3\log T\sqrt{T} + \sum_{t=0}^{T-1}\frac{1}{t+1}\sum_{i=0}^{t}8C_M\bar{C}_r^2\left(\frac{2-\gamma}{1-\gamma}\right)^2\rho^i\gamma^{2i},\\
&\leq D_1\log T + D_2\sqrt{T} + D_3\log T\sqrt{T} + \sum_{t=0}^{T-1}\frac{1}{t+1}\cdot 8C_M\bar{C}_r^2\left(\frac{2-\gamma}{1-\gamma}\right)^2\frac{1}{1-\rho\gamma^2},\\
&\leq \left(D_1+\frac{8C_M\bar{C}_r^2(2-\gamma)^2}{(1-\rho\gamma^2)(1-\gamma)^2}\right)\log T + D_2\sqrt{T} + D_3\log T\sqrt{T},
\end{aligned}
$$

where (s) follows (19) and Proposition 1. Note that to achieve this local regret rate, we actually need to take an extra expectation over the dynamics P because we need to roll out π_t to formulate g_t. Here, we omit the expectation over the dynamics.

B.6 Proof of Theorem 2

Suppose the expert reward function r_E and the parameterized reward function r_θ are both linear, i.e., $r_E = \theta_E^\top\phi$ and $r_\theta = \theta^\top\phi$, where $\phi : \mathcal{S}\times\mathcal{A}\to\mathbb{R}_+^n$ is an n-dimensional feature vector such that $||\phi(s,a)||\leq\bar{C}_r$ for any $(s,a)\in\mathcal{S}\times\mathcal{A}$. The proof follows the similar idea with that of Theorem 1 where follow two-step process: step (i) quantifies the regret under the stationary distribution $\mu^{\pi_E}(\cdot,\cdot)$ and step (ii) quantifies the difference between the correlated distribution $\mathbb{P}_t^{\pi_E}$ and the stationary distribution μ^{π_E}. We start our proof with the following claim:

Claim 6. *If the parameterized reward function r_θ is linear, the function $\bar{L}(\theta)$ is λ-strongly convex for any θ.*

Proof. Recall that $\bar{L}(\theta) = E_{(\bar{S},\bar{A})\sim\mu^{\pi_E}(\cdot,\cdot)}[-\log\pi_\theta(\bar{A}|\bar{S})] + \frac{\lambda}{2}||\theta-\bar{\theta}||^2$. Define $\bar{L}(\theta;(S,A)) \triangleq -\log\pi_\theta(\bar{A}|\bar{S}) + \frac{\lambda}{2}||\theta-\bar{\theta}||^2$, from Lemma 3, we can see that

$$\nabla\bar{L}(\theta;(\bar{S},\bar{A})) = E_{S,A}^{\pi_\theta}\left[\sum_{i=0}^{\infty}\gamma^i\phi(S_i,A_i)\middle|S_0=\bar{S}\right] - E_{S,A}^{\pi_\theta}\left[\sum_{i=0}^{\infty}\gamma^i\phi(S_i,A_i)\middle|S_0=\bar{S},A_0=\bar{A}\right] + \lambda(\theta-\bar{\theta}).$$

Therefore, we have that

$$
\begin{aligned}
&\nabla_{\theta\theta}^2\bar{L}(\theta;(\bar{S},\bar{A})),\\
&= \nabla_\theta E_{S,A}^{\pi_\theta}\left[\sum_{i=0}^{\infty}\gamma^i\phi(S_i,A_i)\middle|S_0=\bar{S}\right] - \nabla_\theta E_{S,A}^{\pi_\theta}\left[\sum_{i=0}^{\infty}\gamma^i\phi(S_i,A_i)\middle|S_0=\bar{S},A_0=\bar{A}\right] + \lambda.
\end{aligned}
$$

Now, we take a look at

$$
\begin{aligned}
&\nabla_\theta E_{S,A}^{\pi_\theta}\left[\sum_{i=0}^{\infty}\gamma^i\phi(S_i,A_i)\middle|S_0=\bar{S},A_0=\bar{A}\right] = \nabla_\theta\Bigg\{\phi(\bar{S},\bar{A})\\
&+\int_{s_1\in\mathcal{S}}P(s_1|\bar{S},\bar{A})\int_{a_{t+1}\in\mathcal{A}}\pi_\theta(a_1|s_1)E_{S,A}^{\pi_\theta}\left[\sum_{i=1}^{\infty}\gamma^i\phi(S_i,A_i)\middle|S_1=s_1,A_1=a_1\right]da_1ds_1\Bigg\},\\
&= \int_{s_1\in\mathcal{S}}P(s_1|\bar{S},\bar{A})\int_{a_1\in\mathcal{A}}\Bigg\{\nabla_\theta\pi_\theta(a_1|s_1)\cdot E_{S,A}^{\pi_\theta}\left[\sum_{i=1}^{\infty}\gamma^i\phi(S_i,A_i)\middle|S_1=s_1,A_1=a_1\right]\\
&+\pi_\theta(a_1|s_1)\cdot\nabla_\theta E_{S,A}^{\pi_\theta}\left[\sum_{i=1}^{\infty}\gamma^i\phi(S_i,A_i)\middle|S_1=s_1,A_1=a_1\right]\Bigg\}da_1ds_1.
\end{aligned}
$$

Keep the expansion, we can see that

$$\nabla_\theta E_{S,A}^{\pi_\theta}\left[\sum_{i=0}^{\infty}\gamma^i\phi(S_i,A_i)\middle|S_0=\bar{S},A_0=\bar{A}\right],$$

$$
\begin{aligned}
&= E^{\pi_\theta}_{S,A}\bigg\{\sum_{i=1}^{\infty}\nabla_\theta\pi_\theta(A'_i|S'_i)\cdot E^{\pi_\theta}_{S,A}\bigg[\sum_{i=1}^{\infty}\gamma^i\phi(S_i,A_i)\bigg|S_1=S'_1,A_1=A'_1\bigg]\bigg|S'_0=\bar{S},A'_0=\bar{A}\bigg\},\\
&\nabla_\theta E^{\pi_\theta}_{S,A}\bigg[\sum_{i=0}^{\infty}\gamma^i\phi(S_i,A_i)\bigg|S_0=\bar{S}\bigg],\\
&= E^{\pi_\theta}_{S,A}\bigg\{\sum_{i=0}^{\infty}\nabla_\theta\pi_\theta(A'_i|S'_i)\cdot E^{\pi_\theta}_{S,A}\bigg[\sum_{i=0}^{\infty}\gamma^i\phi(S_i,A_i)\bigg|S_0=S'_0,A_0=A'_0\bigg]\bigg|S'_0=\bar{S}\bigg\},
\end{aligned}
$$

Thus we have that

$$
\begin{aligned}
&E_{(\bar{S},\bar{A})\sim\mu^{\pi_E}(\cdot,\cdot)}\bigg\{\nabla_\theta E^{\pi_\theta}_{S,A}\bigg[\sum_{i=0}^{\infty}\gamma^i\phi(S_i,A_i)\bigg|S_0=\bar{S}\bigg]-\nabla_\theta E^{\pi_\theta}_{S,A}\bigg[\sum_{i=0}^{\infty}\gamma^i\phi(S_i,A_i)\bigg|S_0=\bar{S},A_0=\bar{A}\bigg]\bigg\},\\
&= E_{(\bar{S},\bar{A})\sim\mu^{\pi_E}(\cdot,\cdot)}E^{\pi_\theta}_{S',A'}\bigg\{\nabla_\theta\pi_\theta(A'_0|S'_0)\cdot E^{\pi_\theta}_{S,A}\bigg[\sum_{i=0}^{\infty}\gamma^i\phi(S_i,A_i)\bigg|S_0=S'_0,A_0=A'_0\bigg]\bigg|S'_0=\bar{S}_0\bigg\}\\
&+ E_{(\bar{S},\bar{A})\sim\mu^{\pi_E}(\cdot,\cdot)}\bigg\{E^{\pi_\theta}_{S',A'}\bigg\{\sum_{i=1}^{\infty}\nabla_\theta\pi_\theta(A'_i|S'_i)\cdot E^{\pi_\theta}_{S,A}\bigg[\sum_{i=1}^{\infty}\gamma^i\phi(S_i,A_i)\bigg|S_1=S'_1,A_1=A'_1\bigg]\bigg|S'_0=\bar{S}\bigg\}\\
&- E^{\pi_\theta}_{S,A}\bigg\{\sum_{i=1}^{\infty}\nabla_\theta\pi_\theta(A'_i|S'_i)\cdot E^{\pi_\theta}_{S,A}\bigg[\sum_{i=1}^{\infty}\gamma^i\phi(S_i,A_i)\bigg|S_1=S'_1,A_1=A'_1\bigg]\bigg|S_0=\bar{S},A_0=\bar{A}\bigg\}\bigg\},\\
&= E_{(\bar{S},\bar{A})\sim\mu^{\pi_E}(\cdot,\cdot)}E^{\pi_\theta}_{S',A'}\bigg\{\nabla_\theta\pi_\theta(A'_0|S'_0)\cdot E^{\pi_\theta}_{S,A}\bigg[\sum_{i=0}^{\infty}\gamma^i\phi(S_i,A_i)\bigg|S_0=S'_0,A_0=A'_0\bigg]\bigg|S'_0=\bar{S}_0\bigg\}\\
&\overset{(a)}{=} E_{(\bar{S},\bar{A})\sim\mu^{\pi_E}(\cdot,\cdot)}E^{\pi_\theta}_{S',A'}\bigg\{\pi_\theta(A'_0|S^E_0)[X-EX]X|S_0=\bar{S}\bigg\},\\
&= \mathrm{Cov}(X)\succ 0, \qquad (20)
\end{aligned}
$$

