ON COMPUTATION AND GENERALIZATION OF GENERATIVE ADVERSARIAL IMITATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative Adversarial Imitation Learning (GAIL) is a powerful and practical approach for learning sequential decision-making policies. Different from Reinforcement Learning (RL), GAIL takes advantage of demonstration data by experts (e.g., human), and learns both the policy and reward function of the unknown environment. Despite the significant empirical progresses, the theory behind GAIL is still largely unknown. The major difficulty comes from the underlying temporal dependency of the demonstration data and the minimax computational formulation of GAIL without convex-concave structure. To bridge such a gap between theory and practice, this paper investigates the theoretical properties of GAIL. Specifically, we show: (1) For GAIL with general reward parameterization, the generalization can be guaranteed as long as the class of the reward functions is properly controlled; (2) For GAIL, where the reward is parameterized as a reproducing kernel function, GAIL can be efficiently solved by stochastic first order optimization algorithms, which attain sublinear convergence to a stationary solution. To the best of our knowledge, these are the first results on statistical and computational guarantees of imitation learning with reward/policy function approximation. Numerical experiments are provided to support our analysis.

1 INTRODUCTION

As various robots (Tail et al., 2018), self-driving cars (Kuefler et al., 2017), unmanned aerial vehicles (Pfeiffer et al., 2018) and other intelligent agents are applied to complex and unstructured environments, programming their behaviors/policy has become increasingly challenging. These intelligent agents need to accommodate a huge number of tasks with unique environmental demands. To address these challenges, many reinforcement learning (RL) methods have been proposed for learning sequential decision-making policies (Sutton et al., 1998; Kaelbling et al., 1996; Mnih et al., 2015). These RL methods, however, heavily rely on human expert domain knowledge to design proper reward functions. For complex tasks, which are often difficult to describe formally, these RL methods become impractical.

The Imitation Learning (IL, Argall et al. (2009); Abbeel & Ng (2004)) approach is a powerful and practical alternative to RL. Rather than having a human expert handcrafting a reward function for learning the desired policy, the imitation learning approach only requires the human expert to demonstrate the desired policy, and then the intelligent agent (a.k.a. learner) learns to match the demonstration. Most of existing imitation learning methods fall in the following two categories:

• Behavioral Cloning (BC, Pomerleau (1991)). BC treats the IL problem as supervised learning. Specifically, it learns a policy by fitting a regression model over expert demonstrations, which directly maps states to actions. Unfortunately, BC has a fundamental drawback. Recall that in supervised learning, the distribution of the training data is decoupled from the learned model, whereas in imitation learning, the agent’s policy affects what state is queried next. The mismatch between training and testing distributions, also known as covariate shift (Ross & Bagnell, 2010; Ross et al., 2011), yields significant compounding errors. Therefore, BC often suffers from poor generalization.

• Inverse Reinforcement Learning (IRL, Russell (1998); Ng et al. (2000); Finn et al. (2016); Levine & Koltun (2012)). IRL treats the IL problem as bi-level optimization. Specifically, it finds a reward function, under which the expert policy is uniquely optimal. Though IRL does not have the error compounding issue, its computation is very inefficient. Many existing IRL methods need to solve
a sequence of computationally expensive reinforcement learning problems, due to their bi-level optimization nature. Therefore, they often fail to scale to large and high dimensional environments.

More recently, Ho & Ermon (2016) propose a Generative Adversarial Imitation Learning (GAIL) method, which obtains significant performance gains over existing IL methods in imitating complex expert policies in large and high-dimensional environments. GAIL generalizes IRL by formulating the IL problem as minimax optimization, which can be solved by alternating gradient-type algorithms in a more scalable and efficient manner.

Specifically, we consider an infinite horizon Markov Decision Process (MDP), where \( \mathcal{S} \) denotes the state space, \( \mathcal{A} \) denotes the action space, \( P \) denotes the Markov transition kernel, \( r^* \) denotes the reward function, and \( p_0 \) denotes the distribution of the initial state. We assume that the Markov transition kernel \( P \) is fixed and there is an unknown expert policy \( \pi^* : \mathcal{S} \rightarrow \mathcal{A} \), where \( \mathcal{P}(\mathcal{A}) \) denotes the set of distributions over the action space. As can be seen, \( \{s_t\}_{t=0}^{T-1} \) essentially forms a Markov chain with the transition kernel induced by \( \pi^* \) as \( P^{\pi^*}(s, s') = \sum_{a \in \mathcal{A}} \pi^*(a | s) \cdot P(s' | s, a) \).

Given \( n \) demonstration trajectories from \( \pi^* \) denoted by \( \{s_t^{(i)}, a_t^{(i)}\}_{t=0}^{T-1} \), where \( i = 1, ..., n \), \( s_0 \sim p_0 \), \( a_t \sim \pi^*(\cdot | s_t) \), and \( s_{t+1} \sim P(\cdot | s_t, a_t) \), GAIL aims to learn \( \pi^* \) by solving the following minimax optimization problem,

\[
\min_{\pi} \max_{\mathcal{R} \in \mathcal{R}} [\mathbb{E}_\pi r(s, a) - \mathbb{E}_{\pi^*} r(s, a)],
\]

where \( \mathbb{E}_\pi [r(s, a)] = \lim_{T \rightarrow \infty} \mathbb{E}[\frac{1}{T} \sum_{t=0}^{T-1} r(s_t, a_t) | \pi] \) denotes the average reward under the policy \( \pi \) when the reward function is \( r \), and \( \mathbb{E}_{\pi^*} [r(s, a)] = \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{n} \mathbb{E}[r(s_t^{(i)}, a_t^{(i)})] \) denotes the empirical average reward over the demonstration trajectories. As shown in (1), GAIL aims to find a policy, which attains an average reward similar to that of the expert policy with respect to any reward belonging to the function class \( \mathcal{R} \).

For large and high-dimensional imitation learning problems, we often encounter infinitely many states. To ease computation, we need to consider function approximations. Specifically, suppose that for every \( s \in \mathcal{S} \) and \( a \in \mathcal{A} \), there are feature vectors \( \psi_s \in \mathbb{R}^{d_S} \) and \( \psi_a \in \mathbb{R}^{d_A} \) associated with \( a \) and \( s \), respectively. Then we can approximate the policy and reward as

\[
\pi(\cdot | s) = \tilde{\pi}_\omega(\psi_s) \quad \text{and} \quad r(s, a) = \tilde{r}_\theta(\psi_s, \psi_a),
\]

where \( \tilde{\pi} \) and \( \tilde{r} \) belong to certain function classes (e.g. reproducing kernel Hilbert space or deep neural networks,Ormoneit & Sen (2002); LeCun et al. (2015)) associated with parameters \( \omega \) and \( \theta \), respectively. Accordingly, we can optimize (1) with respect to the parameters \( \omega \) and \( \theta \) by scalable alternating gradient-type algorithms.

Although GAIL has achieved significant empirical progress, its theoretical properties are still largely unknown. There are three major difficulties when analyzing GAIL: 1). There exists temporal dependency in the demonstration trajectories/data due to their sequential nature (Howard, 1960; Puterman, 2014; Abounadi et al., 2001); 2). GAIL is formulated as a minimax optimization problem. Most of existing learning theories, however, focus on empirical risk minimization problems, and therefore are not readily applicable (Vapnik, 2013; Mohri et al., 2018; Anthony & Bartlett, 2009); 3). The minimax optimization problem in (1) does not have a convex-concave structure, and therefore existing theories in convex optimization literature cannot be applied for analyzing the alternating stochastic gradient-type algorithms (Willem, 1997; Ben-Tal & Nemirovski, 1998; Murray & Overton, 1980; Chambolle & Pock, 2011; Chen et al., 2014). Some recent results suggest to use stage-wise stochastic gradient-type algorithms (Rafique et al., 2018; Dai et al., 2017). More specifically, at every iteration, they need to solve the inner maximization problem up to a high precision, and then apply stochastic gradient update to the outer minimization problem. Such algorithms, however, are rarely used by practitioners, as they are inefficient in practice (due to the computationally intensive inner maximization).

To bridge such a gap between practice and theory, we establish the generalization properties of GAIL and the convergence properties of the alternating mini-batch stochastic gradient algorithm for solving (1). Specifically, our contributions can be summarized as follows:

- We formally define the generalization of GAIL under the “so-called” \( \mathcal{R} \)-reward distance, and then show that the generalization of GAIL can be guaranteed under reward distance as long as the class of the reward functions is properly controlled;
We provide sufficient conditions, under which an alternating mini-batch stochastic gradient algorithm can efficiently solve the minimax optimization in (1), and attains sublinear convergence to a stationary solution.

To the best of our knowledge, these are the first results on statistical and computational theories of imitation learning with reward/policy function approximations.

Our work is related to Syed et al. (2008); Cai et al. (2019). Syed et al. (2008) study the generalization and computational properties of apprenticeship learning. Since they assume that the state space of the underlying Markov decision process is finite, they do not consider any reward/policy function approximations; Cai et al. (2019) study the computational properties of imitation learning under a simple control setting. Their assumption on linear policy and quadratic reward is very restrictive, and does not hold for many real applications.

**Notation.** Given a vector $x = (x_1, ..., x_d) \in \mathbb{R}^d$, we define $\|x\|_2^2 = \sum_{j=1}^d x_j^2$. Given a function $f : \mathbb{R}^d \mapsto \mathbb{R}$, we denote its $\ell_\infty$ norm as $\|f\|_\infty = \max_x |f(x)|$.

## 2 Generalization of GAIL

To analyze the generalization properties of GAIL, we first assume that we can access an infinite number of the expert’s demonstration trajectories (underlying population), and that the reward function is chosen optimally within some large class of functions. This allows us to remove the maximum operation from (1), which leads to an interpretation of how and in what sense the resulting policy is close to the true expert policy. Before we proceed, we first introduce some preliminaries.

**Definition 1** (Stationary Distribution). Note that any policy $\pi$ induces a Markov chain on $S \times A$. The transition kernel is given by

$$P_\pi(s', a' | s, a) = \pi(a' | s') \cdot P(s' | s, a), \quad \forall (s, a), (s', a') \in S \times A.$$  

When such a Markov chain is aperiodic and recurrent, we denote its stationary distribution as $\rho_\pi$.

Note that a policy $\pi$ is uniquely determined by its stationary distribution $\rho_\pi$ in the sense that

$$\pi(a | s) = \rho_\pi(s, a)/\sum_{a \in A} \rho_\pi(s, a).$$

Then we can write the expected average reward of $r(s, a)$ under the policy $\pi$ as

$$\mathbb{E}_\pi[r(s, a)] = \lim_{T \to \infty} \mathbb{E}\left[\frac{1}{T} \sum_{t=0}^{T-1} r(s_t, a_t) | \pi_\pi\right] = \mathbb{E}_{\rho_\pi}[r(s, a)] = \sum_{(s, a) \in S \times A} \rho_\pi(s, a) \cdot r(s, a).$$

We further define the $\mathcal{R}$-distance between two policies $\pi$ and $\pi'$ as follows.

**Definition 2.** Let $\mathcal{R}$ denote a class of symmetric reward functions from $S \times A$ to $\mathbb{R}$, i.e., if $r \in \mathcal{R}$, then $-r \in \mathcal{R}$. Given two policy $\pi'$ and $\pi$, the $\mathcal{R}$-distance for GAIL is defined as

$$d_\mathcal{R}(\pi, \pi') = \sup_{r \in \mathcal{R}} \left(\mathbb{E}_{\pi}(r(s, a)) - \mathbb{E}_{\pi'}(r(s, a))\right).$$

The $\mathcal{R}$-distance of Markov decision processes is essentially an Integral Probability Metric (IPM) over stationary distributions (Müller, 1997). For different choices of $\mathcal{R}$, we have various $\mathcal{R}$-distances. For example, we can choose $\mathcal{R}$ as the class of all 1-Lipschitz continuous functions, which yields that $d_\mathcal{R}(\pi, \pi')$ is the Wasserstein distance between $\rho_\pi$ and $\rho_{\pi'}$ (Vallender, 1974). For computational convenience, GAIL and its variants usually choose $\mathcal{R}$ as a class of functions from some reproducing kernel Hilbert space, or a class of neural network functions.

**Definition 3.** Given $n$ demonstration trajectories from time 0 to $T - 1$ obtained by an expert policy $\pi^*$ denoted by $(s_t^{(i)}, a_t^{(i)})_{i=0}^{T-1}$, where $i = 1, ..., n$, a policy $\hat{\pi}$ learned by GAIL generalizes under the $\mathcal{R}$-distance $d_\mathcal{R}(\cdot, \cdot)$ with generalization error $\epsilon$, if with high probability, we have

$$|d_\mathcal{R}(\pi^*_n, \hat{\pi}) - d_\mathcal{R}(\pi^*, \hat{\pi})| \leq \epsilon,$$

where $d_\mathcal{R}(\pi^*_n, \hat{\pi})$ is the empirical $\mathcal{R}$-distance between $\pi^*$ and $\hat{\pi}$ defined as

$$d_\mathcal{R}(\pi^*_n, \hat{\pi}) = \sup_{r \in \mathcal{R}} \left(\mathbb{E}_{\pi^*_n}(r(s, a)) - \mathbb{E}_\hat{\pi}(r(s, a))\right) \quad \text{with} \quad \mathbb{E}_{\pi^*_n}[r(s, a)] = \frac{1}{nT} \sum_{i=1}^n \sum_{t=0}^{T-1} r(s_t^{(i)}, a_t^{(i)}).$$

The generalization of GAIL implies that the $\mathcal{R}$-distance between the expert policy $\pi^*$ and the learned policy $\hat{\pi}$ is close to the empirical $\mathcal{R}$-distance between them. Our analysis aims to prove the former distance to be small, whereas the latter one is what we attempts to minimize in practice.

We then introduce the assumptions on the underlying Markov decision process and expert policy.
**Assumption 1.** Under the expert policy $\pi^*$, $(s_t, a_t)_{t=0}^{T-1}$ forms a stationary and exponentially $\beta$-mixing Markov chain, i.e.,
\[
\beta(k) = \sup_{n} E_{B \in \sigma_0} \sup_{A \in \sigma_{n+k}} |P(A|B) - P(A)| \leq \beta_0 \exp(-\beta_1 k^\alpha),
\]
where $\beta_0, \beta_1, \alpha$ are positive constants, and $\sigma_j$ is the $\sigma$-algebra generated by $(s_t, a_t)_{t=1}^i$ for $i \leq j$.

Moreover, for every $s \in S$ and $a \in A$, there are feature vectors $\psi_s \in \mathbb{R}^{dS}$ and $\psi_a \in \mathbb{R}^{dA}$ associated with $a$ and $s$, respectively, and $\psi_s$ and $\psi_a$ are uniformly bounded, where
\[
\|\psi_s\|_2 \leq 1 \quad \text{and} \quad \|\psi_a\|_2 \leq 1, \quad \forall s \in S \quad \text{and} \quad \forall a \in A.
\]

Assumption 1 requires the underlying MDP to be ergodic (Levin & Peres, 2017), which is a commonly studied assumption in exiting reinforcement learning literature on maximizing the expected average reward (Strehl & Littman, 2005; Li et al., 2011; Brafman & Tennenholtz, 2002; Kearns & Singh, 2002). The feature vectors associated with $a$ and $s$ allow us to apply function approximations to parameterize the reward and policy functions. Accordingly, we write the reward function as $r(s, a) = \tilde{r}(\psi_s, \psi_a)$, which is assumed to be bounded.