where (a) follows Lemma 3, $X \triangleq E^{\pi_\theta}_{S,A}[\sum_{i=0}^{\infty}\gamma^i\phi(S_i,A_i)|S_0=\bar{S},A_0=A]$ is a random vector, $EX \triangleq E^{\pi_\theta}_{S,A}[\sum_{i=0}^{\infty}\gamma^i\phi(S_i,A_i)|S_0=\bar{S}]$ is the expectation of X over the action distribution, and $\mathrm{Cov}(X)$ is the covariance matrix of the random vector X with itself, which is always positive definite because the policy π_θ is always stochastic.

Therefore, we can see that

$$
||\nabla^2_{\theta\theta}\bar{L}(\theta)|| = ||\mathrm{Cov}(X)+\lambda|| \overset{(b)}{\geq} \lambda,
$$

where (b) follows (20). □

Step (i). We first quantify the regret under the stationary distribution $\mu^{\pi_E}(\cdot,\cdot)$. Suppose $\theta^* \in \arg\min\bar{L}(\theta) = \arg\min\frac{1-\gamma^T}{1-\gamma}\bar{L}(\theta) = \arg\min E_{(S^E_t,A^E_t)\sim\mu^{\pi_E}}[\sum_{t=0}^{T-1}L_t(\theta;(S^E_t,A^E_t))]$.

$$
\begin{aligned}
&||\theta_{t+1}-\theta^*||^2 = ||\theta_t-\alpha_t g_t-\theta^*||^2 = ||\theta_t-\theta^*||^2+\alpha_t^2||g_t||^2-2\alpha_t\langle g_t,\theta_t-\theta^*\rangle,\\
&\Rightarrow E[||\theta_{t+1}-\theta^*||^2],\\
&\overset{(c)}{\leq} E[||\theta_t-\theta^*||^2]+\frac{16\alpha_t^2\bar{C}_r^2}{(1-\gamma)^2}-2\alpha_t\langle\frac{1-\gamma^{t+1}}{1-\gamma}\nabla\bar{L}(\theta_t),\theta_t-\theta^*\rangle-2\alpha_t\langle g_t-\frac{1-\gamma^{t+1}}{1-\gamma}\nabla\bar{L}(\theta_t),\theta_t-\theta^*\rangle, \qquad (21)
\end{aligned}
$$

where (c) follows (8)

Since $\bar{L}(\theta)$ is λ-strongly convex, we have that

$$
\begin{aligned}
&\bar{L}(\theta^*) \geq \bar{L}(\theta_t)+\langle\nabla\bar{L}(\theta_t),\theta^*-\theta_t\rangle+\frac{\lambda}{2}||\theta_t-\theta^*||^2,\\
&\Rightarrow \bar{L}(\theta_t)-\bar{L}(\theta^*) \leq \langle\nabla\bar{L}(\theta_t),\theta_t-\theta^*\rangle-\frac{\lambda}{2}||\theta_t-\theta^*||^2,
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(d)}{\leq} \frac{1-\gamma}{2\alpha_t(1-\gamma^{t+1})}\Big[||\theta_t-\theta^*||^2-||\theta_{t+1}-\theta^*||^2\Big]-\frac{1-\gamma}{1-\gamma^{t+1}}\langle g_t-\frac{1-\gamma^{t+1}}{1-\gamma}\nabla\bar{L}(\theta_t),\theta_t-\theta^*\rangle\\
&+\frac{8\alpha_t\bar{C}_r^2}{(1-\gamma)(1-\gamma^{t+1})}-\frac{\lambda}{2}||\theta_t-\theta^*||^2,\\
&\Rightarrow E_{(S_t^E,A_t^E)\sim\mu^{\pi_E}(\cdot,\cdot)}[\bar{L}(\theta_t)-\bar{L}(\theta^*)],\\
&\leq \frac{1-\gamma-\lambda\alpha_t(1-\gamma^{t+1})}{2\alpha_t(1-\gamma^{t+1})}E_{(S_t^E,A_t^E)\sim\mu^{\pi_E}(\cdot,\cdot)}[||\theta_t-\theta^*||^2]-\frac{1-\gamma}{2\alpha_t(1-\gamma^{t+1})}E_{(S_t^E,A_t^E)\sim\mu^{\pi_E}(\cdot,\cdot)}[||\theta_{t+1}-\theta^*||^2]\\
&+\frac{8\alpha_t\bar{C}_r^2}{(1-\gamma)(1-\gamma^{t+1})}+\frac{1-\gamma}{1-\gamma^{t+1}}E_{(S_t^E,A_t^E)\sim\mu^{\pi_E}(\cdot,\cdot)}\Big[||g_t-\frac{1-\gamma^{t+1}}{1-\gamma}\nabla\bar{L}(\theta_t)||\cdot||\theta_t-\theta^*||\Big],\\
&\overset{(e)}{\leq} \frac{1-\gamma-\lambda\alpha_t(1-\gamma^{t+1})}{2\alpha_t(1-\gamma^{t+1})}E_{(S_t^E,A_t^E)\sim\mu^{\pi_E}(\cdot,\cdot)}[||\theta_t-\theta^*||^2]-\frac{1-\gamma}{2\alpha_t(1-\gamma^{t+1})}E_{(S_t^E,A_t^E)\sim\mu^{\pi_E}(\cdot,\cdot)}[||\theta_{t+1}-\theta^*||^2]\\
&+\frac{8\alpha_t\bar{C}_r^2}{(1-\gamma)(1-\gamma^{t+1})}+\frac{\hat{C}(1-\gamma)}{1-\gamma^{t+1}}E_{(S_t^E,A_t^E)\sim\mu^{\pi_E}(\cdot,\cdot)}\Big[||g_t-\frac{1-\gamma^{t+1}}{1-\gamma}\nabla\bar{L}(\theta_t)||\Big], \qquad (22)
\end{aligned}
$$

where (d) follows (21), and (e) follows Claim 1 which shows that the trajectory of θ_t is bounded and thus there is a positive constant $\hat{C}$ such that $||\theta_t-\theta^*||\leq\hat{C}$.

Telescoping (22) from $t=0$ to $t=T-1$, we can see that

$$
\begin{aligned}
&E_{(S_t^E,A_t^E)\sim\mu^{\pi_E}(\cdot,\cdot)}\Big[\sum_{t=0}^{T-1}L_t(\theta_t;(S_t^E,A_t^E))-\sum_{t=0}^{T-1}L_t(\theta^*;(S_t^E,A_t^E))\Big],\\
&=E_{(S_t^E,A_t^E)\sim\mu^{\pi_E}(\cdot,\cdot)}\Big[\sum_{t=0}^{T-1}\gamma^t\bar{L}(\theta_t)-\sum_{t=0}^{T-1}\gamma^t\bar{L}(\theta^*)\Big],\\
&\leq E_{(S_t^E,A_t^E)\sim\mu^{\pi_E}(\cdot,\cdot)}\Big[\sum_{t=0}^{T-1}\bar{L}(\theta_t)-\sum_{t=0}^{T-1}\bar{L}(\theta^*)\Big],\\
&\overset{(f)}{\leq}\frac{1-\gamma-\lambda\alpha_0}{2\alpha_0}E_{(S_t^E,A_t^E)\sim\mu^{\pi_E}(\cdot,\cdot)}[||\theta_t-\theta^*||^2]-\frac{1-\gamma}{2\alpha_{T-1}(1-\gamma^T)}E_{(S_t^E,A_t^E)\sim\mu^{\pi_E}(\cdot,\cdot)}[||\theta_T-\theta^*||^2]\\
&+\frac{4\bar{C}_rC_d\hat{C}}{(1-\gamma)^2}\Big[|Q^{\text{soft}}_{\theta_0,\pi_0}(S,A)-Q^{\text{soft}}_{\theta_0}(S,A)|-|Q^{\text{soft}}_{\theta_T,\pi_T}(S,A)-Q^{\text{soft}}_{\theta_T}(S,A)|+\sum_{i=0}^{t-1}\frac{8\alpha_iC_Q\bar{C}_r}{1-\gamma}\Big],\\
&\leq\bar{D}_4+\frac{32C_QC_d\hat{C}\bar{C}_r^2}{\lambda(1-\gamma)^3}(\log T+1), \qquad (23)
\end{aligned}
$$

where (f) follows (16), and $\bar{D}_4=\frac{1-\gamma-\lambda\alpha_0}{2\alpha_0}E_{(S_t^E,A_t^E)\sim\mu^{\pi_E}(\cdot,\cdot)}[||\theta_t-\theta^*||^2]-\frac{1-\gamma}{2\alpha_{T-1}(1-\gamma^T)}E_{(S_t^E,A_t^E)\sim\mu^{\pi_E}(\cdot,\cdot)}[||\theta_T-\theta^*||^2]+\frac{4\bar{C}_rC_d\hat{C}}{(1-\gamma)^2}[|Q^{\text{soft}}_{\theta_0,\pi_0}(S,A)-Q^{\text{soft}}_{\theta_0}(S,A)|-|Q^{\text{soft}}_{\theta_T,\pi_T}(S,A)-Q^{\text{soft}}_{\theta_T}(S,A)|]$.

Step (ii). We now quantify the difference between the stationary distribution $\mu^{\pi_E}(\cdot,\cdot)$ and the correlated distribution $\mathbb{P}_t^{\pi_E}(\cdot,\cdot)$. Note that $||\theta_t||$ is bounded (proved in Claim 1), the soft Bellman policy π_θ is continuous in θ and (s,a), and we assume that the state-action space is bounded, there is a positive constant C_π such that $||\log\pi_{\theta_t}(a|s)||\leq C_\pi$ for any $(s,a)\in\mathcal{S}\times\mathcal{A}$. We start with the following relation:

$$
\begin{aligned}
&\Big|\Big|E_{(S_t^E,A_t^E)\sim\mu^{\pi_E}(\cdot,\cdot)}\Big[L_t(\theta_t;(S_t^E,A_t^E))\Big]-E_{(S_t^E,A_t^E)\sim\mathbb{P}_t^{\pi_E}(\cdot,\cdot)}\Big[L_t(\theta_t;(S_t^E,A_t^E))\Big]\Big|\Big|,\\
&=\gamma^t\Big|\Big|\int_{s\in\mathcal{S}}\int_{a\in\mathcal{A}}|\mathbb{P}_t^{\pi_E}(s,a)-\mu^{\pi_E}(s,a)|\cdot||\log\pi_{\theta_t}(a|s)||dads\Big|\Big|,\\
&\leq C_\pi C_M\gamma^t\rho^t,\\
&\Rightarrow\Big|\Big|E_{(S_t^E,A_t^E)\sim\mu^{\pi_E}(\cdot,\cdot)}\Big[\sum_{t=0}^{T-1}L_t(\theta_t;(S_t^E,A_t^E))\Big]-E_{(S_t^E,A_t^E)\sim\mathbb{P}_t^{\pi_E}(\cdot,\cdot)}\Big[\sum_{t=0}^{T-1}L_t(\theta_t;(S_t^E,A_t^E))\Big]\Big|\Big|,
\end{aligned}
$$

$$\leq \frac{C_\pi C_M}{1-\gamma\rho}. \tag{24}$$

Therefore, we have that

$$
\begin{aligned}
&E_{(S_t^E,A_t^E)\sim\mathbb{P}_t^{\pi_E}}\left[\sum_{t=0}^{T-1} L_t(\theta_t;(S_t^E,A_t^E))\right] - \min_\theta E_{(S_t^E,A_t^E)\sim\mathbb{P}_t^{\pi_E}}\left[\sum_{t=0}^{T-1} L_t(\theta;(S_t^E,A_t^E))\right],\\
&\leq E_{(S_t^E,A_t^E)\sim\mathbb{P}_t^{\pi_E}}\left[\sum_{t=0}^{T-1} L_t(\theta_t;(S_t^E,A_t^E))\right] - E_{(S_t^E,A_t^E)\sim\mu^{\pi_E}}\left[\sum_{t=0}^{T-1} L_t(\theta_t;(S_t^E,A_t^E))\right]\\
&+ E_{(S_t^E,A_t^E)\sim\mu^{\pi_E}}\left[\sum_{t=0}^{T-1} L_t(\theta_t;(S_t^E,A_t^E)) - \sum_{t=0}^{T-1} L_t(\theta^*;(S_t^E,A_t^E))\right]\\
&+ E_{(S_t^E,A_t^E)\sim\mu^{\pi_E}}\left[\sum_{t=0}^{T-1} L_t(\theta^*;(S_t^E,A_t^E))\right] - \min_\theta E_{(S_t^E,A_t^E)\sim\mathbb{P}_t^{\pi_E}}\left[\sum_{t=0}^{T-1} L_t(\theta;(S_t^E,A_t^E))\right],\\
&\leq E_{(S_t^E,A_t^E)\sim\mathbb{P}_t^{\pi_E}}\left[\sum_{t=0}^{T-1} L_t(\theta_t;(S_t^E,A_t^E))\right] - E_{(S_t^E,A_t^E)\sim\mu^{\pi_E}}\left[\sum_{t=0}^{T-1} L_t(\theta_t;(S_t^E,A_t^E))\right]\\
&+ E_{(S_t^E,A_t^E)\sim\mu^{\pi_E}}\left[\sum_{t=0}^{T-1} L_t(\theta_t;(S_t^E,A_t^E)) - \sum_{t=0}^{T-1} L_t(\theta^*;(S_t^E,A_t^E))\right]\\
&+ E_{(S_t^E,A_t^E)\sim\mu^{\pi_E}}\left[\sum_{t=0}^{T-1} L_t(\theta^*;(S_t^E,A_t^E))\right] - E_{(S_t^E,A_t^E)\sim\mathbb{P}_t^{\pi_E}}\left[\sum_{t=0}^{T-1} L_t(\theta^*;(S_t^E,A_t^E))\right],\\
&\overset{(g)}{\leq} \frac{C_\pi C_M}{1-\gamma\rho} + \bar{D}_4 + \frac{32 C_Q C_d \hat{C}\bar{C}_r^2}{\lambda(1-\gamma)^3}(\log T + 1) + \frac{C_\pi C_M}{1-\gamma\rho},\\
&= D_4 + D_5(\log T + 1),
\end{aligned}
$$

where (f) follows (23) and (24), $D_4 = \frac{2C_\pi C_M}{1-\gamma\rho} + \bar{D}_4$, and $D_5 = \frac{32 C_Q C_d \hat{C}\bar{C}_r^2}{\lambda(1-\gamma)^3}$.

C Meta-Regularization Algorithm and Convergence Guarantee

To solve problem (5), we use a double-loop method, i.e., we first solve the lower-level problem for K iterations to get an approximate $\phi_{j,K}$ of ϕ_j and then use the approximate $\phi_{j,K}$ to solve the upper-level problem. We do not use a single-loop method to solve (5) because we need to get the task-specific adaptation ϕ_j for each task $\mathcal{T}_j$. The single-loop method only partially solves the lower-level problem by one-step gradient descent and thus the obtained parameter can be far away from ϕ_j.

Lemma 5. *The gradient of the lower-level problem in* (5) *is* $E_{S,A}^{\pi_\phi}\big[\sum_{t=0}^\infty \gamma^t \nabla_\phi r_\phi(S_t,A_t)\big|S_0 = s_0^{\text{tr}}\big] - \sum_{t=0}^\infty \gamma^t \nabla_\phi r_\phi(s_t^{\text{tr}}, a_t^{\text{tr}}) + \frac{\lambda}{1-\gamma}(\phi - \bar{\theta})$ *where* $(s_t^{\text{tr}}, a_t^{\text{tr}}) \in \mathcal{D}_j^{\text{tr}}$.

We roll out the policy π_ϕ from s_t^{tr} to get a trajectory $s_0^\phi, a_0^\phi, \cdots$ where $s_0^\phi = s_t^{\text{tr}}$ and approximate the lower-level gradient by $g_\phi = \sum_{t=0}^\infty \gamma^t \nabla_\phi r_\phi(s_t^\phi, a_t^\phi) - \sum_{t=0}^\infty \gamma^t \nabla_\phi r_\phi(s_t^{\text{tr}}, a_t^{\text{tr}}) + \frac{\lambda}{1-\gamma}(\phi - \bar{\theta})$. We update $\phi_{j,k+1} = \phi_{j,k} - \beta_k g_{\phi_{j,k}}$ for K times to get $\phi_{j,K}$ where β_k is the step size and $\phi_{j,k}$ is the parameter at time k, and use $\phi_{j,K}$ to calculate the approximate of the hyper-gradient $\frac{d}{d\theta}L(\phi_j, \mathcal{D}_j^{\text{eval}})$.