**Assumption 2.** The reward function class is uniformly bounded, i.e., $\|r\|_\infty \leq B_r$ for any $r \in \mathcal{R}$.

Now we proceed with our main result on generalization properties of GAIL. We use $N'(\mathcal{R}, \epsilon, \|\cdot\|_\infty)$ to denote the covering number of the function class $\mathcal{R}$ under the $\ell_\infty$ distance $\|\cdot\|_\infty$.

**Theorem 1 (Main Result).** Suppose Assumptions 1-2 hold, and the policy learned by GAIL satisfies
\[
d_{\mathcal{R}}(\pi^*, \hat{\pi}) - \inf_{\pi} d_{\mathcal{R}}(\pi^*, \pi) < \epsilon,
\]
where the infimum is taken over all possible learned policies. Then with probability at least $1 - \delta$ over the joint distribution of $\{(a_t^{(i)}, s_t^{(i)})_{t=0}^{T-1}\}_{i=1}^n$, we have
\[
d_{\mathcal{R}}(\pi^*, \hat{\pi}) - \inf_{\pi} d_{\mathcal{R}}(\pi^*, \pi) \leq O\left(\frac{B_r}{\sqrt{nT/\zeta}} \log N\left(\mathcal{R}, \sqrt{\frac{\zeta}{nT}}, \|\cdot\|_\infty\right) + B_r \sqrt{\frac{\log(1/\delta)}{nT/\zeta}}\right) + \epsilon,
\]
where $\zeta = (\beta_1^{-1} \log \frac{B_r}{\delta})^\frac{2}{\alpha}$.

Theorem 1 implies that the policy $\hat{\pi}$ learned by GAIL generalizes as long as the complexity of the function class $\mathcal{R}$ is well controlled. To the best of our knowledge, this is the first result on the generalization of imitation learning with function approximations. As the proof of Theorem 1 is involved, we only present a sketch due to space limit. More details are provided in Appendix A.1.

**Proof Sketch.** Our analysis relies on characterizing the concentration property of the empirical average reward under the expert policy. For notational simplicity, we define
\[
\phi = \mathbb{E}_{\pi^*} r(s, a) - \frac{1}{nT} \sum_{i=1}^n \sum_{t=0}^{T-1} r(s_t^{(i)}, a_t^{(i)}).
\]

The key challenge comes from the fact that $(s_t^{(i)}, a_t^{(i)})'$s are dependent. To handle such a dependency, we adopt the independent block technique from Yu (1994). Specifically, we partition every trajectory into disjoint blocks (where the block size is of the order $O((\log(T) + \log(1/\delta))^{1/\alpha})$, and construct two separable trajectories: One contains all blocks with odd indices (denoted by $B_{\text{odd}}$), and the other contains all those with even indices (denoted by $B_{\text{even}}$). We define
\[
\phi_1 = \mathbb{E}_{\pi^*} r(s, a) - \frac{2}{nT} \sum_{i=1}^n \sum_{(s_t^{(i)}, a_t^{(i)}) \in B_{\text{odd}}} r(s_t^{(i)}, a_t^{(i)}),
\]
and analogously for $\phi_2$ with $(s_t^{(i)}, a_t^{(i)}) \in B_{\text{even}}$. Then we have
\[
P(\sup_{r \in \mathcal{R}} \phi \geq \epsilon) \leq P(\sup_{r \in \mathcal{R}} \phi_1 \geq \epsilon) + P(\sup_{r \in \mathcal{R}} \phi_2 \geq \epsilon).
\]

We consider a block-wise independent counterpart of $\phi_1$ denoted by $\tilde{\phi}_1$, where each block is sampled independently from the same Markov chain as $\phi_1$, i.e., $\tilde{\phi}_1$ has independent blocks of samples from the same exponentially $\beta$-mixing Markov chain. Accordingly, we denote $\phi_1$ as
\[
\tilde{\phi}_1 = \mathbb{E}_{\pi^*} r(s, a) - \frac{2}{nT} \sum_{i=1}^n \sum_{(s_t^{(i)}, a_t^{(i)}) \in B_{\text{odd}}} r(s_t^{(i)}, a_t^{(i)}),
\]
where $\overline{B}_{\text{odd}}$ denotes i.i.d. blocks of samples. Now we bound the difference between $\phi_1$ and $\tilde{\phi}_1$ by

$$
P(\sup_{r \in \mathcal{R}} \phi_1 - \mathbb{E}[\sup_{r \in \mathcal{R}} \tilde{\phi}_1] \geq \varepsilon - \mathbb{E}[\sup_{r \in \mathcal{R}} \tilde{\phi}_1]) \leq P(\sup_{r \in \mathcal{R}} \tilde{\phi}_1 - \mathbb{E}[\sup_{r \in \mathcal{R}} \tilde{\phi}_1] \geq \varepsilon - \mathbb{E}[\sup_{r \in \mathcal{R}} \tilde{\phi}_1]) + C\beta T/(\log(T) + \log(1/\delta))^{1/\alpha},$$

where $C$ is a constant, and $\beta$ is the mixing coefficient, and $P(\sup_{r \in \mathcal{R}} \phi_1 - \mathbb{E}[\sup_{r \in \mathcal{R}} \tilde{\phi}_1] \geq \varepsilon - \mathbb{E}[\sup_{r \in \mathcal{R}} \tilde{\phi}_1])$ can be bounded using the empirical process technique for independent random variables. The details of the above inequality can be found in Corollary 3 in Appendix A.1, where the proof technique is adapted from Lemma 1 in Mohri & Rostamizadeh (2009). Let $\tilde{\phi}_2$ be defined analogously as $\phi_1$. With a similar argument further applied to $\phi_2$ and $\tilde{\phi}_2$, we obtain

$$
P(\sup_{r \in \mathcal{R}} \phi \geq \varepsilon) \leq 2P(\sup_{r \in \mathcal{R}} \tilde{\phi}_1 - \mathbb{E}[\sup_{r \in \mathcal{R}} \tilde{\phi}_1] \geq \varepsilon - \mathbb{E}[\sup_{r \in \mathcal{R}} \tilde{\phi}_1]) + 2C\beta T/(\log(T) + \log(1/\delta))^{1/\alpha}.
$$

The rest of our analysis follows the PAC-learning framework using Rademacher complexity and is omitted (Mohri et al., 2018). We complete the proof sketch.

Example 1: Reproducing Kernel Reward Function. One popular option to parameterize the reward by functions is the reproducing kernel Hilbert space (RKHS, Kim & Park (2018); Li et al. (2018)). There have been several implementations of RKHS, and we consider the feature mapping approach. Specifically, we consider $g : \mathbb{R}^{2d} \times \mathbb{R}^{d_A} \rightarrow \mathbb{R}^q$, and the reward can be written as

$$r(s, a) = r_0(\psi_s, \psi_a) = \theta^T g(\psi_s, \psi_a),$$

where $\theta \in \mathbb{R}^q$. We require $g$ to be Lipschitz continuous with respect to $(\psi_s, \psi_a)$.

Assumption 3. The feature mapping $g$ satisfies $g(0, 0) = 0$, and there exists a constant $\rho_g$ such that for any $\psi_s, \psi'_s, \psi_a$ and $\psi'_a$, we have

$$\|g(\psi_s, \psi_a) - g(\psi'_s, \psi'_a)\|_2^2 \leq \rho_g \sqrt{\|\psi_s - \psi'_s\|_2^2 + \|\psi_a - \psi'_a\|_2^2}.$$}

Assumption 3 is mild and satisfied by popular feature mappings, e.g., random Fourier feature mapping\footnote{More precisely, Assumption 3 actually holds with overwhelming probability over the distribution of the random mapping.} (Rahimi & Recht, 2008; Bach, 2017). The next corollary presents the generalization bound of GAIL using feature mapping.

Corollary 1. Suppose $\|\theta\|_2 \leq B_\theta$. For large enough $n$ and $T$, with probability at least $1 - \delta$ over the joint distribution of $\{(a_i, s_i)_{i=1}^{T-1}\}_{i=1}^n$, we have

$$d_\mathcal{R}(\pi^*, \hat{\pi}) - \inf_{\pi} d_\mathcal{R}(\pi^*, \pi) \leq O \left( \frac{\rho_g B_\theta}{\sqrt{nT/\zeta}} \sqrt{q \log \left( \rho_g B_\theta \sqrt{nT/\zeta} \right) + \rho_g B_\theta \sqrt{\log(1/\delta) / nT/\zeta}} \right) + \epsilon.$$

Corollary 1 indicates that with respect to a class of properly normalized reproducing kernel reward functions, GAIL generalizes in terms of the $R$-distance.

Example 2: Neural Network Reward Function. Another popular option to parameterize the reward function is to use neural networks. Specifically, let $\sigma(v) = [\max\{v_1, 0\}, ..., \max\{v_d, 0\}]^T$ denote the ReLU activation for $v \in \mathbb{R}^d$. We consider a $D$-layer feedforward neural network with ReLU activation as follows,

$$r(s, a) = r_0(\psi_s, \psi_a) = W_D^T \sigma(W_{D-1}\sigma(...\sigma(W_1[\psi_1^T, \psi_s^T]^T)),$$

where $\mathcal{W} = \{W_i | W_i \in \mathbb{R}^{d_{i-1} \times d_i}, i = 1, ..., D - 1, W_D \in \mathbb{R}^{d_{D-1}}\}$ and $d_0 = d_A + d_S$. The next corollary presents the generalization bound of GAIL using neural networks.

Corollary 2. Suppose $\|W_i\|_2 \leq 1$, where $i = 1, ..., D$. For large enough $n$ and $T$, with probability at least $1 - \delta$ over the joint distribution of $\{(a_i, s_i)_{i=1}^{T-1}\}_{i=1}^n$, we have

$$d_\mathcal{R}(\pi^*, \hat{\pi}) - \inf_{\pi} d_\mathcal{R}(\pi^*, \pi) \leq O \left( \frac{1}{\sqrt{nT/\zeta}} \sqrt{d^2D \log \left( D\sqrt{d^2nT/\zeta} \right) + \frac{\log(1/\delta)}{nT/\zeta}} \right) + \epsilon.$$

Corollary 2 indicates that with respect to a class of properly normalized neural network reward functions, GAIL generalizes in terms of the $R$-distance.
Remark 1 (The Tradeoff between Generalization and Representation of GAIL). As can be seen from Definition 2, the \( R \)-distances are essentially differentiating two policies. For the Wasserstein-type distance, i.e., \( R \) contains all 1-Lipschitz continuous functions, if \( d_R(\pi, \pi') \) is small, it is safe to conclude that two policies \( \pi \) and \( \pi' \) are nearly the same almost everywhere. However, when we choose \( R \) to be the reproducing kernel Hilbert space or the class of neural networks with relatively small complexity, \( d_R(\pi, \pi') \) can be small even if \( \pi \) and \( \pi' \) are not very close. Therefore, we need to choose a sufficiently diverse class of reward functions to ensure that we recover the expert policy.

As Theorem 1 suggests, however, that we need to control the complexity of the function class \( R \) to guarantee the generalization. This implies that when parameterizing the reward function, we need to carefully choose the function class to attain the optimal tradeoff between generalization and representation of GAIL.

3 COMPUTATION OF GAIL

To investigate the computational properties of GAIL, we parameterize the reward by functions belonging to some reproducing kernel Hilbert space. The implementation is based on feature mapping, as mentioned in the previous section. The policy can be parameterized by functions belonging to some reproducing kernel Hilbert space or some class of deep neural networks with parameter \( \omega \). Specifically, we denote \( \pi(a|s) = \tilde{\pi}_\omega(a|s) \), where \( \tilde{\pi}_\omega(a|s) \) is the parameterized policy mapping from \( \mathbb{R}^{d_\omega} \) to a simplex in \( \mathbb{R}_+^d \), where \( |A| = d \). For computational convenience, we consider solving a slightly modified minimax optimization problem:

\[
\min_{\omega} \max_{\|\theta\|_2 \leq \kappa} \mathbb{E}_{\tilde{\pi}_\omega} [r(s, a)] - \mathbb{E}_{\pi} [r(s, a)] - \lambda H(\tilde{\pi}_\omega) - \frac{\mu}{2} \|\theta\|_2^2,
\]

(2)

where \( \tilde{\pi}_\omega(s, a) = \theta^T g(s, a) \), \( H(\tilde{\pi}_\omega) \) is some regularizer for the policy (e.g., causal entropy regularizer, Ho & Ermon (2016)), and \( \lambda \) and \( \mu > 0 \) are tuning parameters. Compared with (1), the additional regularizers in (2) can improve the optimization landscape, and help mitigate computational instability in practice.

We apply the alternating mini-batch stochastic gradient algorithm to (2). Specifically, we denote the objective function in (2) as \( F(\omega, \theta) \) for notational simplicity. At the \( (t+1) \)-th iteration, we take

\[
\theta^{(t+1)} = \Pi_\kappa (\theta^{(t)} + \frac{\eta_\theta}{q_\theta} \sum_{j \in M_\theta^{(t)}} \nabla \theta j_f(\omega^{(t)}, \theta^{(t)})) \quad \text{and} \quad \omega^{(t+1)} = \omega^{(t)} - \frac{\eta_\omega}{q_\omega} \sum_{j \in M_\omega^{(t)}} \nabla \omega j_f(\omega^{(t)}, \theta^{(t+1)}),
\]

(3)

(4)

where \( \eta_\theta \) and \( \eta_\omega \) are learning rates, the projection \( \Pi_\kappa(v) = 1(\|v\|_2 \leq \kappa) \cdot v + 1(\|v\|_2 > \kappa) \cdot \kappa \cdot v / \|v\|_2 \), \( j_f(\cdot, \cdot) \)’s and \( \omega j_f(\cdot, \cdot) \)’s are independent stochastic approximations of \( \nabla F \) (Sutton et al., 2000), and \( M_\theta^{(t)} \), \( M_\omega^{(t)} \) are mini-batches with sizes \( q_\theta \) and \( q_\omega \), respectively. Before we proceed with the convergence analysis, we impose the follow assumptions on the problem.

Assumption 4. There are two positive constants \( M_\omega \) and \( M_\theta \) such that for any \( \omega \) and \( \|\theta\|_2 \leq \kappa \),

\( \text{Unbiased: } \mathbb{E} \nabla j_f(\omega, \theta) = \mathbb{E} \nabla \omega j_f(\omega, \theta) = \nabla F(\omega, \theta) \),

\( \text{Bounded: } \mathbb{E} \|\nabla \omega j_f(\omega, \theta) - \nabla \omega F(\omega, \theta)\|_2^2 \leq M_\omega \) and \( \mathbb{E} \|\nabla \theta j_f(\omega, \theta) - \nabla \theta F(\omega, \theta)\|_2^2 \leq M_\theta \).

Assumption 4 requires the stochastic gradient to be unbiased with a bounded variance, which is a common assumption in existing optimization literature (Nemirovski et al., 2009; Ghadimi & Lan, 2013; Duchi et al., 2011; Bottou, 2010).