Lemma 6. *The hyper-gradient (i.e., gradient of the upper-level problem in* (5)*) is* $\frac{d}{d\theta}L(\phi_j, \mathcal{D}_j^{\text{eval}}) = \big[I + \frac{1-\gamma}{\lambda}\nabla^2_{\phi\phi}L(\phi_j, \mathcal{D}_j^{\text{tr}})\big]^{-1}\nabla_\phi L(\phi_j, \mathcal{D}_j^{\text{eval}})$*, where*

$$\nabla_\phi L(\phi_j, \mathcal{D}_j^{\text{eval}}) = |\mathcal{D}_j^{\text{eval}}| \cdot E_{S,A}^{\pi_{\phi_j}}\left[\sum_{t=0}^\infty \gamma^t \nabla_\phi r_{\phi_j}(S_t,A_t)\middle| S_0 \sim P_0\right] - \sum_{v=1}^{|\mathcal{D}_j^{\text{eval}}|}\sum_{t=0}^\infty \gamma^t \nabla_\phi r_{\phi_j}(s_t^v, a_t^v),$$

$$\nabla^2_{\phi\phi}L(\phi_j, \mathcal{D}_j^{\text{tr}}) = E_{S,A}^{\pi_{\phi_j}}\left[\sum_{t=0}^\infty \gamma^t \nabla^2_{\phi\phi} r_{\phi_j}(S_t,A_t)\middle| S_0 = s_0^{\text{tr}}\right] - \sum_{t=0}^\infty \gamma^t \nabla^2_{\phi\phi} r_{\phi_j}(s_t^{\text{tr}}, a_t^{\text{tr}}) + e.$$

The $|\mathcal{D}_j^{\text{eval}}|$ is the number of trajectories in $\mathcal{D}_j^{\text{eval}}$, $(s_t^v, a_t^v) \in \mathcal{D}_j^{\text{eval}}$, and e is an extra term whose expression can be found in Appendix C.2.

To approximate the hyper-gradient, we first roll out $\pi_{\phi_{j,K}}$ to get two trajectories $s_0^{\phi_{j,K}}, a_0^{\phi_{j,K}}, \cdots$ and $\bar{s}_0^{\phi_{j,K}}, \bar{a}_0^{\phi_{j,K}}, \cdots$ where $s_0^{\phi_{j,K}}$ is drawn from P_0 and $\bar{s}_0^{\phi_{j,K}} = s_0^{\text{tr}}$. We estimate $\nabla_\phi L(\phi_j, \mathcal{D}_j^{\text{eval}})$ via $\bar{\nabla}_\phi L(\phi_{j,K}, \mathcal{D}_j^{\text{eval}}) = |\mathcal{D}_j^{\text{eval}}| \cdot \sum_{t=0}^{\infty} \gamma^t \nabla_\phi r_{\phi_{j,K}}(s_t^{\phi_{j,K}}, a_t^{\phi_{j,K}}) - \sum_{v=1}^{|\mathcal{D}_j^{\text{eval}}|} \sum_{t=0}^{\infty} \gamma^t \nabla_\phi r_{\phi_{j,K}}(s_t^v, a_t^v)$ and estimate the term $\nabla^2_{\phi\phi} L(\phi_j, \mathcal{D}_j^{\text{tr}})$ via $\bar{\nabla}^2_{\phi\phi} L(\phi_{j,K}, \mathcal{D}_j^{\text{tr}}) = \sum_{t=0}^{\infty} \gamma^t \nabla^2_{\phi\phi} r_{\phi_{j,K}}(\bar{s}_t^{\phi_{j,K}}, \bar{a}_t^{\phi_{j,K}}) - \sum_{t=0}^{\infty} \gamma^t \nabla^2_{\phi\phi} r_{\phi_{j,K}}(s_t^{\text{tr}}, a_t^{\text{tr}})$. Therefore, we can approximate the hyper-gradient term $\frac{d}{d\bar{\theta}} L(\phi_j, \mathcal{D}_j^{\text{eval}})$ via $h_j = [I + \frac{1-\gamma}{\lambda} \bar{\nabla}^2_{\phi\phi} L(\phi_{j,K}, \mathcal{D}_j^{\text{tr}})]^{-1} \bar{\nabla}_\phi L(\phi_{j,K}, \mathcal{D}_j^{\text{eval}})$. We omit the extra term e in the approximate h_j and its impact on the convergence can be bounded (proved in Appendix C.4).

To solve the upper-level problem in (5), at each iteration n, we sample a batch of B tasks, and compute the task-specific adaptation $\phi_{j,K}$ and the hyper-gradient h_j for each task. The update law to solve the upper-level problem is: $\bar{\theta}_{n+1} = \bar{\theta}_n - \frac{\tau_n}{B} \sum_{j=1}^{B} h_j$ where τ_n is the step size. Note that the time index k is for the lower-level problem and n is for the upper-level problem.

Algorithm 2 Meta-regularization

```
Input: Initialized meta-prior $\bar{\theta}_0$ and task-specific adaptation $\phi_{j,0}$
Output: Learned prior $\bar{\theta}_N$
for $n = 0, 1, \cdots, N-1$ do
    Sample a batch of $B$ tasks $\{\mathcal{T}_j\}_{j=1}^{B} \sim P_{\mathcal{T}}$
    for each task $\mathcal{T}_j$ do
        for $k = 0, 1, \cdots, K-1$ do
            Compute the soft Bellman policy $\pi_{\phi_{j,k}}$ via soft Q-learning or soft actor-critic
            Roll out the policy $\pi_{\phi_{j,k}}$ to get a trajectory $s_0^{\phi_{j,k}}, a_0^{\phi_{j,k}}, \cdots$
            Compute the gradient $g_{\phi_{j,k}} = \sum_{t=0}^{\infty} \gamma^t \nabla_\phi r_{\phi_{j,k}}(s_t^{\phi_{j,k}}, a_t^{\phi_{j,k}}) - \sum_{t=0}^{\infty} \gamma^t \nabla_\phi r_{\phi_{j,k}}(s_t^{\text{tr}}, a_t^{\text{tr}})$
            $+ \frac{\lambda}{1-\gamma}(\phi_{j,k} - \bar{\theta}_n)$
            Update $\phi_{j,k+1} = \phi_{j,k} - \beta_k g_{\phi_{j,k}}$
        end for
        Compute the soft Bellman policy $\pi_{\phi_{j,K}}$ via soft Q-learning or soft actor-critic
        Roll out the policy $\pi_{\phi_{j,K}}$ to get two trajectories $s_0^{\phi_{j,K}}, a_0^{\phi_{j,K}}, \cdots$ and $\bar{s}_0^{\phi_{j,K}}, \bar{a}_0^{\phi_{j,K}}, \cdots$
        Compute the hyper-gradient $h_j = [I + \frac{1-\gamma}{\lambda} \bar{\nabla}^2_{\phi\phi} L(\phi_{j,K}, \mathcal{D}_j^{\text{tr}})]^{-1} \bar{\nabla}_\phi L(\phi_{j,K}, \mathcal{D}_j^{\text{eval}})$
    end for
    Update $\bar{\theta}_{n+1} = \bar{\theta}_n - \frac{\tau_n}{B} \sum_{j=1}^{B} h_j$
end for
```

Lemma 7 (Convergence of the lower-level problem)**.** *Suppose Assumptions 1-2 hold and $\lambda \geq \frac{C_L}{2} + \eta$ where $\eta \in (0, \frac{C_L}{2})$. Let $\beta_k = \frac{1-\gamma}{\eta(k+1)}$, then we have*

$$E[||\phi_{j,K} - \phi_j||^2] \leq O(\frac{1}{K}).$$

Assumption 3. *The parameterized reward function r_θ has bounded third-order gradient, i.e., $||\nabla^3_{\theta\theta\theta} r_\theta(s,a)|| \leq \hat{C}_r$ for any (s,a) where $\hat{C}_r$ is a positive constant.*

Theorem 3 (Convergence of the upper-level problem)**.** *Suppose Assumption 3 and the condition in Lemma 7 hold. Let $\tau_n = (n+1)^{-1/2}$ and define $F(\bar{\theta})$ as $E_{j\sim P_{\mathcal{T}}}[L(\phi_j, \mathcal{D}_j^{eval})]$ under the meta-prior $\bar{\theta}$. Then we have the following convergence:*

$$\frac{1}{N} \sum_{n=0}^{N-1} E[||\nabla F(\bar{\theta}_n)||^2] \leq O(\frac{1}{\sqrt{N}} + \frac{\log N}{\sqrt{N}} + \frac{1}{K}) + C_1,$$

and the expression of C_1 can be found in (27).

C.1 Proof of Lemma 5

The proof is similar to that of Lemma 1. We first prove the case of deterministic dynamics and the corresponding result can be an unbiased estimate in cases of stochastic dynamics (proved in Subsection B.2).

$$\begin{aligned}
\nabla_\phi L(\phi, \mathcal{D}_j^{\text{tr}}) &= -\sum_{t=0}^{\infty} \gamma^t \nabla_\phi \log \pi_\phi(a_t^{\text{tr}}|s_t^{\text{tr}}),\\
&= -\sum_{t=0}^{\infty} \gamma^t \Big[\nabla_\phi Q_\phi^{\text{soft}}(s_t^{\text{tr}}, a_t^{\text{tr}}) - \nabla_\phi V_\phi^{\text{soft}}(s_t^{\text{tr}})\Big],\\
&= -\sum_{t=0}^{\infty} \gamma^t \Big[\nabla_\phi r_\phi(s_t^{\text{tr}}, a_t^{\text{tr}}) + \gamma \nabla_\phi V_\phi^{\text{soft}}(s_{t+1}^{\text{tr}}) - \nabla_\phi V_\phi^{\text{soft}}(s_t^{\text{tr}})\Big],\\
&= \nabla_\phi V_\phi^{\text{soft}}(s_0^{\text{tr}}) - \sum_{t=0}^{\infty} \gamma^t \nabla_\phi r_\phi(s_t^{\text{tr}}, a_t^{\text{tr}}),\\
&\overset{(a)}{=} E_{S,A}^{\pi_\phi}\Big[\sum_{t=0}^{\infty} \gamma^t \nabla_\phi r_\phi(S_t, A_t)\Big| S_0 = s_0^{\text{tr}}\Big] - \sum_{t=0}^{\infty} \gamma^t \nabla_\phi r_\phi(s_t^{\text{tr}}, a_t^{\text{tr}}),
\end{aligned}$$

where (a) follows the proof of Lemma 3.

C.2 Proof of Lemma 6

Since the lower-level problem of (5) is unconstrained and ϕ_i is the optimal solution, we know that

$$\nabla_\phi L(\phi_j, \mathcal{D}_j^{\text{tr}}) + \frac{\lambda}{1-\gamma}(\phi_j - \bar{\theta}) = 0 \tag{25}$$

We further take derivative of both sides in (25) with respect to $\bar{\theta}$ and we get that

$$\Big[\nabla_{\phi\phi}^2 L(\phi_j, \mathcal{D}_j^{\text{tr}}) + \frac{\lambda}{1-\gamma}\Big]\nabla_{\bar{\theta}}\phi_j - \frac{\lambda}{1-\gamma} = 0 \Rightarrow \nabla_{\bar{\theta}}\phi_j = \Big[I + \frac{1-\gamma}{\lambda}\nabla_{\phi\phi}^2 L(\phi_j, \mathcal{D}_j^{\text{tr}})\Big]^{-1}.$$

Therefore, we have that

$$\frac{d}{d\bar{\theta}} L(\phi_j, \mathcal{D}_j^{\text{eval}}) = (\nabla_{\bar{\theta}}\phi_j)^\top \nabla_\phi L(\phi_j, \mathcal{D}_j^{\text{eval}}) = \Big[I + \frac{1-\gamma}{\lambda}\nabla_{\phi\phi}^2 L(\phi_j, \mathcal{D}_j^{\text{tr}})\Big]^{-1} \nabla_\phi L(\phi_j, \mathcal{D}_j^{\text{eval}}).$$

Similar to Lemma 5, we can know that

$$\nabla_\phi L(\phi_j, \mathcal{D}_j^{\text{eval}}) = \sum_{v=1}^{|\mathcal{D}_j^{\text{eval}}|} E_{S,A}^{\pi_{\phi_j}}\Big[\sum_{t=0}^{\infty} \gamma^t \nabla_\phi r_{\phi_j}(S_t, A_t)\Big| S_0 = s_0^v\Big] - \sum_{v=1}^{|\mathcal{D}_j^{\text{eval}}|}\sum_{t=0}^{\infty} \gamma^t \nabla_\phi r_{\phi_j}(s_t^v, a_t^v), \tag{26}$$

where $(s_t^v, a_t^v) \in \zeta^v$ and $\zeta^v \in \mathcal{D}_j^{\text{eval}}$.

Since there is usually abundant data in $\mathcal{D}_j^{\text{eval}}$ and $S_0 \sim P_0$, we can reformulate (26) as the following:

$$\nabla_\phi L(\phi_j, \mathcal{D}_j^{\text{eval}}) \approx |\mathcal{D}_j^{\text{eval}}| \cdot E_{S,A}^{\pi_{\phi_j}}\Big[\sum_{t=0}^{\infty} \gamma^t \nabla_\phi r_{\phi_j}(S_t, A_t)\Big| S_0 \sim P_0\Big] - \sum_{v=1}^{|\mathcal{D}_j^{\text{eval}}|}\sum_{t=0}^{\infty} \gamma^t \nabla_\phi r_{\phi_j}(s_t^v, a_t^v).$$

Claim 7. *The second-order information $\nabla_{\phi\phi}^2 Q_\phi^{\text{soft}}(s,a) = \Delta(s,a) + E_{S,A}^{\pi_\phi}[\sum_{t=0}^{\infty}\gamma^t Cov(S_t)|S_0 = s, A_0 = a]$ and $\nabla_{\phi\phi}^2 V_\phi^{\text{soft}}(s) = \Delta(s) + E_{S,A}^{\pi_\phi}[\sum_{t=0}^{\infty}\gamma^t Cov(S_t)|S_0 = s]$ where $\Delta(s,a) = E_{S,A}^{\pi_\phi}[\sum_{t=0}^{\infty}\gamma^t \nabla_{\phi\phi}^2 r_\phi(S_t, A_t)|S_0 = s, A_0 = a]$, $\Delta(s) = E_{S,A}^{\pi_\phi}[\sum_{t=0}^{\infty}\gamma^t \nabla_{\phi\phi}^2 r_\phi(S_t, A_t)|S_0 = s]$, and $Cov(s) \triangleq \int_{a\in\mathcal{A}} \pi_\phi(a|s)[\Delta(s,a) - \Delta(s)]\Delta(s)da$ is the covariance matrix of $\Delta(s,\cdot)$ at state s.*

The proof of Claim 7 follows the proof of Lemma 4

Now we take a look at the term $\nabla^2_{\phi\phi} L(\phi_j, \mathcal{D}^{\text{tr}}_j)$ and consider the case of deterministic dynamics:

$$\begin{aligned}
&\nabla^2_{\phi\phi} L(\phi_j, \mathcal{D}^{\text{tr}}_j) = -\sum_{t=0}^{\infty} \gamma^t \nabla^2_{\phi\phi} \log \pi_{\phi_j}(a^{\text{tr}}_t | s^{\text{tr}}_t), \\
&= -\sum_{t=0}^{\infty} \gamma^t \Big[\nabla^2_{\phi\phi} Q^{\text{soft}}_{\phi_j}(s^{\text{tr}}_t, a^{\text{tr}}_t) - \nabla^2_{\phi\phi} V^{\text{soft}}_{\phi_j}(s^{\text{tr}}_t) \Big], \\
&= -\sum_{t=0}^{\infty} \gamma^t \Big[\nabla^2_{\phi\phi} r_{\phi_j}(s^{\text{tr}}_t, a^{\text{tr}}_t) + \gamma \nabla^2_{\phi\phi} V^{\text{soft}}_{\phi_j}(s^{\text{tr}}_{t+1}) - \nabla^2_{\phi\phi} V^{\text{soft}}_{\phi_j}(s^{\text{tr}}_t) \Big], \\
&= \nabla^2_{\phi\phi} V^{\text{soft}}_{\phi_j}(s^{\text{tr}}_0) - \sum_{t=0}^{\infty} \gamma^t \nabla^2_{\phi\phi} r_{\phi_j}(s^{\text{tr}}_t, a^{\text{tr}}_t), \\
&\overset{(a)}{=} E^{\pi_{\phi_j}}_{S,A}[\sum_{t=0}^{\infty} \gamma^t \nabla^2_{\phi\phi} r_{\phi_j}(S_t, A_t) | S_0 = s^{\text{tr}}_0] - \sum_{t=0}^{\infty} \gamma^t \nabla^2_{\phi\phi} r_{\phi_j}(s^{\text{tr}}_t, a^{\text{tr}}_t) \\
&+ E^{\pi_{\phi_j}}_{S,A}[\sum_{t=0}^{\infty} \gamma^t \text{Cov}(S_t) | S_0 = s^{\text{tr}}_0],
\end{aligned}$$

where (a) follows Claim 7. As proved in Subsection B.2, we can still use this expression in the case of stochastic dynamics. We define $e \triangleq E^{\pi_{\phi_j}}_{S,A}[\sum_{t=0}^{\infty} \gamma^t \text{Cov}(S_t) | S_0 = s^{\text{tr}}_0]$. However, e is intractable to compute because it requires to compute COV(S_t) at every visited state. To approximate the covariance matrix COV(S_t), we need to empirically roll out the policy π_{ϕ_j} from S_t for enough times to get enough samples. We need to do these policy roll-outs at every state S_t. For example, suppose we roll out the policy π_{ϕ_j} ten times to get ten samples at each state S_t. Empirically, we need to do these roll-outs $10 \times \bar{T}$ where $\bar{T}$ is a very large integer that we regard as infinity because the trajectory horizon is infinite. This is intractable because we need to roll out the policy for too many times. Moreover, we cannot guarantee that we can approximate COV(S_t) well given that we only use ten samples to approximate it.