Assumption 5. (i) For any \( \omega \), there exists some constant \( \chi > 0 \) and \( v \in (0, 1) \) such that

\[
\|P_{\tilde{\pi}_\omega}(s', a' | s, a) - P_{\pi}(s' | s, a)\|_{TV} \leq \chi v^t,
\]

where \( P_{\tilde{\pi}_\omega}(s', a' | s, a) = \tilde{\pi}_\omega(a' | s) P(s' | s, a) \) is the transition kernel induced by \( \tilde{\pi}_\omega \), \( P_{\pi} \) is the initial distribution of \((s_0, a_0)\), and \( \rho_{\pi_\omega} \) is the stationary distribution induced by \( \tilde{\pi}_\omega \).

(ii) There exist constants \( S_\pi, B_\omega, L_\rho, L_Q \) such that for any \( \omega, \omega' \), we have

\[
\|\nabla \log(\tilde{\pi}_\omega(a|s)) - \nabla \log(\tilde{\pi}_\omega'(a|s))\|_2 \leq S_\pi \|\omega - \omega'\|_2, \quad \|\nabla \log(\tilde{\pi}_\omega(a|s))\|_2 \leq B_\omega, \quad \|\rho_{\pi_\omega} - \rho_{\pi_\omega}'\|_{TV} \leq L_\rho \|\omega - \omega'\|_2, \quad \|Q^{\tilde{\pi}_\omega} - Q^{\tilde{\pi}_\omega}'\|_{TV} \leq L_Q \|\omega - \omega'\|_2,
\]

where \( Q^{\tilde{\pi}_\omega}(s, a) = \sum_{t=0}^\infty \mathbb{E} [\tilde{\pi}_\omega(s_t, a_t) | s_0 = s, a_0 = a, \tilde{\pi}_\omega] \) is the action-value function.

(iii) There exist constants \( B_H \) and \( S_H \) such that for any \( \omega, \omega' \), we have

\[
H(\tilde{\pi}_\omega) \leq B_H, \quad \|\nabla H(\tilde{\pi}_\omega) - \nabla H(\tilde{\pi}_\omega')\|_2 \leq S_H \|\omega - \omega'\|_2.
\]
Note that (i) of Assumption 5 requires the Markov Chain to be geometrically mixing. (ii) and (iii) state some commonly used regularity conditions for policies (Sutton et al., 2000; Pirotta et al., 2015).

We then define $L$-stationary points of $F$. Specifically, we say that $(\omega^*, \theta^*)$ is a stationary point of $F$, if and only if, for any fixed $\alpha > 0$,
$$
\nabla_\omega F(\omega^*, \theta^*) = 0 \quad \text{and} \quad \theta^* - \Pi_\kappa(\theta^* + \alpha \nabla_\theta F(\omega^*, \theta^*)) = 0.
$$
The $L$-stationarity is a generalization of the stationary point for unconstrained optimization, and is a necessary condition for optimality. Accordingly, we take $\alpha = 1$ and measure the sub-stationarity of the algorithm at the iteration $N$ by
$$
J_N = \min_{1 \leq t \leq N} \mathbb{E}[\|\theta(t) - \Pi_\kappa(\theta(t) + \nabla_\theta F(\omega(t), \theta(t)))\|^2_2 + \mathbb{E}[\|\nabla_\omega F(\omega(t), \theta(t+1))\|^2_2].
$$
We then state the global convergence of the alternating mini-batch stochastic gradient algorithm.

**Theorem 2.** Suppose Assumptions 1-5 hold. We choose step sizes $\eta_\theta, \eta_\omega$ satisfying
$$
\eta_\omega \leq \min \left\{ \frac{L_\omega}{S_\omega(8L_\omega + 2)}, \frac{1}{2L_\omega} \right\}, \quad \eta_\theta \leq \min \left\{ \frac{1}{150\mu}, \frac{7L_\omega + 1}{150S_\omega^2}, \frac{1}{100(2\mu + S_\omega)} \right\},
$$
and meanwhile $\eta_\omega/\eta_\theta \leq \mu/(30L_\omega + 5)$, where $L_\omega = 2\sqrt{2}(S_\pi + 2B_\omega L_\rho)\kappa_p\rho_\chi/(1 - v) + B_\omega L_Q$, and $S_\omega = 2\sqrt{2}Q\kappa_p\rho_\chi B_\omega/(1 - v)$. Given any $\epsilon > 0$, we choose batch sizes $q_\theta = O(1/\epsilon)$ and $q_\omega = O(1/\epsilon)$. Then we need at most
$$
N = \mathcal{O}(C_0 + 4\sqrt{2}\rho_\kappa + \mu\kappa^2 + 2\lambda B_H)\epsilon^{-1}
$$
iterations such that $J_N \leq \epsilon$, where $C_0$ depends on the initialization, and $\eta$ depends on $\eta_\omega$ and $\eta_\theta$.

Here $\tilde{O}$ hides linear or quadratic dependence on some constants in Assumptions 1-5. Theorem 2 shows that the minimax optimization problem in (2) does not have a convex-concave structure, the alternating mini-batch stochastic gradient algorithm still guarantees to converge to a stationary point. We are not aware of any similar results for GAIL in existing literature.

**Proof Sketch.** We prove the convergence by showing
$$
\mathbb{E}[\|\theta(t) - \Pi_\kappa(\theta(t) + \nabla_\theta F(\omega(t), \theta(t)))\|^2_2 + \mathbb{E}[\|\nabla_\omega F(\omega(t), \theta(t+1))\|^2_2] \leq C + N\epsilon/2,
$$
where $C$ is a constant and $N\epsilon/2$ is the accumulation of noise in stochastic approximations of $\nabla F$. Then we straightforwardly have $NJ_N \leq C + N\epsilon/2$. Dividing both sizes by $N$, we can derive the desired result. The main difficulty of showing (5) comes from the fact that the outer minimization problem is nonconvex and we cannot solve the inner maximization problem exactly. To overcome this difficulty, we construct a monotonically decreasing potential function:
$$
\mathcal{E}(t) = \mathbb{E} F(\omega(t), \theta(t)) + s\left( (1 + 2\eta_\omega L_\omega)/2 \cdot \mathbb{E}[\|\omega(t) - \omega(t-1)\|^2_2 + \mathbb{E}[\|\nabla_\omega F(\omega(t), \theta(t+1))\|^2_2] \leq C + N\epsilon/2,
$$
for a constant $s$ to be chosen later. Denote $\xi_\theta(t)$ and $\xi_\omega(t)$ as the i.i.d. noise of the stochastic gradients. The following lemma characterizes the decrement of the potential function at each iteration.

**Lemma 1.** With the step sizes $\eta_\theta$ and $\eta_\omega$ chosen as in Theorem 2, we have
$$
\mathcal{E}(t+1) - \mathcal{E}(t) \leq - k_1 \mathbb{E}[\|\omega(t+1) - \omega(t)\|^2_2] - k_2 \mathbb{E}[\|\theta(t+1) - \theta(t)\|^2_2] + k_3 \mathbb{E}[\|\theta(t+1) - \theta(t)\|^2_2] + k_4 \mathbb{E}[\|\nabla_\omega F(\omega(t), \theta(t+1))\|^2_2]
$$
for a constant $\nu$ depending on $F$, $\eta_\omega$, and $\eta_\omega$. Moreover, we have constants $k_1, k_2, k_3, k_4, k_5 > 0$ for $s = 8/(\kappa_\omega^2(58L_\omega + 9))$.

Let $k = 1/\min\{k_1, k_2, k_3\}$ and $\phi = \max\{1, 1/\eta_\theta^2, 1/\eta_\omega^2\}$. We obtain
$$
\mathbb{E}[\|\theta(t+1) - \Pi_\kappa(\theta(t+1) + \nabla_\theta F(\omega(t), \theta(t+1)))\|^2_2 + \mathbb{E}[\|\nabla_\omega F(\omega(t), \theta(t+1))\|^2_2]
$$
\leq k \phi \mathbb{E}[\|\xi_\theta(t+1) - \xi_\theta(t)\|^2_2 + \|\xi_\omega(t+1) - \omega(t)\|^2_2]
\leq k \phi (\mathcal{E}(t) - \mathcal{E}(N)) + k \phi N \nu \mathbb{E}[\|\xi_\theta(t+1)\|^2_2 + \|\xi_\omega(t+1)\|^2_2],
$$
where (i) follows from plugging in the update (3) as well as the contraction property of projection, and (ii) follows from Lemma 1. Choosing $q_\theta = 4k\phi\nu M_\theta/\epsilon$ and $q_\omega = 4k\phi\nu M_\omega/\epsilon$, we obtain
$$
\mathbb{E}[\|\theta(t+1) - \Pi_\kappa(\theta(t+1) + \nabla_\theta F(\omega(t), \theta(t+1)))\|^2_2 + \mathbb{E}[\|\nabla_\omega F(\omega(t), \theta(t+1))\|^2_2]
\leq k \phi (\mathcal{E}(t) - \mathcal{E}(N)) + N\epsilon.
$$
We have $\mathcal{E}(N) \geq \mathbb{E} F(\omega(N), \theta(N))$ by the construction of $\mathcal{E}(N)$. It is easy to verify that $F$ is lower bounded (Lemma 10 in Appendix B). Eventually, we complete the proof by substituting the lower bound and choosing $N = k \phi (2\mathcal{E}(1) + 4\sqrt{2}\rho_\kappa + \mu\kappa^2 + 2\lambda B_H)\epsilon^{-1}$. \(\square\)
4 Experiment

To verify our theory in Section 3, we conduct experiments in three reinforcement learning tasks: Acrobot, MountainCar, and Hopper. For each task, we first train an expert policy using the proximal policy optimization (PPO) algorithm in (Schulman et al., 2017) for 500 iterations, and then use the expert policy to generate the demonstration data. The demonstration data for every task contains 500 trajectories, each of which is a series of state action pairs throughout one episode in the environment. When training GAIL, we randomly select a mini-batch of trajectories, which contain at least 8192 state action pairs. We use PPO to update the policy parameters. This avoids the instability of the policy gradient algorithm, and improves the reproducibility of our experiments.

When training GAIL, we randomly select a mini-batch of trajectories, which contain at least 8192 state action pairs. We use PPO to update the policy parameters. This avoids the instability of the policy gradient algorithm, and improves the reproducibility of our experiments.

We use the same neural network architecture for all the environments. For policy, we use a fully connected neural network with two hidden layers of 128 neurons in each layer and tanh activation. For reward, we use a fully connected ReLU neural network with two hidden layers of 1024 and 512 neurons, respectively. To implement the kernel reward, we fix the first two layers of the neural network after random initialization and only update the third layer, i.e., the first two layers mimic the random feature mapping. We choose $\kappa = 1$ and $\mu = 0.3$. When updating the neural network reward, we use weight normalization in each layer (Salimans & Kingma, 2016).

When updating the kernel reward at each iteration, we choose to take the stochastic gradient ascent step for either once (i.e., alternating update in Section 3) or 10 times. When updating the neural network reward at each iteration, we choose to take the stochastic gradient ascent step for only once. We tune step size parameters for updating the policy and reward, and summarize the numerical results of the step sizes attaining the maximal average episode reward in Figure 1.

As can be seen, using multiple stochastic gradient ascent steps for updating the reward at each iteration yields similar performance as that of one step. We present the convergence analysis of using multiple stochastic gradient ascent steps for updating the reward in Appendix C. Moreover, we observe that parameterizing the reward by neural networks slightly outperform that of the kernel reward. However, its training process tends to be unstable and takes longer time to converge.

5 Discussions

Our proposed theories of GAIL are closely related to Generative Adversarial Networks (Goodfellow et al., 2014; Arjovsky et al., 2017): (1) The generalization of GANs is defined based on the integral probabilistic metric (IPM) between the synthetic distribution obtained by the generator network and the distribution of the real data (Arora et al., 2017). As the real data in GANs are considered as independent realizations of the underlying distribution, the generalization of GANs can be analyzed using commonly used empirical process techniques for i.i.d. random variables. GAIL, however, involves dependent demonstration data from experts, and therefore the analysis is more involved. (2) Our computational theory of GAIL can be applied to MMD-GAN and its variants, where the IPM is induced by some reproducing kernel Hilbert space (Li et al., 2017; Bińkowski et al., 2018; Arbel et al., 2018). The alternating mini-batch stochastic gradient algorithm attains a similar sublinear rate of convergence to a stationary solution.

Moreover, our computational theory of GAIL only considers the policy gradient update when learning the policy (Sutton et al., 2000). Extending to other types of updates such as natural policy gradient (Kakade, 2002), proximal policy gradient (Schulman et al., 2017) and trust region policy optimization (Schulman et al., 2015) is a challenging, but important future direction.
REFERENCES


11

A Proofs in Section 2

A.1 Proof of Theorem 1

We first consider \( n = 1 \). For notational simplicity, we denote \( x_t = (s_t, a_t) \) and \( \tau = \{ x_t \}_{t=0}^{T-1} \). Then the generalization gap is bounded by

\[
\begin{align*}
\Phi(\tau) &= d_R(\pi', \hat{\pi}) - \inf_{\pi'} d_R(\pi^*, \pi) - d_R(\pi^*, \hat{\pi}) \\
&= \inf_{\pi'} d_R(\pi^*, \pi) + \sup_{\pi'} d_R(\pi^*, \pi) \\
&\leq 2 \left( \sup_{\pi'} \mathbb{E}_{\pi'} r(s, a) - \frac{1}{T} \sum_{t=0}^{T-1} r(s_t, a_t) \right) + \epsilon.
\end{align*}
\]

Denote \( \Phi(\tau) = \sup_{\pi \in \cal{R}} \mathbb{E}_{\pi} r(s, a) \). We utilize the independent block technique first proposed in Yu (1994) to show the concentration of \( \Phi(\tau) \). Specifically, we partition \( \tau \) into \( 2m \) blocks of equal size \( b \). We denote two alternating sequences as

\[
\tau_0 = (X_1, X_2, \ldots, X_m), \quad \tau_1 = (X_1^{(1)}, X_2^{(1)}, \ldots, X_m^{(1)}), \quad X_i = (x_{2i-1}b+1, \ldots, x_{2i-1}b+b),
\]

where \( x_{2i-1} \) is i.i.d blocks of size \( b \). According to Lemma 2, we define \( \tilde{\tau}_0 = (\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_m) \), where \( \tilde{X}_i \) follows the same distribution as \( X_i \).

Before we proceed, we need to introduce a lemma which characterizes the relationship between the expectations of a bounded measurable function with respect to \( \tau_0 \) and \( \tilde{\tau}_0 \).

Lemma 2. (Yu, 1994) Suppose \( h \) is a measurable function bounded by \( M > 0 \) defined over the blocks \( X_i \), then the following inequality holds:

\[
|\mathbb{E}_{\tau_0}(h) - \mathbb{E}_{\tilde{\tau}_0}(h)| \leq (m - 1)M \beta(b),
\]

where \( \mathbb{E}_{\tau_0} \) denotes the expectation with respect to \( \tau_0 \), and \( \mathbb{E}_{\tilde{\tau}_0} \) denotes the expectation with respect to \( \tilde{\tau}_0 \).