C.3 Proof of Lemma 7

We know that $||\nabla^2_{\phi\phi} L(\phi, \mathcal{D}^{\text{tr}}_j) + \frac{\lambda}{1-\gamma}|| \leq ||\nabla^2_{\phi\phi} L(\phi, \mathcal{D}^{\text{tr}}_j)|| + \frac{\lambda}{1-\gamma} \leq \sum_{t=0}^{\infty} \Big(||\nabla^2 L_t(\phi)|| + \lambda \gamma^t \Big) \overset{(a)}{\leq} \frac{1}{1-\gamma} C_L$ where (a) follows (6). Therefore, the lower-level objective function in (8) is $\frac{C_L}{1-\gamma}$-smooth. Moreover, since $\lambda \geq \frac{C_L}{2} + \eta$, then $||\nabla^2_{\phi\phi} L(\phi, \mathcal{D}^{\text{tr}}_j)|| \leq \frac{C_L - 2\eta}{2(1-\gamma)}$. Therefore, the lower-level objective function in (8) is $\frac{2\eta}{1-\gamma}$-strongly convex. Following the standard result for strongly-convex and smooth stochastic optimization, we can reach the result in Lemma 7.

C.4 Proof of Theorem 3

In this proof, we first bound the hyper-gradient approximation error (i.e., $||\frac{d}{d\bar{\theta}} L(\phi_j, \mathcal{D}^{\text{eval}}_j) - h_j||$) and then prove the convergence. Define $\bar{h}_j \triangleq [I + \frac{1-\gamma}{\lambda} \bar{\nabla}^2_{\phi\phi} L(\phi_j, \mathcal{D}^{\text{tr}}_j)]^{-1} \nabla_\phi L(\phi_j, \mathcal{D}^{\text{eval}}_j)$ where $\bar{\nabla}^2_{\phi\phi} L(\phi_j, \mathcal{D}^{\text{tr}}_j) \triangleq E^{\pi_{\phi_j}}_{S,A}[\sum_{t=0}^{\infty} \gamma^t \nabla^2_{\phi\phi} r_{\phi_j}(S_t, A_t) | S_0 = s^{\text{tr}}_0] - \sum_{t=0}^{\infty} \gamma^t \nabla^2_{\phi\phi} r_{\phi_j}(s^{\text{tr}}_t, a^{\text{tr}}_t)$. Therefore, we have that

$$\begin{aligned}
&||\bar{h}_j - \frac{d}{d\bar{\theta}} L(\phi_j, \mathcal{D}^{\text{eval}}_j)||, \\
&\leq \Bigg|\Bigg| [I + \frac{1-\gamma}{\lambda} \nabla^2_{\phi\phi} L(\phi_j, \mathcal{D}^{\text{tr}}_j)]^{-1} - [I + \frac{1-\gamma}{\lambda} \bar{\nabla}^2_{\phi\phi} L(\phi_j, \mathcal{D}^{\text{tr}}_j)]^{-1} \Bigg|\Bigg| \cdot ||\nabla_\phi L(\phi_j, \mathcal{D}^{\text{eval}}_j)||, \\
&\leq \Bigg[\Bigg|\Bigg| [I + \frac{1-\gamma}{\lambda} \nabla^2_{\phi\phi} L(\phi_j, \mathcal{D}^{\text{tr}}_j)]^{-1} \Bigg|\Bigg| + \Bigg|\Bigg| [I + \frac{1-\gamma}{\lambda} \bar{\nabla}^2_{\phi\phi} L(\phi_j, \mathcal{D}^{\text{tr}}_j)]^{-1} \Bigg|\Bigg| \Bigg] \cdot ||\nabla_\phi L(\phi_j, \mathcal{D}^{\text{eval}}_j)||, \\
&\overset{(a)}{\leq} \left(\frac{2\lambda}{2\lambda + 2\eta - C_L} + \frac{\lambda}{\lambda - 2\tilde{C}_r} \right) \cdot \frac{2|\mathcal{D}^{\text{eval}}_j| \bar{C}_r}{1-\gamma} \triangleq C_1 > 0, \qquad (27)
\end{aligned}$$

where (a) follows the fact that $||I + \frac{1-\gamma}{\lambda}\nabla^2_{\phi\phi}L(\phi_j, \mathcal{D}^{\text{tr}}_j)|| \geq 1 - ||\frac{1-\gamma}{\lambda}\nabla^2_{\phi\phi}L(\phi_j, \mathcal{D}^{\text{tr}}_j)|| \geq 1 - \frac{C_L - 2\eta}{2\lambda}$ given that $||\nabla^2_{\phi\phi}L(\phi, \mathcal{D}^{\text{tr}}_j)|| \leq \frac{C_L-2\eta}{2(1-\gamma)}$ (proved in the proof of Lemma 7). Therefore $||[I + \frac{1-\gamma}{\lambda}\nabla^2_{\phi\phi}L(\phi_j, \mathcal{D}^{\text{tr}}_j)]^{-1}|| \leq \frac{2\lambda}{2\lambda+2\eta-C_L}$. Similarly, we can bound $||[I + \frac{1-\gamma}{\lambda}\bar{\nabla}^2_{\phi\phi}L(\phi_j, \mathcal{D}^{\text{tr}}_j)]^{-1}|| \leq \frac{\lambda}{\lambda - 2\tilde{C}_r}$ given that $||\bar{\nabla}^2_{\phi\phi}L(\phi_j, \mathcal{D}^{\text{tr}}_j)|| \leq \frac{2\tilde{C}_r}{1-\gamma}$. Note that $\lambda > \frac{C_L}{2}$ and $C_L > \tilde{C}_r$ (proved in (6)), therefore C_1 is a positive constant.

Now, we bound the term $||h_j - \bar{h}_j||$. We define $\Delta_{\phi_j} = I + \frac{1-\gamma}{\lambda}\bar{\nabla}^2_{\phi\phi}L(\phi_j, \mathcal{D}^{\text{tr}}_j)$ and $\Delta_{\phi_j} = I + \frac{1-\gamma}{\lambda}\bar{\nabla}^2_{\phi\phi}L(\phi_{j,K}, \mathcal{D}^{\text{tr}}_j)$. Thus we have $||\Delta^{-1}_{\phi_j}|| \leq \frac{\lambda}{\lambda-2\tilde{C}_r}$ (follows (27)) and similarly $||\Delta^{-1}_{\phi_{j,K}}|| \leq \frac{\lambda}{\lambda-2\tilde{C}_r}$. Therefore,

$$
\begin{aligned}
&E[||h_j - \bar{h}_j||] = E\bigg[\bigg|\bigg|\Delta^{-1}_{\phi_{j,K}}\nabla_\phi L(\phi_{j,K}, \mathcal{D}^{\text{eval}}_j) - \Delta^{-1}_{\phi_j}\nabla_\phi L(\phi_j, \mathcal{D}^{\text{eval}}_j)\bigg|\bigg|\bigg], \\
&\leq E\bigg[\bigg|\bigg|\Delta^{-1}_{\phi_{j,K}}\nabla_\phi L(\phi_{j,K}, \mathcal{D}^{\text{eval}}_j) - \Delta^{-1}_{\phi_{j,K}}\nabla_\phi L(\phi_j, \mathcal{D}^{\text{eval}}_j)\bigg|\bigg| \\
&+ \bigg|\bigg|\Delta^{-1}_{\phi_{j,K}}\nabla_\phi L(\phi_j, \mathcal{D}^{\text{eval}}_j) - \Delta^{-1}_{\phi_j}\nabla_\phi L(\phi_j, \mathcal{D}^{\text{eval}}_j)\bigg|\bigg|\bigg], \\
&\leq E\bigg[||\Delta^{-1}_{\phi_{j,K}}|| \cdot ||\nabla_\phi L(\phi_{j,K}, \mathcal{D}^{\text{eval}}_j) - \nabla_\phi L(\phi_j, \mathcal{D}^{\text{eval}}_j)|| + ||\Delta^{-1}_{\phi_{j,K}} - \Delta^{-1}_{\phi_j}|| \cdot ||\nabla_\phi L(\phi_j, \mathcal{D}^{\text{eval}}_j)||\bigg], \\
&\overset{(b)}{\leq} \frac{\lambda}{\lambda - 2\tilde{C}_r} \cdot |\mathcal{D}^{\text{eval}}_j| \cdot \frac{C_L - 2\eta}{2(1-\gamma)}||\phi_{j,K} - \phi_j|| + E[||\Delta^{-1}_{\phi_{j,K}} - \Delta^{-1}_{\phi_j}||] \cdot |\mathcal{D}^{\text{eval}}_j| \cdot \frac{2\bar{C}_r}{1-\gamma}, \\
&\leq O(\frac{1}{K}) + E[||\Delta^{-1}_{\phi_{j,K}}|| \cdot ||\Delta^{-1}_{\phi_j}|| \cdot ||\Delta^{-1}_{\phi_{j,K}} - \Delta^{-1}_{\phi_j}||] \cdot |\mathcal{D}^{\text{eval}}_j| \cdot \frac{2\bar{C}_r}{1-\gamma}, \\
&\leq O(\frac{1}{K}) + \bar{C}_L E[||\Delta^{-1}_{\phi_{j,K}}|| \cdot ||\Delta^{-1}_{\phi_j}||] \cdot |\mathcal{D}^{\text{eval}}_j| \cdot \frac{2\bar{C}_r}{1-\gamma}||\phi_{j,K} - \phi_j|| \leq O(\frac{1}{K}),
\end{aligned} \tag{28}
$$