Corollary 3. Applying Lemma 2, we have

\[
P_{\tau}(\Phi(\tau) > \epsilon) \leq 2P_{\tau_0}(\Phi(\tilde{\tau}_0) - \mathbb{E}_{\tilde{\tau}_0}[\Phi(\tilde{\tau}_0)]) > \epsilon - \mathbb{E}_{\tau_0}[\Phi(\tilde{\tau}_0)]) + 2(m - 1)\beta(b).
\]

Proof. Consider \( P(\Phi(\tau) > \epsilon) \), we have

\[
P_{\tau}(\Phi(\tau) > \epsilon) = P_{\tau}(\sup_{\pi \in \cal{R}} \mathbb{E}_{\pi} r(x) - \frac{1}{T} \sum_{t=0}^{T-1} r(x_t) > \epsilon)
\]

\[
\leq P_{\tau} \left( \sup_{\pi \in \cal{R}} \mathbb{E}_{\pi} r(x) - \frac{1}{2} \sum_{t=0}^{T-1} r(x_t) > \epsilon \right) + P_{\tau} \left( \sup_{\pi \in \cal{R}} \mathbb{E}_{\pi} r(x) - \frac{1}{2} \sum_{t=0}^{T-1} r(x_t) > \epsilon \right)
\]

\[
= \sup_{\pi} \mathbb{E}_{\pi} r(x) > 2\epsilon \]

\[
\leq P_{\tau_0}(\Phi(\tau_0) > \epsilon) + P_{\tau_1}(\Phi(\tau_1) > \epsilon)
\]

\[
= 2P_{\tau_0}(\Phi(\tau_0) > \epsilon)
\]

\[
= 2P_{\tau_0}(\Phi(\tau_0) - \mathbb{E}_{\tilde{\tau}_0}[\Phi(\tilde{\tau}_0)]) > \epsilon - \mathbb{E}_{\tau_0}[\Phi(\tilde{\tau}_0)])
\]

where the first inequality (8) follows from the convexity of supremum.

Applying Lemma 2 and setting \( h = \mathbb{1} \{ (\Phi(\tau_0) - \mathbb{E}_{\tilde{\tau}_0}[\Phi(\tilde{\tau}_0)]) > \epsilon - \mathbb{E}_{\tau_0}[\Phi(\tilde{\tau}_0)]) \}, \) we obtain

\[
P_{\tau_0}(\Phi(\tau_0) - \mathbb{E}_{\tilde{\tau}_0}[\Phi(\tilde{\tau}_0)]) > \epsilon - \mathbb{E}_{\tau_0}[\Phi(\tilde{\tau}_0)])
\]

\[
\leq P_{\tau_0}(\Phi(\tau_0) - \mathbb{E}_{\tilde{\tau}_0}[\Phi(\tilde{\tau}_0)]) > \epsilon - \mathbb{E}_{\tau_0}[\Phi(\tilde{\tau}_0)]) + 2(m - 1)\beta(b).
\]
Now combining (7) and (9), we obtain

\[ \Phi(\tau) = \sup_{r_b \in \mathcal{R}_b} E_{\tau^b} r_b(\tilde{X}_i) - \frac{1}{m} \sum_{i=1}^{m} r_b(\tilde{X}_i). \]

Given samples \( \tilde{X}_1, \ldots, \tilde{X}_i, \ldots, \tilde{X}_m \) and \( \tilde{X}_1, \ldots, \tilde{X}_i', \ldots, \tilde{X}_m \), we have

\[ \Phi(\tilde{X}_1, \ldots, \tilde{X}_i, \ldots, \tilde{X}_m) - \Phi(\tilde{X}_1, \ldots, \tilde{X}_i', \ldots, \tilde{X}_m) \leq \frac{2}{m} \left| r_b(\tilde{X}_i) \right| \leq \frac{2B_r}{m}. \]

Then by McDiarmid’s inequality, we have

\[ P_{\tau} (\Phi(\tau) > \epsilon) \leq 2 \exp \left( \frac{-m(\epsilon - \mathbb{E}_{\tau^b}[\Phi(\tau_0)])^2}{2B_r^2} \right) + 2(m - 1)\beta(b). \] (10)

By the argument of symmetrization, we have

\[ \mathbb{E}_{\tau^b}[\Phi(\tau_0)] \leq 2 \mathbb{E}_{\tau^b,\sigma} \left[ \frac{1}{m} \sup_{r_b \in \mathcal{R}_b} \sum_{i=1}^{m} \sigma_i r_b(\tilde{X}_i) \right], \] (11)

where \( \sigma_i \)’s are i.i.d. Rademacher random variables. Now we relate the Rademacher complexity (11) to its counterpart taking i.i.d samples. Specifically, we denote \( x_{i,j}^{(t)} \) as the \( t \)-th point of the \( j \)-th block.

Denote \( \tau_0^j \) as the collection of the \( j \)-th sample from each independent block \( \tilde{X}_i \) for \( i = 1, \ldots, m \). Plugging in the definition of \( r_b \), we have

\[
2 \mathbb{E}_{\tau^b,\sigma} \left[ \frac{1}{m} \sup_{r \in \mathcal{R}} \sum_{t=1}^{m} \sigma_i \frac{1}{b} \sum_{j=1}^{b} r(x_{i,j}^{(t)}) \right] \leq 2 \mathbb{E}_{\tau_0^j,\sigma} \left[ \frac{1}{b} \sum_{j=1}^{b} \frac{1}{r} \sup_{t \in \mathcal{T}} \sum_{t=1}^{m} \sigma_i r(x_{i,j}^{(t)}) \right] 
\]

\[
\leq \frac{2}{b} \sum_{j=1}^{b} \mathbb{E}_{\tau_0^j,\sigma} \left[ \frac{1}{r} \sup_{t \in \mathcal{T}} \sum_{t=1}^{m} \sigma_i r(x_{i,j}^{(t)}) \right] 
\]

\[
= \frac{2}{b} \sum_{j=1}^{b} \mathbb{E}_{\tau_0^j,\sigma} \left[ \frac{1}{r} \sum_{t \in \mathcal{T}} \sigma_i r(x_{i,j}^{(t)}) \right] 
\]

\[ = 2 \mathbb{E}_{\tau_0^j,\sigma} \left[ \frac{1}{m} \sup_{r \in \mathcal{R}} \sum_{t=1}^{m} \sigma_i r(x_{i,j}^{(t)}) \right]. \]

(12)

Setting the right-hand side of (10) to be \( \frac{\delta}{2} \) and substituting (12), we obtain, with probability at least \( 1 - \frac{\delta}{2} \), for all \( r \in \mathcal{R} \),

\[
\Phi(\tau) \leq 2 \mathbb{E}_{\tau_0^j,\sigma} \left[ \frac{1}{m} \sup_{r \in \mathcal{R}} \sum_{t=1}^{m} \sigma_i r(x_{i,j}^{(t)}) \right] + 2B_r \sqrt{\frac{\log \frac{4}{\delta}}{2m}}, \]

(13)

where \( \delta' = \delta - 4(m-1)\beta(b) \).

Then we denote

Rademacher complexity for \( \tilde{X}_i \)'s: \( \mathcal{R}_{\tilde{X}_i} = \mathbb{E}_{\tau_0^j,\sigma} \left[ \frac{1}{m} \sup_{r \in \mathcal{R}} \sum_{t=1}^{m} \sigma_i r(x_{i,j}^{(t)}) \right] \),

Empirical Rademacher complexity for \( \tilde{X}_i \)'s: \( \hat{\mathcal{R}}_{\tilde{X}_i} = \mathbb{E}_{\sigma} \left[ \frac{1}{m} \sup_{r \in \mathcal{R}} \sum_{t=1}^{m} \sigma_i r(x_{i,j}^{(t)}) \right] \),

Empirical Rademacher complexity for \( X_i \)'s: \( \hat{\mathcal{R}}_m = \mathbb{E}_{\sigma} \left[ \frac{1}{m} \sup_{r \in \mathcal{R}} \sum_{t=1}^{m} \sigma_i r(x_{i,j}^{(t)}) \right] \).
Applying Lemma 2 to the indicator function \( I \{ \mathcal{R}^\beta_m - \hat{\mathcal{R}}_m > \epsilon \} \), we obtain
\[
P(\mathcal{R}^\beta_m - \hat{\mathcal{R}}_m > \epsilon) \leq \mathbb{P}(\mathcal{R}^\beta_m - \hat{\mathcal{R}}_m^\beta > \epsilon) + (m - 1)\beta(2b - 1) \leq \mathbb{P}(\mathcal{R}^\beta_m - \hat{\mathcal{R}}_m^\beta > \epsilon) + (m - 1)\beta(b).
\]
It is straightforward to verify that given \( x_1^{(1)}, \ldots, x_1^{(b)} \) and \( x_1^{(1)}, \ldots, x_1^{(m)}, \ldots, x_1^{(m)} \), the Rademacher complexity satisfies
\[
|\mathcal{R}^\beta_m - \hat{\mathcal{R}}_m^\beta| \leq \frac{2B_r}{m}.
\]
Then by applying McDiarmid’s Inequality again, we obtain
\[
P(\mathcal{R}^\beta_m - \hat{\mathcal{R}}_m > \epsilon) \leq \exp\left(\frac{-m\epsilon^2}{2B_r^2}\right) + (m - 1)\beta(b).
\]
Thus with probability at least \( 1 - \frac{\delta}{2} \), we have
\[
\mathcal{R}^\beta_m - \hat{\mathcal{R}}_m \leq 2B_r \sqrt{\frac{\log \frac{2m}{(m - 1)\beta(b)}}{2m}} \leq 2B_r \sqrt{\frac{\log \frac{4}{2m}}{2m}}.
\]
Combining (13) and (14), we have with probability at least \( 1 - \delta \),
\[
\Phi(\tau) \leq 2\hat{\mathcal{R}}_m + 6B_r \sqrt{\frac{\log \frac{4}{2m}}{2m}}.
\]
We apply the Dudley’s entropy integral to bound \( \hat{\mathcal{R}}_m \). Specifically, we have
\[
\hat{\mathcal{R}}_m \leq \frac{4\alpha}{\sqrt{m}} + \frac{12}{m} \int_0^{\sqrt{m}B_r} \sqrt{\log N(\mathcal{R}, \epsilon, \| \cdot \|_\infty)} \, de
\leq \frac{4\alpha}{\sqrt{m}} + \frac{12B_r}{\sqrt{m}} \sqrt{\log N(\mathcal{R}, \alpha, \| \cdot \|_\infty)}.
\]
It suffices to pick \( \alpha = \frac{1}{\sqrt{m}} \). By combining (6), (15), and (16), we have with probability at least \( 1 - \delta \),
\[
d_\mathcal{R}(\pi^*, \hat{\pi}) - \inf_{\pi} d_\mathcal{R}(\pi^*, \pi) \leq \frac{16}{m} + \frac{48B_r}{\sqrt{m}} \sqrt{\log N(\mathcal{R}, 1/\sqrt{m}, \| \cdot \|_\infty)} + \frac{12B_r}{\sqrt{m}} \sqrt{\frac{\log \frac{4}{2m}}{2m}} + \epsilon,
\]
where \( \delta' = \delta - 4(m - 1)\beta(b) \) and \( 2bm = T \). Substituting \( m = T/2b \), we have
\[
d_\mathcal{R}(\pi^*, \hat{\pi}) - \inf_{\pi} d_\mathcal{R}(\pi^*, \pi) \leq \frac{32b}{T} + \frac{48B_r}{\sqrt{T/2b}} \sqrt{\log N(\mathcal{R}, 1/\sqrt{T/2b}, \| \cdot \|_\infty)} + \frac{12B_r}{\sqrt{T/2b}} \sqrt{\frac{\log \frac{4}{2m}}{2m}} + \epsilon.
\]
Now we instantiate a choice of \( b \) and \( m \) for exponentially \( \beta \)-mixing sequences, where the mixing coefficient \( \beta(b) \leq \beta_0 \exp(-\beta_1 b^\alpha) \) for constants \( \beta_0, \beta_1, \alpha > 0 \). We set \( \delta' > \delta - 4(m - 1)\beta(b) = \frac{\delta}{2} \).
By a simple calculation, it is enough to choose \( b = (\log(4\beta_b T/\delta))^{1/\alpha} \). Substituting such a \( b \) into (17), we have with probability at least \( 1 - \delta \),
\[
d_\mathcal{R}(\pi^*, \hat{\pi}) - \inf_{\pi} d_\mathcal{R}(\pi^*, \pi) \leq \mathcal{O}\left(\frac{B_r}{\sqrt{T/\zeta}} \sqrt{\log N(\mathcal{R}, \sqrt{T/\zeta}, \| \cdot \|_\infty)} + \frac{B_r}{\sqrt{T/\zeta}} \sqrt{\log(1/\delta)} + \epsilon, \right).
\]
where \( \zeta = \beta_1^{-1} \log \frac{\beta_b T}{\delta} \).
When \( n > 1 \), we concatenate \( n \) trajectories to form a sequence of length \( nT \), and such a sequence is still exponentially \( \beta \)-mixing. Applying the same technique, we partition the whole sequence into \( 2nm \) blocks of equal size \( b \). Then with probability at least \( 1 - \delta \), we have
\[
d_\mathcal{R}(\pi^*, \hat{\pi}) - \inf_{\pi} d_\mathcal{R}(\pi^*, \pi) \leq \mathcal{O}\left(\frac{B_r}{nT/\zeta} \sqrt{\log N(\mathcal{R}, \sqrt{nT/\zeta}, \| \cdot \|_\infty)} + \frac{B_r}{nT/\zeta} \sqrt{\log(1/\delta)} + \epsilon, \right).
\]
A.2 Proof of Corollary 1

The reward function can be bounded by

\[ |r(s, a)| = |\theta^\top g(\psi_s, \psi_a)| \leq \|\theta\|_2 g(\psi_s, \psi_a) \|_2 \leq \sqrt{2}B_\theta \rho_g, \]

where the first inequality comes from Cauchy-Schwartz inequality.

To compute the covering number, we exploit the Lipschitz continuity of \( r(s, a) \) with respect to parameter \( \theta \). Specifically, for two different parameters \( \theta \) and \( \theta' \), we have

\[ \|r(s, a) - r'(s, a)\|_\infty = \| (\theta - \theta')^\top g(\psi_s, \psi_a) \|_\infty \]

\[ \leq (i) \| \theta - \theta' \|_2 \sup_{(s, a) \in S \times A} \| g(\psi_s, \psi_a) \|_2 \]

\[ \leq \sqrt{2} \| \theta - \theta' \|_2 \rho_g \sup_{(s, a) \in S \times A} \| \psi_s \|_2 + \| \psi_a \|_2 \leq \sqrt{2} \rho_g \| \theta - \theta' \|_2, \]

where (i) comes from Cauchy-Schwartz inequality, (ii) comes from the Lipschitz continuity of \( g \), and (iii) comes from the boundedness of \( \psi_s \) and \( \psi_a \).

Denote \( \Theta = \{ \theta \in \mathbb{R}^q : \| \theta \|_2 \leq B_\theta \} \). By the standard argument of the volume ratio, we have

\[ \mathcal{N}(\Theta, \epsilon, \| \cdot \|_2) \leq \left( 1 + \frac{2B_\theta}{\epsilon} \right)^q. \]

Accordingly, we have

\[ \mathcal{N}\left(R, \sqrt{\frac{2(\beta_1^{-1} \log(4\beta_0 T/\delta))}{T}}, \| \cdot \|_\infty \right) \leq \mathcal{N}\left(\Theta, \frac{1}{\sqrt{2} \rho_g} \sqrt{\frac{2(\beta_1^{-1} \log(4\beta_0 T/\delta))}{T}} \| \cdot \|_2 \right), \]

\[ \leq \left( 1 + 2\sqrt{2} \rho_g \frac{T}{\beta_1^{-1} \log(4\beta_0 T/\delta)} \right)^q. \quad (19) \]

Plugging (19) into (18), we have

\[ d_R(\pi^*, \tilde{\pi}) = \inf_{\pi} d_R(\pi^*, \pi) \]

\[ = O\left( \frac{\rho_g B_\theta}{\sqrt{T/\zeta}} \sqrt{q \log \left( \frac{T}{\zeta} \right)} + \rho_g B_\theta \sqrt{\frac{\log(1/\delta)}{T/\zeta}} + \epsilon \right) \]

hold, with probability at least \( 1 - \delta \).

A.3 Proof of Corollary 2

We investigate the Lipschitz continuity of \( r \) with respect to the weight matrices \( W_1, \cdots, W_D \).

Specifically, given two different sets of matrices \( W_1, \cdots, W_D \) and \( W'_1, \cdots, W'_D \), we have

\[ \|r(s, a) - r'(s, a)\|_\infty \]

\[ \leq \| W_D^T \sigma(W_{D-1}\sigma(\cdots \sigma(W_1[\psi_a^T, \psi_s^T]^T)\cdots)) - (W_D')^T \sigma(W_{D-1}\sigma(\cdots \sigma(W_1[\psi_a^T, \psi_s^T]^T)\cdots)) \|_2 \]

\[ \leq \| W_D^T \sigma(W_{D-1}\sigma(\cdots \sigma(W_1[\psi_a^T, \psi_s^T]^T)\cdots)) - (W_D')^T \sigma(W_{D-1}\sigma(\cdots \sigma(W_1[\psi_a^T, \psi_s^T]^T)\cdots)) \|_2 \]

\[ + \| (W_D')^T \sigma(W_{D-1}\sigma(\cdots \sigma(W_1[\psi_a^T, \psi_s^T]^T)\cdots)) - (W_D')^T \sigma(W_{D-1}\sigma(\cdots \sigma(W_1[\psi_a^T, \psi_s^T]^T)\cdots)) \|_2 \]

\[ \leq \| W_D - W_D' \|_2 \| \sigma(W_{D-1}\sigma(\cdots \sigma(W_1[\psi_a^T, \psi_s^T]^T)\cdots)) \|_2 \]

\[ + \| W_D' \|_2 \| \sigma(W_{D-1}\sigma(\cdots \sigma(W_1[\psi_a^T, \psi_s^T]^T)\cdots)) - \sigma(W_D'\sigma(\cdots \sigma(W_1[\psi_a^T, \psi_s^T]^T)\cdots)) \|_2 \].

Note that we have

\[ \| \sigma(W_{D-1}\sigma(\cdots \sigma(W_1[\psi_a^T, \psi_s^T]^T)\cdots)) \|_2 \leq (i) \| W_D - W_D' \|_2 \| \sigma(W_{D-1}\sigma(\cdots \sigma(W_1[\psi_a^T, \psi_s^T]^T)\cdots)) \|_2 \]

\[ \leq \| W_D - W_D' \|_2 \| \sigma(\cdots \sigma(W_1[\psi_a^T, \psi_s^T]^T)\cdots) \|_2 \leq \| \psi_a^T, \psi_s^T \|_2 \| \psi_s^T \|_2 \leq \sqrt{2}, \]
where (i) comes from the definition of the ReLU activation, (ii) comes from \( \|W_i\|_2 \leq 1 \) and recursion, and (iii) comes from the boundedness of \( \psi_s \) and \( \psi_a \). Accordingly, we have
\[
\|r(s, a) - r'(s, a)\|_\infty \leq \sqrt{2} \|W_D - W_D'\|_2 + \|W_D'\|_2 \|\sigma(W_{D-1}\sigma(...)) - \sigma(W_{D-1}\sigma(...))\|_2
\]
\[
\leq \sqrt{2} \|W_D - W_D'\|_2 + \|W_{D-1}\sigma(...)) - W_{D-1}'\sigma(...))\|_2
\]
\[
\leq \sum_{i=1}^D \sqrt{2} \|W_i - W_i'\|_2 ,
\]
where (i) comes from the Lipschitz continuity of the ReLU activation, and (ii) comes from the recursion. We then derive the covering number of \( \mathcal{R} \) by the Cartesian product of the matrix covering of \( W_1, ..., W_D \):
\[
\mathcal{N}(\mathcal{R}, \epsilon, \| \cdot \|_\infty) \leq \prod_{i=1}^D \mathcal{N}(W_i, \epsilon \|\cdot\|_2) \leq \left( 1 + \frac{\sqrt{2}D\sqrt{d}}{\epsilon} \right)^{d^2D} ,
\]
where the second inequality comes from the standard argument of the volume ratio. Plugging (20) into (18), we have
\[
\begin{align*}
&d_{\mathcal{R}}(\pi^*, \tilde{\pi}) - \inf_{\pi} d_{\mathcal{R}}(\pi^*, \pi) \\
&= O \left( \frac{1}{\sqrt{T/\xi}} \right) \left[ d^2D \log \left( D \frac{dT}{\xi} \right) + \sqrt{\frac{\log(1/\delta)}{T/\xi}} \right] + \epsilon
\end{align*}
\]
hold, with probability at least \( 1 - \delta \).

\section{Proof of Theorem 2}

The proof is arranged as follows. We first prove the boundedness of \( Q \) function and characterize the Lipschitz properties of the gradients of \( F \) with respect to \( \omega \) and \( \theta \), respectively, in Section B.2. Then we provide some important lemmas in Section B.3. Using these lemmas, we prove Lemma 1 in Section B.4. We prove Theorem 2 in Section B.5. For notational simplicity, we denote \( \langle \cdot, \cdot \rangle \) as the vector inner product throughout the rest of our analysis.

\subsection{Boundedness of \( Q \) Function}

\textbf{Lemma 3.} For any \( \omega \), we have
\[
\|Q^\tilde{\omega}\|_\infty \leq B_Q ,
\]
where \( B_Q = \frac{2\sqrt{2}\kappa_0\chi}{1 - \nu} \).

\textbf{Proof.}
\[
Q^\tilde{\omega}(s, a) = \sum_{t=0}^\infty \mathbb{E}[\tilde{r}_\theta(s_t, a_t) - \mathbb{E}[\tilde{r}_\theta | s_0 = s, a_0 = a, \tilde{\pi}_\omega]]
\]
\[
= \sum_{t=0}^\infty \left[ \int_{S \times A} \tilde{r}_\theta(s, a) \rho_0(s, a)(P_{\pi_\omega})^t d(s, a) - \int_{S \times A} \tilde{r}_\theta(s, a) \rho_{\tilde{\pi}_\omega}(s, a) d(s, a) \right]
\]
\[
= \sum_{t=0}^\infty \int_{S \times A} \tilde{r}_\theta(s, a) \rho_0(s, a)(P_{\pi_\omega})^t - \rho_{\tilde{\pi}_\omega}(s, a) d(s, a)
\]
\[
\leq \sum_{t=0}^\infty 2 \|\tilde{r}_\theta\|_\infty \|\rho_0(P_{\pi_\omega})^t - \rho_{\tilde{\pi}_\omega}\|_{TV}
\]
\[
\leq 2\sqrt{2}\kappa_0 \sum_{t=0}^\infty \chi \nu^t = \frac{2\sqrt{2}\kappa_0\chi}{1 - \nu} ,
\]
where the first inequality comes from the definition of Total Variance distance of probability measures and the second inequality results from (i) of Assumption 5. \qed
\[ \| \nabla_{\omega} F(\omega, \theta) - \nabla_{\omega} F(\omega', \theta) \|_2 \leq L_{\omega} \| \omega - \omega' \|_2, \]
\[ \| \nabla_{\theta} F(\omega, \theta) - \nabla_{\theta} F(\omega, \theta') \|_2 \leq \mu \| \theta - \theta' \|_2, \]
where \( L_{\omega} = \frac{2\sqrt{S_{\omega} + 2B_{\omega}L_{\rho}}}{1 - c} + B_{\omega}L_{Q}. \)

**Proof.** By the Policy Gradient Theorem (Sutton et al., 2000), we have
\[ \nabla_{\omega} F(\omega, \theta) = \mathbb{E}_{\pi_{\omega}} \nabla \log(\tilde{\pi}_\omega(a | s))Q^\omega(s, a). \]

Therefore,
\[ \| \nabla_{\omega} F(\omega, \theta) - \nabla_{\omega} F(\omega', \theta) \|_2 = \left\| \mathbb{E}_{\pi_{\omega}} \nabla \log(\tilde{\pi}_\omega(a | s))Q^\omega(s, a) - \mathbb{E}_{\pi_{\omega'}} \nabla \log(\tilde{\pi}_{\omega'}(a | s))Q^{\omega'}(s, a) \right\|_2 \]
\[ \leq \left\| \mathbb{E}_{\pi_{\omega}} \nabla \log(\tilde{\pi}_\omega(a | s))Q^\omega(s, a) - \mathbb{E}_{\pi_{\omega'}} \nabla \log(\tilde{\pi}_{\omega'}(a | s))Q^{\omega'}(s, a) \right\|_2 \]
\[ + \left\| \mathbb{E}_{\pi_{\omega}} \nabla \log(\tilde{\pi}_\omega(a | s))Q^\omega(s, a) - \mathbb{E}_{\pi_{\omega'}} \nabla \log(\tilde{\pi}_{\omega'}(a | s))Q^{\omega'}(s, a) \right\|_2 \]
\[ \leq (S_{\omega}B_{Q} + B_{\omega}L_{Q}) \| \omega - \omega' \|_2 + 2B_{\omega}B_{Q} \| \rho_{\omega} - \rho_{\omega'} \|_{TV} \]
\[ \leq (S_{\omega}B_{Q} + B_{\omega}L_{Q} + 2B_{\omega}B_{Q}L_{\rho}) \| \omega - \omega' \|_2, \tag{21} \]
where the second and the third inequality results from (ii) of Assumption 5. Plugging \( B_{Q} = \frac{2\sqrt{2\pi_{\omega}X}}{1 - c} \) into (21) yields the desired result.

Similarly, we have
\[ \| \nabla_{\theta} F(\omega, \theta) - \nabla_{\theta} F(\omega, \theta') \|_2 \leq \| -\mu(\theta - \theta') \|_2 \leq \mu \| \theta - \theta' \|_2. \]

We then characterize the Lipschitz continuity of \( \nabla_{\theta} F \) with respect to \( \omega. \)

**Lemma 5.** Suppose Assumptions 1, 3 and 5 hold. For any \( \omega, \omega' \) and \( \theta \), we have
\[ \| \nabla_{\theta} F(\omega, \theta) - \nabla_{\theta} F(\omega', \theta) \|_2 \leq S_{\omega} \| \omega - \omega' \|_2, \]
where \( S_{\omega} = \frac{2\sqrt{2\pi_{\omega}X}}{1 - c}B_{\omega}. \)

**Proof.** We have
\[ \nabla_{\theta} F(\omega, \theta) = -\mu\theta - \nabla_{\theta} \left[ \mathbb{E}_{\rho_{\omega}} \left[ \theta^T g(\psi_{s_t}, \psi_{a_t}) \right] - \mathbb{E}_{\rho_{\omega'}} \left[ \theta^T g(\psi_{s_t}, \psi_{a_t}) \right] \right]. \]

Therefore, we have
\[ \| \nabla_{\theta} F(\omega, \theta) - \nabla_{\theta} F(\omega', \theta) \|_2 \]
\[ = \left\| \mathbb{E}_{\rho_{\omega}} \left[ g(\psi_{s_t}, \psi_{a_t}) \right] - \mathbb{E}_{\rho_{\omega'}} \left[ g(\psi_{s_t}, \psi_{a_t}) \right] \right\|_2. \]

Suppose \( j^* = \arg\max_{1 \leq j \leq q} \mathbb{E}_{\rho_{\omega}} g(\psi_{s_t}, \psi_{a_t})_j - \mathbb{E}_{\rho_{\omega'}} g(\psi_{s_t}, \psi_{a_t})_j \), by Mean Value Theorem, there exists vector \( \zeta \), which is some interpolation between vectors \( \omega \) and \( \omega' \), such that
\[ \mathbb{E}_{\rho_{\omega}} g(\psi_{s_t}, \psi_{a_t})_j - \mathbb{E}_{\rho_{\omega'}} g(\psi_{s_t}, \psi_{a_t})_j = (\nabla_{\omega} \mathbb{E}_{\rho_{\omega}} g(\psi_{s_t}, \psi_{a_t})_j, \omega - \omega'). \tag{23} \]

By Policy Gradient Theorem, we have
\[ \left\| \nabla_{\omega} \mathbb{E}_{\rho_{\omega}} g(\psi_{s_t}, \psi_{a_t})_j \right\|_2 = \left\| \mathbb{E}_{\rho_{\omega}} \nabla \log(\pi_{\omega}(a | s))Q^\omega_{\omega}(s, a) \right\|_2 \]
\[ \leq \sup_{(s, a) \in S \times \mathcal{A}} | \nabla \log(\pi_{\omega}(a | s))Q^\omega_{\omega}(s, a) | \]
\[ \leq B_{Q}B_{\omega}, \tag{24} \]
where \( Q^\omega_{\omega}(s, a) = \sum_{t=0}^{\infty} \mathbb{E}[g(s_t, a_t)_j - \mathbb{E}_{\pi_{\omega}} g(s_t, a_t)_j | s_0 = s, a_0 = a, \pi_{\omega}] \). Combining (22), (23), (24) and using Cauchy-Schwartz Inequality, we prove the lemma. \( \square \)
B.3 Some Important Lemmas for Proving Lemma 1

We denote
\[ \xi_t^{\omega(t)} = \nabla_\theta F(\omega(t), \theta(t)) - \frac{1}{q_\theta} \sum_{j \in \mathcal{M}_t} \nabla_\theta f_j(\omega(t), \theta(t)) \]
and
\[ \xi_t^{\theta(t)} = \nabla_\omega F(\omega(t), \theta(t+1)) - \frac{1}{q_\omega} \sum_{j \in \mathcal{M}_t} \nabla_\omega f_j(\omega(t), \theta(t+1)) \]
as the i.i.d. noise of the stochastic gradient, respectively. Throughout the rest of the analysis, the expectation \( \mathbb{E} \) is taken with respect to all the noise in each iteration of the alternating mini-batch stochastic gradient descent algorithm. The next lemma characterizes the progress at the \((t + 1)\)-th iteration.