where $||\nabla^3_{\phi\phi\phi}L(\phi_j, \mathcal{D}^{\text{tr}}_j)|| \leq \tilde{C}_L$. The expression of $\tilde{C}_L$ can be derived following the proof of Lemma 4 by notifying Assumption 3. The (b) holds because (i) $||\nabla^2_{\phi\phi}L(\phi, \mathcal{D}^{\text{eval}}_j)|| = \frac{|\mathcal{D}^{\text{eval}}_j|}{|\mathcal{D}^{\text{tr}}_j|}||\nabla^2_{\phi\phi}L(\phi, \mathcal{D}^{\text{tr}}_j)|| \leq |\mathcal{D}^{\text{eval}}_j| \cdot \frac{C_l - 2\eta}{2(1-\gamma)}$ and (ii) $||\nabla_\phi L(\phi_j, \mathcal{D}^{\text{eval}}_j)|| \leq |\mathcal{D}^{\text{eval}}_j| \cdot \frac{2\bar{C}_r}{1-\gamma}$ (see the expression of $\nabla_\phi L(\phi_j, \mathcal{D}^{\text{eval}}_j)$ in Lemma 5).

$$
E[||h_j - \frac{d}{d\bar{\theta}}L(\phi_j, \mathcal{D}^{\text{eval}}_j)||] \leq E[||h_j - \bar{h}_j|| + ||\bar{h}_j - \frac{d}{d\bar{\theta}}L(\phi_j, \mathcal{D}^{\text{eval}}_j)||] \overset{(c)}{\leq} C_1 + O(\frac{1}{K}), \tag{29}
$$

where (c) follows (27)-(28). Define $F(\bar{\theta})$ as $E_{j\sim P_{\mathcal{T}}}[L(\phi_j, \mathcal{D}^{\text{eval}}_j)]$ under the meta-prior $\bar{\theta}$. Note that $||h_j|| \leq ||\Delta^{-1}_{\phi_{j,K}}|| \cdot ||\nabla_\phi L(\phi_{j,K}, \mathcal{D}^{\text{eval}}_j)|| \leq \frac{\lambda}{\lambda-2\tilde{C}_r} \cdot |\mathcal{D}^{\text{eval}}_j| \cdot \frac{2\bar{C}_r}{1-\gamma}$. Therefore,

$$
\begin{aligned}
&F(\bar{\theta}_{n+1}) \geq F(\bar{\theta}_n) + (\nabla F(\bar{\theta}_n))^\top(\bar{\theta}_{n+1} - \bar{\theta}_n) - \frac{|\mathcal{D}^{\text{eval}}|(C_L - 2\eta)}{4(1-\gamma)}||\bar{\theta}_{n+1} - \bar{\theta}_n||^2, \\
&\geq F(\bar{\theta}_n) + \frac{\tau_n}{B}\sum_{j=1}^{B}(\nabla F(\bar{\theta}_n))^\top h_j - \frac{|\mathcal{D}^{\text{eval}}|(C_L - 2\eta)\tau_n^2}{4(1-\gamma)}||\frac{1}{B}\sum_{j=1}^{B}h_j||^2, \\
&\geq F(\bar{\theta}_n) + \frac{\tau_n}{B}\sum_{j=1}^{B}(\nabla F(\bar{\theta}_n))^\top h_j - \frac{|\mathcal{D}^{\text{eval}}|(C_L - 2\eta)\tau_n^2}{4(1-\gamma)} \cdot \frac{\lambda}{\lambda - 2\tilde{C}_r} \cdot |\mathcal{D}^{\text{eval}}_j| \cdot \frac{2\bar{C}_r}{1-\gamma}, \\
&\geq F(\bar{\theta}_n) + \frac{\tau_n}{B}\sum_{j=1}^{B}(\nabla F(\bar{\theta}_n))^\top h_j - \frac{2\lambda\bar{C}_r|\mathcal{D}^{\text{eval}}|^2(C_L - 2\eta)\tau_n^2}{4(1-\gamma)^2(\lambda - 2\tilde{C}_r)}, \\
&\Rightarrow E[F(\bar{\theta}_{n+1})] \geq E[F(\bar{\theta}_n)] + \tau_n E[||\nabla F(\bar{\theta}_n)||^2] + \frac{\tau_n}{B}\sum_{j=1}^{B}E[h_j - \frac{d}{d\bar{\theta}}L(\phi_j, \mathcal{D}^{\text{eval}}_j)] \\
&+ \frac{\tau_n}{B}\sum_{j=1}^{B}E[\frac{d}{d\bar{\theta}}L(\phi_j, \mathcal{D}^{\text{eval}}_j) - \nabla F(\bar{\theta}_n)] - \frac{2\lambda\bar{C}_r|\mathcal{D}^{\text{eval}}|^2(C_L - 2\eta)\tau_n^2}{4(1-\gamma)^2(\lambda - 2\tilde{C}_r)},
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(d)}{\geq} E[F(\bar{\theta}_n)] + \tau_n E[||\nabla F(\bar{\theta}_n)||^2] + \tau_n(C_1 + O(\frac{1}{K})) - \frac{2\lambda\bar{C}_r|\mathcal{D}^{\text{eval}}|^2(C_L - 2\eta)\tau_n^2}{4(1-\gamma)^2(\lambda - 2\tilde{C}_r)},\\
&\Rightarrow \frac{1}{N}\sum_{n=0}^{N-1}\tau_N E[||\nabla F(\bar{\theta}_n)||^2] \leq \frac{1}{N}\sum_{n=0}^{N-1}\tau_n E[||\nabla F(\bar{\theta}_n)||^2],\\
&\leq \frac{1}{N}[F(\bar{\theta}_N) - F(\bar{\theta}_0)] + \frac{1}{N}\sum_{n=0}^{N-1}\tau_n(C_1 + O(\frac{1}{K})) + \frac{1}{N}\sum_{n=0}^{N-1}\frac{2\lambda\bar{C}_r|\mathcal{D}^{\text{eval}}|^2(C_L - 2\eta)\tau_n^2}{4(1-\gamma)^2(\lambda - 2\tilde{C}_r)},\\
&\Rightarrow \frac{1}{N}\sum_{n=0}^{N-1}E[||\nabla F(\bar{\theta}_n)||^2] \leq O(\frac{1}{\sqrt{N}} + \frac{\log N}{\sqrt{N}} + \frac{1}{K}) + C_1,
\end{aligned}
$$

where $|\mathcal{D}^{\text{eval}}| \triangleq \sup_{i\sim P_{\mathcal{T}}}\{|\mathcal{D}_j^{\text{eval}}|\}$ and (d) follows (29).

D Experiment details

The code was running on a laptop whose processor is AMD Ryzen 7 4700U with Radeon Graphics, 2.00GHz, and the installed RAM is 20.0GB. The operating system is Ubuntu 18.04. We use a neural network to parameterize the learned reward function. The neural network has two hidden layers where each hidden layer has 64 neurons. The activation functions are respectively ReLU and Tanh.

D.1 Baselines

Here we provide the update rule for each baseline. Given a learned reward function r_θ, the policy updates of the four baselines are the same with that of MERIT-IRL, i.e., one-step policy iteration. The difference is the reward update.

IT-IRL: IT-IRL is MERIT-IRL without the meta-regularization term. Therefore, the reward update of IT-IRL is $\theta_{t+1} = \theta_t - \alpha_t g'_t$ where $g'_t = \sum_{i=0}^{\infty}\gamma^i\nabla_\theta r_{\theta_t}(s'_i, a'_i) - \sum_{i=0}^{\infty}\gamma^i\nabla_\theta r_{\theta_t}(s''_i, a''_i)$. Recall from Subsection 4.1 that $\{(s'_i, a'_i)\}_{i\geq 0}$ is generated by the learned policy π_{θ_t} starting from the initial state $s'_0 = s^E_0$, and $\{(s''_i, a''_i)\}_{i\geq 0}$ is generated by the learned policy π_{θ_t} starting from (s''_t, a''_t) where $(s''_i, a''_i) = (s^E_i, a^E_i)$ for $0 \leq i \leq t$.