For notational simplicity, we define vector function
\[ G(\pi) = \mathbb{E}_{\mu, g(\psi), \psi} \]

Lemma 6. At the \((t + 1)\)-th iteration, we have
\[
\mathbb{E} F(\omega(t+1), \theta(t+1)) - \mathbb{E} F(\omega(t), \theta(t)) \\
\leq \left( L_\omega - \frac{1}{\eta_\omega} \right) \mathbb{E} \|\omega(t+1) - \omega(t)\|_2^2 + \frac{S_\omega}{2} \cdot \mathbb{E} \|\omega(t) - \omega(t-1)\|_2^2 \\
+ \left( \frac{1}{2\eta_\omega} + \frac{S_\omega}{2} + \mu \right) \mathbb{E} \|\theta(t+1) - \theta(t)\|_2^2 + \left( \frac{1}{2\eta_\omega} + \frac{L_\theta}{2} \right) \mathbb{E} \|\theta(t) - \theta(t-1)\|_2^2 \\
+ \eta_\omega \mathbb{E} \|\xi_t^{\omega(t)}\|_2^2 + \frac{1}{2\mu} \mathbb{E} \|\xi_t^{\theta(t)}\|_2^2.
\]

Proof. We have
\[
\mathbb{E} F(\omega(t+1), \theta(t+1)) - \mathbb{E} F(\omega(t), \theta(t)) \\
= \mathbb{E} F(\omega(t+1), \theta(t+1)) - \mathbb{E} F(\omega(t), \theta(t+1)) + \mathbb{E} F(\omega(t), \theta(t+1)) - \mathbb{E} F(\omega(t), \theta(t)).
\]
By the mean value theorem, we have
\[
\langle \nabla_\omega F(\tilde{\omega}(t), \theta(t+1)), \omega(t+1) - \omega(t) \rangle = F(\omega(t+1), \theta(t+1)) - F(\omega(t), \theta(t+1)),
\]
where \( \tilde{\omega}(t) \) is some interpolation between \( \omega(t+1) \) and \( \omega(t) \). Then we have
\[
\mathbb{E} F(\omega(t+1), \theta(t+1)) - \mathbb{E} F(\omega(t), \theta(t+1)) \\
= \mathbb{E} \langle \nabla_\omega F(\tilde{\omega}(t), \theta(t+1)) - \nabla_\omega F(\omega(t), \theta(t+1)), \omega(t+1) - \omega(t) \rangle + \mathbb{E} \langle \nabla_\omega F(\omega(t), \theta(t+1)), \omega(t+1) - \omega(t) \rangle. \tag{26}
\]
By Cauchy-Schwartz inequality, we have
\[
\mathbb{E} \|\nabla_\omega F(\tilde{\omega}(t), \theta(t+1)) - \nabla_\omega F(\omega(t), \theta(t+1)), \omega(t+1) - \omega(t) \|_2 \\
\leq \mathbb{E} \|\nabla_\omega F(\tilde{\omega}(t), \theta(t+1)) - \nabla_\omega F(\omega(t), \theta(t+1))\|_2 \|\omega(t+1) - \omega(t)\|_2 \\
\leq L_\omega \mathbb{E} \|\omega(t+1) - \omega(t)\|_2^2,
\]
where the last inequality comes from Lemma 5. Moreover, (4) implies
\[ \omega(t+1) - \omega(t) = -\eta_\omega \langle \nabla_\omega F(\omega(t), \theta(t+1)) + \xi_t^{\omega(t)} \rangle. \]
Therefore, we have
\[
\mathbb{E} \langle \nabla_\omega F(\omega(t), \theta(t+1)), \omega(t+1) - \omega(t) \rangle = -\frac{1}{\eta_\omega} \mathbb{E} \|\omega(t+1) - \omega(t)\|_2^2 - \mathbb{E} \langle \xi_t^{\omega(t)}, \omega(t+1) - \omega(t) \rangle \\
= -\mathbb{E} \langle \xi_t^{\omega(t)}, -\eta_\omega \langle \nabla_\omega F(\omega(t), \theta(t+1)) + \xi_t^{\omega(t)} \rangle \rangle \\
- \frac{1}{\eta_\omega} \mathbb{E} \|\omega(t+1) - \omega(t)\|_2^2 \\
= -\frac{1}{\eta_\omega} \mathbb{E} \|\omega(t+1) - \omega(t)\|_2^2 + \eta_\omega \mathbb{E} \|\xi_t^{\omega(t)}\|_2^2.
\]
Thus, we have
\[
\mathbb{E}F(\omega^{(t+1)}, \theta^{(t+1)}) - \mathbb{E}F(\omega^{(t)}, \theta^{(t+1)}) \leq (L_\omega - \frac{1}{\eta_\omega})\mathbb{E}\|\omega^{(t+1)} - \omega^{(t)}\|^2 + \eta_\omega \mathbb{E} \|\xi_\omega^{(t)}\|^2. \tag{27}
\]

By (3), the increment of $F(\omega, \theta)$ takes the form
\[
F(\omega^{(t)}, \theta^{(t+1)}) - F(\omega^{(t)}, \theta^{(t)}) = \left\langle \bar{G}(\pi^{(t)}), \theta^{(t+1)} - \theta^{(t)} \right\rangle - \frac{\mu}{2} (\|\theta^{(t+1)}\|^2 - \|\theta^{(t)}\|^2)
\]
\[
\leq \left\langle \bar{G}(\pi^{(t)}), \theta^{(t+1)} - \mu \theta^{(t)}, \theta^{(t+1)} - \theta^{(t)} \right\rangle. \tag{28}
\]

For notational simplicity, we define
\[
\epsilon^{(t+1)} = \theta^{(t+1)} - \theta^{(t)} - \left\langle \nabla_\theta F(\omega^{(t)}, \theta^{(t)}), \xi_\theta^{(t)} \right\rangle. \tag{29}
\]

Note that we have
\[
\nabla_\theta F(\omega^{(t)}, \theta^{(t)}) = G(\pi^{(t)}) - G(\pi^*) - \mu \theta^{(t)}. \tag{30}
\]

Plugging (29) and (30) into (28), we obtain
\[
F(\omega^{(t)}, \theta^{(t+1)}) - F(\omega^{(t)}, \theta^{(t)}) \leq \left\langle \frac{\theta^{(t+1)} - \theta^{(t)} - \epsilon^{(t+1)} - \epsilon^{(t)}}{\eta_\theta} - \xi_\theta^{(t)}, \theta^{(t+1)} - \theta^{(t)} \right\rangle.
\]

Since $\theta$ belongs to the convex set $\{\theta \mid \|\theta\|_2 \leq \kappa\}$, we have
\[
\langle \epsilon^{(t)}, \theta^{(t+1)} - \theta^{(t)} \rangle \geq 0.
\]

Then we obtain
\[
F(\omega^{(t)}, \theta^{(t+1)}) - F(\omega^{(t)}, \theta^{(t)}) \leq \left\langle \frac{\theta^{(t+1)} - \theta^{(t)} - \epsilon^{(t+1)} - \epsilon^{(t)}}{\eta_\theta} - \xi_\theta^{(t)}, \theta^{(t+1)} - \theta^{(t)} \right\rangle
\]
\[
= \frac{1}{\eta_\theta} \epsilon^{(t+1)} - \epsilon^{(t)} + \frac{1}{\eta_\theta} \left\|\theta^{(t+1)} - \theta^{(t)}\right\|^2 - \left\langle \xi_\theta^{(t)}, \theta^{(t+1)} - \theta^{(t)} \right\rangle. \tag{31}
\]

By the definition of $\epsilon^{(t)}$ in (29), we have
\[
\epsilon^{(t+1)} = \theta^{(t+1)} - \left\langle \nabla_\theta F(\omega^{(t)}, \theta^{(t)}), \xi_\theta^{(t)} \right\rangle \tag{32}
\]

and
\[
\epsilon^{(t)} = \theta^{(t)} - \left\langle \nabla_\theta F(\omega^{(t-1)}, \theta^{(t-1)}), \xi_\theta^{(t-1)} \right\rangle. \tag{33}
\]

Subtracting (33) from (32), we obtain
\[
\epsilon^{(t)} - \epsilon^{(t+1)} = \left(\theta^{(t)} - \theta^{(t+1)}\right) - \left(\theta^{(t-1)} - \theta^{(t)}\right) - \eta_\theta \left\langle \nabla_\theta F(\omega^{(t-1)}, \theta^{(t-1)}), \nabla_\theta F(\omega^{(t)}, \theta^{(t)}) \right\rangle - \eta_\theta \xi_\theta^{(t-1)} - \xi_\theta^{(t)}. \tag{34}
\]

Plugging (34) into the first term on the right hand side of (31), we obtain
\[
\frac{1}{\eta_\theta} \left(\epsilon^{(t)} - \epsilon^{(t+1)}\right), \theta^{(t+1)} - \theta^{(t)} \right\rangle
\]
\[
= \frac{1}{\eta_\theta} \left(\theta^{(t)} - \theta^{(t-1)}, \theta^{(t+1)} - \theta^{(t)}\right) + \left\langle \nabla_\theta F(\omega^{(t)}, \theta^{(t)}), \nabla_\theta F(\omega^{(t-1)}, \theta^{(t-1)}), \theta^{(t+1)} - \theta^{(t)} \right\rangle
\]
\[
- \frac{1}{\eta_\theta} \left\|\theta^{(t+1)} - \theta^{(t)}\right\|^2 - \left\langle \xi_\theta^{(t)} - \theta^{(t)} \right\rangle. \tag{35}
\]

For term (A), we apply the Cauchy-Schwarz inequality to obtain an upper bound as follows.
\[
\frac{1}{\eta_\theta} \left(\theta^{(t)} - \theta^{(t-1)}, \theta^{(t+1)} - \theta^{(t)}\right) \leq \frac{1}{\eta_\theta} \left\|\theta^{(t)} - \theta^{(t-1)}\right\|_2 \left\|\theta^{(t+1)} - \theta^{(t)}\right\|_2
\]
\[
\leq \frac{1}{2\eta_\theta} \left\|\theta^{(t)} - \theta^{(t-1)}\right\|_2^2 + \frac{1}{2\eta_\theta} \left\|\theta^{(t+1)} - \theta^{(t)}\right\|_2^2. \tag{36}
\]
To derive the upper bound of (B), we apply Lemma 5 to obtain
\[
\langle \nabla_\theta F(\omega(t), \theta(t)) - \nabla_\theta F(\omega(t), \theta(t+1) - \theta(t)) \rangle \\
= \langle \nabla_\theta F(\omega(t), \theta(t)) - \nabla_\theta F(\omega(t), \theta(t), \theta(t+1) - \theta(t)) \rangle \\
+ \langle \nabla_\theta F(\omega(t-1), \theta(t)) - \nabla_\theta F(\omega(t-1), \theta(t+1), \theta(t+1) - \theta(t)) \rangle \\
\leq S_\omega \|\omega(t) - \omega(t-1)\|_2 \cdot \|\theta(t) - \theta(t-1)\|_2 + \mu \cdot \|\theta(t) - \theta(t-1)\|_2 \cdot \|\theta(t+1) - \theta(t)\|_2 \\
\leq \frac{S_\omega}{2} \|\omega(t) - \omega(t-1)\|_2^2 + \mu \|\theta(t) - \theta(t-1)\|_2^2 \\
+ \frac{\mu}{2} \|\theta(t) - \theta(t-1)\|_2^2 + \frac{\mu}{2} \|\theta(t+1) - \theta(t)\|_2^2.
\]

Plugging (36) and (37) into (35), we obtain
\[
\frac{1}{\eta_\theta} (\epsilon(t) - \epsilon(t+1), \theta(t+1) - \theta(t)) \leq \left( \frac{1}{2\eta_\theta} + \frac{\mu}{2} + \frac{S_\omega}{2} \right) \|\theta(t+1) - \theta(t)\|_2^2 \\
+ \left( \frac{1}{2\eta_\theta} + \frac{\mu}{2} \right) \|\theta(t) - \theta(t-1)\|_2^2 \\
+ \frac{S_\omega}{2} \|\omega(t) - \omega(t-1)\|_2^2 + \langle \xi_\theta^{(t)} - \xi_\theta^{(t-1)}, \theta(t+1) - \theta(t) \rangle. \tag{38}
\]

Further plugging (38) into (31), we obtain
\[
\begin{aligned}
F(\omega(t), \theta(t+1)) - F(\omega(t), \theta(t)) \\
\leq \left( \frac{1}{2\eta_\theta} + \frac{S_\omega}{2} + \frac{\mu}{2} \right) \|\theta(t+1) - \theta(t)\|_2^2 + \left( \frac{1}{2\eta_\theta} + \frac{\mu}{2} \right) \|\theta(t) - \theta(t-1)\|_2^2 \\
+ \frac{S_\omega}{2} \|\omega(t) - \omega(t-1)\|_2^2 - \langle \xi_\theta^{(t-1)}, \theta(t+1) - \theta(t) \rangle \\
\leq \left( \frac{1}{2\eta_\theta} + \frac{S_\omega}{2} + \mu \right) \|\theta(t+1) - \theta(t)\|_2^2 + \left( \frac{1}{2\eta_\theta} + \frac{\mu}{2} \right) \|\theta(t) - \theta(t-1)\|_2^2 \\
+ \frac{S_\omega}{2} \|\omega(t) - \omega(t-1)\|_2^2 + \|\xi_\theta^{(t-1)}\|_2^2. \tag{39}
\end{aligned}
\]

Finally, taking expectation of (39) with respect to the noise and together with (27), we prove the final result.
\[
\begin{aligned}
EF(\omega(t+1), \theta(t+1)) - EF(\omega(t), \theta(t)) \\
\leq \left( L_\omega - \frac{1}{\eta_\omega} \right) E \|\omega(t+1) - \omega(t)\|_2^2 + \frac{S_\omega}{2} E \|\omega(t) - \omega(t-1)\|_2^2 \\
+ \left( \frac{1}{2\eta_\theta} + \frac{S_\omega}{2} + \mu \right) E \|\theta(t+1) - \theta(t)\|_2^2 \\
+ \left( \frac{1}{2\eta_\theta} + \frac{\mu}{2} \right) E \|\theta(t) - \theta(t-1)\|_2^2 + \eta_\omega E \|\xi_\omega^{(t)}\|_2^2 + \frac{1}{2\mu} E \|\xi_\theta^{(t-1)}\|_2^2.
\end{aligned}
\]

We then characterize the update of \( \omega \).

**Lemma 7.** The update of \( \omega \) satisfies
\[
\begin{aligned}
E \left( \omega(t+1) - \omega(t) - (\omega(t) - \omega(t-1)), \omega(t+1) - \omega(t) \right) \\
\leq -\frac{\eta_\omega}{\eta_\theta} \cdot E \left( (\theta(t+2) - \theta(t+1)) - (\theta(t+1) - \theta(t)) - (\epsilon(t+2) - \epsilon(t+1)) \right) \\
- \frac{\mu\eta_\omega}{2} \cdot E \|\theta(t+1) - \theta(t)\|_2^2 + \frac{\eta_\omega(5L_\omega + 1)}{2} \cdot E \|\omega(t+1) - \omega(t)\|_2^2 + \frac{\eta_\omega L_\omega}{2} \cdot E \|\omega(t) - \omega(t-1)\|_2^2 \\
+ \left( \frac{\eta_\omega}{2\mu} + \eta_\omega^2 \right) E \|\xi_\omega^{(t)}\|_2^2 + \frac{\eta_\omega^2}{2} E \|\xi_\theta^{(t-1)}\|_2^2.
\end{aligned}
\]
Proof. By the update policy, we have

\[
E \left< \omega^{(t+1)} - \omega^{(t)} - (\omega^{(t)} - \omega^{(t-1)}), \omega^{(t+1)} - \omega^{(t)} \right>
\]

\[
= - \eta_w E \left< \nabla_\omega F(\omega^{(t)}, \theta^{(t+1)}) + \xi^{(t)} - \nabla_\omega F(\omega^{(t-1)}, \theta^{(t)}) - \xi^{(t-1)}, \omega^{(t+1)} - \omega^{(t)} \right>
\]

\[
= - \eta_w E \left< \nabla_\omega F(\omega^{(t)}, \theta^{(t+1)}) - \nabla_\omega F(\omega^{(t)}, \theta^{(t)}), \omega^{(t+1)} - \omega^{(t)} \right>
\]

\[
- \eta_w E \left< \nabla_\omega F(\omega^{(t)}, \theta^{(t)}) - \nabla_\omega F(\omega^{(t-1)}, \theta^{(t)}), \omega^{(t+1)} - \omega^{(t)} \right>
\]

\[
- \eta_w E \left< \xi^{(t)}, -\eta_w (\nabla_\omega F(\omega^{(t)}, \theta^{(t+1)}) + \xi^{(t)}) \right> + \eta_w E \left< \xi^{(t-1)}, \omega^{(t+1)} - \omega^{(t)} \right>
\]

\[
\leq - \eta_w E \left< \nabla_\omega F(\omega^{(t)}, \theta^{(t+1)}) - \nabla_\omega F(\omega^{(t)}, \theta^{(t)}), \omega^{(t+1)} - \omega^{(t)} \right>
\]

\[
- \eta_w E \left< \nabla_\omega F(\omega^{(t)}, \theta^{(t)}) - \nabla_\omega F(\omega^{(t-1)}, \theta^{(t)}), \omega^{(t+1)} - \omega^{(t)} \right>
\]

\[
+ \eta_w^2 \left\| \xi^{(t)} \right\|_2^2 + \frac{\eta_w}{2} \left\| \xi^{(t-1)} \right\|_2^2 + \frac{\eta_w}{2} \left\| \omega^{(t+1)} - \omega^{(t)} \right\|_2^2.
\]  

(40)

Then for (C), by the definition of objective function, we have

\[
(C) = - \eta_w E \left< \nabla_\omega \sum_j G(\bar{\pi}_{\omega(t)})_j (\theta^{(t+1)} - \theta^{(t)})_j, \omega^{(t+1)} - \omega^{(t)} \right>
\]

\[
= - \eta_w E \sum_j \left< \nabla_\omega G(\bar{\pi}_{\omega(t)})_j - \nabla_\omega G(\bar{\pi}_{\omega^{(t+1)}})_j + \nabla_\omega G(\bar{\pi}_{\omega^{(t+1)}})_j, \omega^{(t+1)} - \omega^{(t)} \right>_j.
\]

\[
(\theta^{(t+1)} - \theta^{(t)})_j
\]

\[
= - \eta_w E \sum_j (\theta^{(t+1)} - \theta^{(t)})_j \left< \nabla_\omega G(\bar{\pi}_{\omega(t)})_j - \nabla_\omega G(\bar{\pi}_{\omega^{(t+1)}})_j, \omega^{(t+1)} - \omega^{(t)} \right>_j.
\]

\[
- \eta_w E \sum_j (\theta^{(t+1)} - \theta^{(t)})_j \left< \nabla_\omega G(\bar{\pi}_{\omega^{(t+1)}})_j, \omega^{(t+1)} - \omega^{(t)} \right>_j,
\]  

(41)

where for every \( j \), \( \bar{\pi}_{\omega^{(t)}}_j \) is some interpolation between vectors \( \bar{\pi}_{\omega(t)} \) and \( \bar{\pi}_{\omega^{(t+1)}} \) such that

\[
\left< \nabla_\omega G(\bar{\pi}_{\omega(t)})_j, \omega^{(t+1)} - \omega^{(t)} \right> = G(\bar{\pi}_{\omega(t+1)})_j - G(\bar{\pi}_{\omega(t)})_j.
\]  

(42)

For the first term in the right hand side of (41), by Lemma 4, we have

\[
- \eta_w E \sum_j (\theta^{(t+1)} - \theta^{(t)})_j \left< \nabla_\omega G(\bar{\pi}_{\omega(t)})_j - \nabla_\omega G(\bar{\pi}_{\omega^{(t+1)}})_j, \omega^{(t+1)} - \omega^{(t)} \right>_j
\]

\[
\leq 2 \eta_w L_\omega E \left\| \omega^{(t+1)} - \omega^{(t)} \right\|_2^2.
\]

(43)

For the second term in the right hand side of (41), by (42) we have

\[
- \eta_w E \sum_j (\theta^{(t+1)} - \theta^{(t)})_j \left< \nabla_\omega G(\bar{\pi}_{\omega^{(t+1)}})_j, \omega^{(t+1)} - \omega^{(t)} \right>_j
\]

\[
= - \eta_w E \left< G(\bar{\pi}_{\omega^{(t+1)}}) - G(\bar{\pi}_{\omega^{(t)}}), \theta^{(t+1)} - \theta^{(t)} \right>.
\]
Then by the definition of objective function, we have
\[
- \eta_\omega \mathbb{E} \sum_j (\theta^{(t+1)} - \theta^{(t)}) \cdot \langle \nabla_\omega G(\Pi_{\omega^{(t)}}) \rangle_j, \omega^{(t+1)} - \omega^{(t)} \rangle
\]
\[
= - \eta_\omega \mathbb{E} \left( \frac{1}{\eta_\theta} \left( \theta^{(t+2)} - \theta^{(t+1)} - \epsilon^{(t+2)} \right) - \xi^{(t+1)}_\theta + \mu \theta^{(t+1)} \right)
- \frac{1}{\eta_\theta} \left( \theta^{(t+1)} - \theta^{(t)} - \epsilon^{(t+1)} \right) + \xi^{(t)}_\theta - \mu \theta^{(t)}, \theta^{(t+1)} - \theta^{(t)} \rangle
\]
\[
= - \frac{\eta_\omega}{\eta_\theta} \mathbb{E} \left( \left( \theta^{(t+2)} - \theta^{(t+1)} \right) - \left( \theta^{(t+1)} - \theta^{(t)} \right) - \left( \epsilon^{(t+2)} - \epsilon^{(t+1)} \right), \theta^{(t+1)} - \theta^{(t)} \right)
+ \eta_\omega \mathbb{E} \left( \xi^{(t+1)}_\theta - \xi^{(t)}_\theta, \theta^{(t+1)} - \theta^{(t)} \right) - \mu \eta_\omega \mathbb{E} \left\| \theta^{(t+1)} - \theta^{(t)} \right\|_2^2
\]
\[
\leq - \frac{\eta_\omega}{\eta_\theta} \mathbb{E} \left( \left( \theta^{(t+2)} - \theta^{(t+1)} \right) - \left( \theta^{(t+1)} - \theta^{(t)} \right) - \left( \epsilon^{(t+2)} - \epsilon^{(t+1)} \right), \theta^{(t+1)} - \theta^{(t)} \right)
- \frac{\mu \eta_\omega}{2} \mathbb{E} \left\| \theta^{(t+1)} - \theta^{(t)} \right\|_2^2 + \frac{\eta_\omega}{2 \mu} \mathbb{E} \left\| \xi^{(t)}_\theta \right\|_2^2. \tag{44}
\]
For (D), applying Cauchy-Schwarz inequality we have
\[
- \eta_\omega \mathbb{E} \langle \nabla_\omega F(\omega^{(t)}), \theta^{(t)} \rangle - \nabla_\omega F(\omega^{(t-1)}, \theta^{(t)}), \omega^{(t+1)} - \omega^{(t)} \rangle
\]
\[
\leq \eta_\omega L_\omega \mathbb{E} \left\| \omega^{(t+1)} - \omega^{(t)} \right\|_2 \left\| \omega^{(t)} - \omega^{(t-1)} \right\|_2
\]
\[
\leq \frac{\eta_\omega L_\omega}{2} \mathbb{E} \left\| \omega^{(t+1)} - \omega^{(t)} \right\|_2^2 + \frac{\eta_\omega L_\omega}{2} \mathbb{E} \left\| \omega^{(t)} - \omega^{(t-1)} \right\|_2^2. \tag{45}
\]
Finally, combining (40)-(45), we prove Lemma 7.
\[
\square
\]
For notational simplicity, we define
\[
\delta^{(t+2)} = (\theta^{(t+2)} - \theta^{(t+1)}) - (\theta^{(t+1)} - \theta^{(t)}). \tag{46}
\]
Lemma 8. The first term on the right hand side of Lemma 7 satisfies
\[
- \frac{\eta_\omega}{\eta_\theta} \mathbb{E} \left( \theta^{(t+2)} - \theta^{(t+1)}, \theta^{(t+1)} - \theta^{(t)} - \epsilon^{(t+2)} - \epsilon^{(t+1)}, \theta^{(t+1)} - \theta^{(t)} \right) \tag{47}
\]
\[
\leq - \frac{\eta_\omega}{2 \eta_\theta} \mathbb{E} \left\| \theta^{(t+2)} - \theta^{(t+1)} \right\|_2^2 + \frac{3 \mu^2 \eta_\omega \eta_\theta}{2} + \frac{\eta_\omega}{2 \eta_\theta} \mathbb{E} \left\| \theta^{(t+1)} - \theta^{(t)} \right\|_2^2
\]
\[
+ \frac{3 \mu \eta_\omega \eta_\theta^2}{2} \mathbb{E} \left\| \omega^{(t+1)} - \omega^{(t)} \right\|_2^2 + \frac{3 \eta_\omega \eta_\theta}{2} \mathbb{E} \left\| \xi^{(t+1)}_\theta \right\|_2^2 + \mathbb{E} \left\| \xi^{(t)}_\theta \right\|_2^2. \tag{48}
\]
Proof. Plugging (46) into (47), we obtain
\[
- \frac{\eta_\omega}{\eta_\theta} \mathbb{E} \left( \theta^{(t+2)} - \theta^{(t+1)}, \theta^{(t+1)} - \theta^{(t)} - \epsilon^{(t+2)} - \epsilon^{(t+1)}, \theta^{(t+1)} - \theta^{(t)} \right)
= \frac{\eta_\omega}{\eta_\theta} \mathbb{E} \left( \delta^{(t+2)} - \delta^{(t+1)}, \delta^{(t+2)} - \delta^{(t+1)} \right)
\]
\[
= \frac{\eta_\omega}{\eta_\theta} \mathbb{E} \left( \delta^{(t+2)} - \delta^{(t+1)}, \delta^{(t+2)} - \delta^{(t+1)} \right) - \frac{\eta_\omega}{\eta_\theta} \mathbb{E} \left( \delta^{(t+2)}, \theta^{(t+2)} - \theta^{(t+1)} \right)
+ \frac{\eta_\omega}{\eta_\theta} \mathbb{E} \left( \delta^{(t+2)} - \delta^{(t+1)}, \theta^{(t+2)} - \theta^{(t+1)} \right). \tag{49}
\]
By applying the equality
\[
\langle u, v \rangle = \frac{1}{2} (\| u \|^2 + \| v \|^2 - \| u - v \|^2)
\]

23
to the first two terms on the right hand side of (49), we obtain

\[
- \frac{\eta_o}{\eta} \mathbb{E} \left( \left( \theta^{(t+2)} - \theta^{(t+1)} \right) - \left( \theta^{(t+1)} - \theta^{(t)} \right) - \left( \epsilon^{(t+2)} - \epsilon^{(t+1)} \right), \theta^{(t+1)} - \theta^{(t)} \right)
\]

\[
= \frac{\eta_o}{2\eta} \mathbb{E} \left( \left\| \delta^{(t+2)} - \left( \epsilon^{(t+2)} - \epsilon^{(t+1)} \right) \right\|^2_2 + \left\| \delta^{(t+2)} - \theta^{(t+1)} \right\|^2_2 - \left\| \epsilon^{(t+2)} - \epsilon^{(t+1)} \right\|^2_2 \right)
\]

\[
- \frac{\eta_o}{2\eta} \mathbb{E} \left( \left\| \delta^{(t+2)} \right\|^2_2 + \left\| \theta^{(t+2)} - \theta^{(t+1)} \right\|^2_2 - \left\| \theta^{(t+1)} - \theta^{(t)} \right\|^2_2 \right)
\]

\[
+ \frac{\eta_o}{\eta} \mathbb{E} \left( \epsilon^{(t+2)} - \epsilon^{(t+1)}, \theta^{(t+2)} - \theta^{(t+1)} \right).
\]

Recall that

\[
\theta^{(t+2)} = \Pi_{\alpha} \left( \theta^{(t+1)} + \eta_o \left( \nabla_\theta F(\omega^{(t+1)} \theta^{(t+1)}) + \xi^{(t+1)}_\theta \right) \right),
\]

and

\[
\theta^{(t+1)} = \Pi_{\alpha} \left( \theta^{(t)} + \eta_o \left( \nabla_\theta F(\omega^{(t)} \theta^{(t)}) + \xi^{(t)}_\theta \right) \right).
\]

Following from the convexity of \( \{ \theta \| \theta \|_2 \leq \kappa \} \), we have

\[
\langle \epsilon^{(t+2)}, \theta^{(t+2)} - \theta^{(t+1)} \rangle \leq 0 \quad \text{and} \quad \langle \epsilon^{(t+1)}, \theta^{(t+2)} - \theta^{(t+1)} \rangle \geq 0.
\]

Thus, the last term on the right side of (50) is negative. By rearranging the terms in (50), we obtain

\[
- \frac{\eta_o}{\eta} \mathbb{E} \left( \left( \theta^{(t+2)} - \theta^{(t+1)} \right) - \left( \theta^{(t+1)} - \theta^{(t)} \right) - \left( \epsilon^{(t+2)} - \epsilon^{(t+1)} \right), \theta^{(t+1)} - \theta^{(t)} \right)
\]

\[
\leq \frac{\eta_o}{2\eta} \mathbb{E} \left( \left\| \delta^{(t+2)} - \left( \epsilon^{(t+2)} - \epsilon^{(t+1)} \right) \right\|^2_2 - \frac{\eta_o}{2\eta} \mathbb{E} \left\| \theta^{(t+2)} - \theta^{(t+1)} \right\|^2_2 + \frac{\eta_o}{2\eta} \mathbb{E} \left\| \theta^{(t+1)} - \theta^{(t)} \right\|^2_2 \right).
\]

By definition of \( \delta^{(t+2)} \) in (46), we have

\[
\delta^{(t+2)} - \left( \epsilon^{(t+2)} - \epsilon^{(t+1)} \right) = \left( \theta^{(t+2)} - \theta^{(t+1)} - \epsilon^{(t+2)} \right) - \left( \theta^{(t+1)} - \theta^{(t)} - \epsilon^{(t+1)} \right)
\]

\[
= \eta_o [G(\pi_{\omega^{(t+1)}}) - G(\pi_{\omega^{(t)}}) - \mu (\theta^{(t+1)} - \theta^{(t)}) + \xi^{(t+1)}_\theta - \xi^{(t)}_\theta].
\]

Using the Cauchy-Schwarz inequality, we obtain

\[
\frac{\eta_o}{2\eta} \mathbb{E} \left( \left\| \delta^{(t+2)} - \left( \epsilon^{(t+2)} - \epsilon^{(t+1)} \right) \right\|^2_2
\]

\[
\leq \frac{3\eta_o \eta_o S^2_o}{2} \mathbb{E} \left( \left\| \epsilon^{(t+1)} - \omega^{(t)} \right\|^2_2 + \frac{3}{2} \mu^2 \eta_o \mathbb{E} \left\| \epsilon^{(t+1)} - \epsilon^{(t)} \right\|^2_2 \right)
\]

\[
\leq \frac{3\eta_o \eta_o S^2_o}{2} \mathbb{E} \left( \left\| \omega^{(t+1)} - \omega^{(t)} \right\|^2_2 + \frac{3}{2} \mu^2 \eta_o \mathbb{E} \left\| \epsilon^{(t+1)} - \epsilon^{(t)} \right\|^2_2 \right).
\]