Naive MERIT-IRL: This method has the meta-regularization term, however, it uses the naive way (depicted in the middle of Figure 1) to update the reward function. In specific, it only compares the partial expert trajectory $\{s^E_i, a^E_i\}_{i=0}^t$ and partial learner trajectory $\{s'_i, a'_i\}_{i=0}^t$. Therefore, the reward update of Naive MERIT-IRL is $\theta_{t+1} = \theta_t - \alpha_t g''_t$ where $g''_t = \sum_{i=0}^{t}\gamma^i\nabla_\theta r_{\theta_t}(s'_i, a'_i) - \sum_{i=0}^{t}\gamma^i\nabla_\theta r_{\theta_t}(s^E_i, a^E_i) + \frac{\lambda(1-\gamma^{t+1})}{1-\gamma}(\theta - \bar{\theta})$.

Naive IT-IRL: This method does not have the meta-regularization term and uses the naive way to update the reward function. Therefore, the reward update of Naive IT-IRL is $\theta_{t+1} = \theta_t - \alpha_t g'''_t$ where $g'''_t = \sum_{i=0}^{t}\gamma^i\nabla_\theta r_{\theta_t}(s'_i, a'_i) - \sum_{i=0}^{t}\gamma^i\nabla_\theta r_{\theta_t}(s^E_i, a^E_i)$.

Hindsight: This method is a standard IRL method with the meta-regularization term where the complete expert trajectory $\{s^E_i, a^E_i\}_{i\geq 0}$ and the complete learner trajectory $\{s'_i, a'_i\}_{i\geq 0}$ are compared to update the reward function. Therefore, the reward update of Hindsight is $\theta_{t+1} = \theta_t - \alpha_t g''''_t$ where $g''''_t = \sum_{i=0}^{\infty}\gamma^i\nabla_\theta r_{\theta_t}(s'_i, a'_i) - \sum_{i=0}^{\infty}\gamma^i\nabla_\theta r_{\theta_t}(s^E_i, a^E_i) + \frac{\lambda(1-\gamma^{t+1})}{1-\gamma}(\theta - \bar{\theta})$.

D.2 MuJoCo

D.2.1 Walker

Figure 2b shows that MERIT can achieve similar performance with the expert after $t = 600$ while the other three in-trajectory learning baselines fail to imitate the expert before the ongoing trajectory terminates. Note that the naive methods (i.e., Naive MERIT-IRL and Naive IT-IRL) have much smaller improvement from $t = 0$ compared to MERIT-IRL and IT-IRL. The reason is that the naive reward update method is flawed. Intuitively, the reward update mechanism of these two baselines are myopic as explained in Subsection 4.1. Theoretically, the gradients g''_t of Naive MERIT-IRL and g'''_t

of Naive IT-IRL are biased estimate of (4) even if π_t approaches π_{θ_t} since (4) includes the trajectory suffix ($i > t$) terms while g_t'' and g_t''' only include the trajectory prefix ($i \leq t$) terms.

MERIT-IRL performs much better than IT-IRL. The reason is that the meta-regularization term restricts the learned reward parameter within a certain neighborhood of the meta-prior $\bar{\theta}$ (proved in Appendix B.3). Given that $\bar{\theta}$ is trained over a family of relevant tasks, it is expected that the actual reward function parameter of our task shall be “close" to $\bar{\theta}$ [24, 43, 56], i.e., inside this neighborhood. Therefore, MERIT-IRL can efficiently learn the expert’s reward function. On the contrary, IT-IRL does not have the meta-prior $\bar{\theta}$ as a guidance and thus has to search over the whole parameter space, which is extremely difficult to learn the expert’s reward function when the data is lacking. Note that MERIT-IRL and Naive MERIT-IRL have better initial performance than IT-IRL and Naive IT-IRL since MERIT-IRL and Naive MERIT-IRL starts at the meta-prior $\bar{\theta}$ while IT-IRL and Naive IT-IRL initializes randomly.

D.2.2 Hopper

Figure 2c shows that MERIT can achieve similar performance with the expert after $t = 500$ while the other three in-trajectory learning baselines fail to imitate the expert before the ongoing trajectory terminates.

D.3 Stock Market

We use the real-world data of 30 constitute stocks in Dow Jones Industrial Average from 2021-01-01 to 2022-01-01. The 30 stocks are respectively: ‘AXP’, ‘AMGN’, ‘AAPL’, ‘BA’, ‘CAT’, ‘CSCO’, ‘CVX’, ‘GS’, ‘HD’, ‘HON’, ‘IBM’, ‘INTC’, ‘JNJ’, ‘KO’, ‘JPM’, ‘MCD’, ‘MMM’, ‘MRK’, ‘MSFT’, ‘NKE’, ‘PG’, ‘TRV’, ‘UNH’, ‘CRM’, ‘VZ’, ‘V’, ‘WBA’, ‘WMT’, ‘DIS’, ‘DOW’.

The state of the stock market MDP is the perception of the stock market, including the open/close price of each stock, the current asset, and some technical indices [47]. The action has the same dimension as the number of stocks where each dimension represents the amount of buying/selling the corresponding stock. The detailed formulation of the MDP can be found in FinRL [47, 57].

The turbulence index is a technical index of stock market and is included as a dimension of the state [47, 57]. The function p_2 is defined as the amount of buying the stocks whose turbulence index is larger than the turbulence threshold. Therefore, the more the target investor buys the stocks whose turbulence index is larger than the turbulence threshold, the larger p_2 will be and thus the smaller reward the target investor will receive.

Discussion on the experiment results. In Figure 2d, MERIT-IRL can achieve the similar cumulative reward with the expert when only the first 60% of the trajectory is observed while IT-IRL can achieve performance close to the expert after $t = 220$. This shows that the meta-regularization can help imitate the expert faster. In contrast, Naive MERIT-IRL and Naive IT-IRL barely improves because the naive reward update method is flawed. Intuitively, the reward update mechanism of these two baselines are myopic as explained in Subsection 4.1. Theoretically, the gradients g_t'' of Naive MERIT-IRL and g_t''' of Naive IT-IRL are biased estimate of (4) even if π_t approaches π_{θ_t} since (4) includes the trajectory suffix ($i > t$) terms while g_t'' and g_t''' only include the trajectory prefix ($i \leq t$) terms.

The last row in Table 1 shows the final results of the algorithms. We can see that MERIT-IRL achieves much better performance than the other in-trajectory learning baselines (i.e., IT-IRL, Naive MERIT-IRL, and Naive IT-IRL). MERIT-IRL achieves comparable performance with Hindsight and the expert. Note that it is not expected that MERIT-IRL outperforms Hindsight since Hindsight has the complete expert trajectory to learn.

E Potential negative societal impact

Since MERIT-IRL can infer the reward function of the expert, potential negative societal impact may occur when the learner is malicious. Take the stock market experiment as an example, private information like preferences or habits of the investors may be leaked by using MERIT-IRL. To avoid this situation, the investors needs to take additional strategies such as protecting its investment data from unsecure resources.

F Limitations

From the objective (2), we can see that the goal of MERIT-IRL is to align with the expert demonstration, i.e., finding a reward function such that its corresponding policy makes the expert trajectory most likely. An ideal case is that we can also directly quantify the reward learning performance and study the reward identifiability issue. Thus, a future work is to study the reward identifiability issue in the context of in-trajectory IRL.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract summarizes our contributions, and the introduction has a "contribution statement" part which elaborates our contributions and mentions the consistency with theoretical and experiment results. The assumptions and limitations are included in "theoretical analysis" (Subsection 4.2) and Appendix F.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitation is mentioned in Appendix F. Under the assumptions, i.e., Assumptions 1 and 2, we have a justification of either the assumption is widely used in literature or the assumption can be satisfied by real scenarios.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We clearly state our assumptions in Assumptions 1 and 2. The theoretical statements are also clearly stated in Subsection 4.2 and the correct proof is included in Appendix B. All the assumptions and theoretical statements are numbered and cross-referenced.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: This paper provides pseudocode and elaborates each step of the algorithm in Subsection 4.1 and Appendix 4.3. We also include the experiment details in Appendix D.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code is submitted in the supplementary materials and we include a document in the code folder to describe how to run the code.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendix D includes the experiment details, and we also submit the code.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: In the experiment results (i.e., Table 1, Table **??**, Table **??**), we include both the mean and standard deviation. In Figure 2, we also plot the mean and standard deviation.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We include the type of operating system, CPU, and RAM used in the first paragraph in Appendix D.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This paper follows NeurIPS Code of Ethnics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer:[Yes]

    Justification: We discuss the potential negative societal impact in Appendix E.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper poses no such risks because we do not have data nor model to release.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use the SAC code from a Github repository, and we clearly state it at the top of that code script.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The submitted code can be considered as an asset, and we provide a file along with the code to document the code.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.