Plugging (52) into (51) yields (48), which concludes the proof of Lemma 8. \( \square \)

**Lemma 9.** For the update of \( \omega \), we have

\[
\frac{1}{2} \mathbb{E} \left\| \omega^{(t+1)} - \omega^{(t)} \right\|^2_2 - \frac{1}{2} \mathbb{E} \left\| \omega^{(t)} - \omega^{(t-1)} \right\|^2_2
\]

\[
\leq \left( \frac{\eta_o (5L + 1)}{2} + \frac{3\eta_o \eta_o S^2_o}{2} \right) \mathbb{E} \left( \left\| \omega^{(t+1)} - \omega^{(t)} \right\|^2_2 + \frac{\eta_o L \omega}{2} \mathbb{E} \left\| \omega^{(t)} - \omega^{(t-1)} \right\|^2_2 \right)
\]

\[
- \frac{\eta_o}{2\eta} \mathbb{E} \left\| \theta^{(t+2)} - \theta^{(t+1)} \right\|^2_2 + \left( \frac{3}{2} \eta_o \eta_o \mu^2 + \frac{\eta_o}{2\eta} - \frac{\mu \eta_o}{2} \right) \mathbb{E} \left\| \theta^{(t+1)} - \theta^{(t)} \right\|^2_2
\]

\[
+ \left( \eta^2_o + \frac{\eta_o}{2\mu} \right) \mathbb{E} \left( \left\| \xi^{(t+1)}_\theta \right\|^2_2 + \frac{\eta_o}{2} \mathbb{E} \left\| \xi^{(t-1)}_\theta \right\|^2_2 + \frac{3}{2} \eta_o \mathbb{E} \left\| \xi^{(t+1)}_\theta \right\|^2_2 + \mathbb{E} \left\| \xi^{(t)}_\theta \right\|^2_2 \right).
\]
Proof. Combing Lemma 7 and Lemma 8, we obtain
\[
E(\omega(t+1) - \omega(t) - (\omega(t) - \omega(t-1)), \omega(t+1) - \omega(t)) \\
\leq -\frac{\eta_1}{2\eta_0}E\left(\frac{3}{2} \eta_1 \eta_0 \mu^2 + \frac{\eta_1}{2\eta_0}\right)E\left(\theta(t+1) - \theta(t)\right)^2 \\
+ \frac{3\eta_1 \eta_0 \omega^2}{2}E\left(\omega(t+1) - \omega(t)\right)^2_2 + \frac{\eta_1 \omega}{2}E\left(\theta(t+1) - \theta(t)\right)^2_2 \\
+ 2\eta_0 \omega E\left(\omega(t+1) - \omega(t)\right)^2_2 + \frac{\eta_1 \omega (L+1)}{2}E\left(\omega(t+1) - \omega(t)\right)^2_2 + \frac{\eta_1 \omega^2}{2}E\left(\omega(t) - \omega(t-1)\right)^2_2 \\
+ (\eta_0^2 + \frac{\eta_0}{2\mu})E\left(\xi^2_0\right)^2_2 + \frac{\eta_0}{2}E\left(\xi_{t-1}^2\right)^2_2 + \frac{3}{2} \eta_1 \eta_0 \cdot \left(E\left(\xi_{t+1}^2\right)^2_2 + E\left(\xi_{t}^2\right)^2_2\right). \tag{53}
\]
Note that we have
\[
E(\omega(t+1) - \omega(t) - (\omega(t) - \omega(t-1)), \omega(t+1) - \omega(t)) \\
= \frac{1}{2}E\left(\omega(t+1) - \omega(t)\right)^2_2 - \frac{1}{2}E\left(\omega(t) - \omega(t-1)\right)^2_2 + \frac{1}{2}E\left(\omega(t+1) - \omega(t) - (\omega(t) - \omega(t-1))\right)^2_2 \\
\geq \frac{1}{2}E\left(\omega(t+1) - \omega(t)\right)^2_2 - \frac{1}{2}E\left(\omega(t) - \omega(t-1)\right)^2_2. \tag{54}
\]
Combining (53) and (54), we conclude the proof of Lemma 9.
\[\square\]

Lemma 10. Suppose Assumption 1, 3 and 5 hold. \(F(\omega(t), \theta(t))\) is lower bounded throughout all iterations.

Proof. By definition of \(F\) in (2), we have
\[
F(\omega(t), \theta(t)) \geq E_{\tilde{\epsilon}(t)} \tilde{r}(s, a) - E_{\pi_t} \tilde{r}(\psi(s), \psi(a)) - \frac{\mu}{2} \left\| \theta \right\|_2^2 - \lambda B_H \\
\geq -\left(2\sqrt{2} \rho_\theta \kappa + \frac{\mu}{2} \kappa^2 + \lambda B_H\right). \tag{55}
\]
\[\square\]

B.4 PROOF OF LEMMA 1

Recall that we construct a potential function that dacays monotonically along the solution path, which takes the form
\[
E^{(t+1)} = EF(\omega(t+1), \theta(t+1)) + s \cdot \left(\frac{1 + 2\eta_1 \omega}{4} \frac{\mu_1}{2} \frac{S_\omega}{2} + \frac{3}{2} \eta_1 \eta_0 \mu^2\right)E\left(\theta(t+1) - \theta(t)\right)^2_2 + \frac{\mu_1}{8} E\left(\theta(t+1) - \theta(t)\right)^2_2. \tag{56}
\]
for some constant \(s > 0\). We define five constants \(k_1, k_2, k_3, k_4, k_5\) as
\[
k_1 = \frac{1}{2\eta_1} - s \cdot \left(\frac{7L_\omega + 1}{2} \frac{3\eta_1 \eta_0 S_{\omega}^2}{2}\right), \quad k_2 = s \cdot \frac{\eta_1 \omega}{2} \frac{S_\omega}{2}, \\
k_3 = s \cdot \left(\frac{\mu_1}{4} - \frac{3}{2} \eta_1 \eta_0 \mu^2\right), \quad k_4 = s \cdot \left(\frac{\mu_1}{8} - \frac{1}{2\eta_1} + \frac{S_\omega + 2\mu}{2}\right), \\
k_5 = s \cdot \frac{\mu_1}{8} - \left(\frac{L_\omega}{8} \frac{S_{\omega} + 2\mu}{2}\right).
\]
Here, we restate Lemma 1 and then prove it.

Lemma 1. We choose step sizes \(\eta_0, \eta_1\) satisfying
\[
\eta_0 \leq \min \left\{ \frac{L_\omega}{S_{\omega}(8L_\omega + 2)}, \frac{1}{2L_\omega} \right\}, \quad \eta_1 \leq \min \left\{ \frac{1}{150\mu}, \frac{7L_\omega + 1}{150S_{\omega}^2}, \frac{1}{100(2\mu + S_\omega)} \right\}.
\]

and meanwhile $\eta_\omega/\eta_\theta \leq \mu/(30L_\omega + 5)$, where $L_\omega = \sqrt{2\kappa_0\rho_0|A| + \lambda S_H}$. Then we have
\[
\mathcal{E}^{(t+1)} - \mathcal{E}^{(t)} \leq -k_1\mathbb{E}\left[\omega^{(t+1)} - \omega^{(t)}\right]^2 - k_2\mathbb{E}\left[\omega^{(t)} - \omega^{(t-1)}\right]^2 - k_3\mathbb{E}\left[\theta^{(t+2)} - \theta^{(t+1)}\right]^2
\]
\[
- k_4\mathbb{E}\left[\omega^{(t+1)} - \omega^{(t)}\right]^2 - k_5\mathbb{E}\left[\omega^{(t)} - \omega^{(t-1)}\right]^2
\]
\[
+ \eta_\omega\mathbb{E}\left[\xi^{(t)}_\omega\right]^2 + \frac{1}{2}\mathbb{E}\left[\xi^{(t-1)}_\theta\right]^2
\]
\[
+ s \cdot \left(\eta_\omega^2\mathbb{E}\left[\xi^{(t)}_\omega\right]^2 + \eta_\theta^2\mathbb{E}\left[\xi^{(t-1)}_\theta\right]^2 + \frac{3\eta_\omega\eta_\theta}{2} \cdot (\mathbb{E}\left[\xi^{(t+1)}_\theta\right]^2 + \mathbb{E}\left[\xi^{(t)}_\theta\right]^2)\right). \quad (57)
\]
Moreover, we have $k_1, k_2, k_3, k_4, k_5 > 0$ for
\[
s = \frac{8}{\eta_\omega^2(58L_\omega + 9)}.
\]

**Proof.** For notational simplicity, we define
\[
K^{(t+1)} = \frac{1}{2} + \frac{\eta_\omega}{2\eta_\theta} \cdot \mathbb{E}\left[\omega^{(t+1)} - \omega^{(t)}\right]^2
\]
\[
+ \left(\frac{\eta_\omega}{2\eta_\theta} - \frac{\mu\eta_\omega}{4} + \frac{3\eta_\theta\eta_\omega}{2}\right) \cdot \mathbb{E}\left[\omega^{(t+1)} - \omega^{(t)}\right]^2 + \frac{\mu\eta_\omega}{8} \mathbb{E}\left[\omega^{(t+1)} - \theta^{(t)}\right]^2. \quad (58)
\]
By rearranging the inequality in Lemma (9), we obtain
\[
K^{(t+1)} - K^{(t)} \leq \left(\frac{\eta_\omega}{2\eta_\theta} + \frac{3\eta_\theta\eta_\omega S_\omega^2}{2}\right) \mathbb{E}\left[\omega^{(t+1)} - \omega^{(t)}\right]^2
\]
\[
- \frac{\eta_\omega}{2\eta_\theta} \mathbb{E}\left[\omega^{(t+1)} - \omega^{(t-1)}\right]^2 - \left(\frac{\mu\eta_\omega}{4} - \frac{3\mu^2\eta_\theta\eta_\omega}{2}\right) \mathbb{E}\left[\omega^{(t+1)} - \theta^{(t)}\right]^2
\]
\[
- \frac{\mu\eta_\omega}{8} \mathbb{E}\left[\omega^{(t+1)} - \theta^{(t)}\right]^2 - \frac{\mu\eta_\omega}{8} \mathbb{E}\left[\omega^{(t)} - \theta^{(t)}\right]^2
\]
\[
+ \left(\eta_\omega^2 + \frac{\eta_\theta^2}{2}\right) \mathbb{E}\left[\omega^{(t)}\right]^2 + \frac{\eta_\theta^2}{2} \mathbb{E}\left[\omega^{(t-1)}\right]^2 + \frac{3\eta_\omega\eta_\theta}{2} \cdot (\mathbb{E}\left[\xi^{(t+1)}_\theta\right]^2 + \mathbb{E}\left[\xi^{(t)}_\theta\right]^2). \quad (59)
\]
By definition of $P^{(t)}$ in (56), we have
\[
\mathcal{E}^{(t)} = F(\omega^{(t)}, \theta^{(t)}) + s \cdot K^{(t)}
\]
for some constant $s > 0$. Combining (59) and Lemma 6, since $\eta_\omega < \frac{1}{15\mu}$, we obtain
\[
\mathcal{E}^{(t+1)} - \mathcal{E}^{(t)} \leq -\left(\frac{2\eta_\omega}{\eta_\theta} - s \cdot \left(\frac{\eta_\omega}{2\eta_\theta} + \frac{3\eta_\theta\eta_\omega S_\omega^2}{2}\right)\right) \mathbb{E}\left[\omega^{(t+1)} - \omega^{(t)}\right]^2
\]
\[
- \left(s \cdot \frac{\eta_\omega}{2\eta_\theta} - \frac{S_\omega}{2}\right) \mathbb{E}\left[\omega^{(t)} - \omega^{(t-1)}\right]^2 - s \cdot \left(\frac{\mu\eta_\omega}{4} - \frac{3\mu^2\eta_\theta\eta_\omega}{2}\right) \mathbb{E}\left[\omega^{(t)} - \omega^{(t)}\right]^2
\]
\[
- \left(s \cdot \frac{\mu\eta_\omega}{8} - \frac{\mu\eta_\omega + 2\mu + S_\omega}{2}\right) \mathbb{E}\left[\omega^{(t)} - \omega^{(t-1)}\right]^2
\]
\[
- \left(s \cdot \frac{\mu\eta_\omega}{8} - \frac{\mu}{2}\right) \mathbb{E}\left[\omega^{(t)} - \omega^{(t)}\right]^2
\]
\[
+ \eta_\omega \mathbb{E}\left[\xi^{(t)}_\omega\right]^2 + \frac{1}{2\mu} \mathbb{E}\left[\xi^{(t-1)}_\theta\right]^2
\]
\[
+ s \cdot \left(\eta_\omega^2 + \frac{\eta_\theta^2}{2}\right) \mathbb{E}\left[\xi^{(t)}_\omega\right]^2 + \frac{\eta_\theta^2}{2} \mathbb{E}\left[\xi^{(t-1)}_\theta\right]^2 + \frac{3\eta_\omega\eta_\theta}{2} \cdot (\mathbb{E}\left[\xi^{(t+1)}_\theta\right]^2 + \mathbb{E}\left[\xi^{(t)}_\theta\right]^2). \quad (60)
\]
Since $\eta_\theta < \frac{1}{15\mu\omega}$, we have $k_3 > 0$. Now we choose a proper constant $s$ such that $k_1, k_2, k_4, k_5$ are positive. Note that they are positive if and only if
\[
1/(2\eta_\omega) > s \cdot ((\eta_\omega(7L_\omega + 1)/2) + \frac{3\eta_\omega\eta_\theta S_\omega^2}{2}), \quad (61)
\]
\[
s \cdot \eta_\omega L_\omega/2 > S_\omega/2, \quad (62)
\]
\[
s \cdot (\mu\eta_\omega) > 1/2\eta_\theta + (2\mu + S_\omega)/2, \quad (63)
\]
\[
s \cdot (\mu\eta_\omega)/8 > 1/(2\eta_\theta) + \mu/2. \quad (64)
\]
Rearranging the terms in (61), (62), (63), (64), we obtain
\[
\frac{S_\omega}{\eta_\omega L_\omega} < s < \frac{1/(2\eta_\omega)}{\eta_\omega(7L_\omega + 1 + 3\eta_\omega S_\omega^2)/2}, \quad (65)
\]
\[
\frac{1/(2\eta_\theta) + (S_\omega + 2\mu)/\mu\eta_\omega/8}{\eta_\omega(7L_\omega + 1 + 3\eta_\omega S_\omega^2)/2} < s < \frac{1/(2\eta_\omega)}{\eta_\omega(7L_\omega + 1 + 3\eta_\omega S_\omega^2)/2}. \quad (66)
\]
Since
\[
\eta_\theta \leq \frac{7L_\omega + 1}{150S_\omega^2} \quad \text{and} \quad \eta_\theta < \frac{1}{100(2\mu + S_\omega)},
\]
by rearranging the terms in (65) and (66) and taking the leading terms, we obtain
\[
\eta_\omega < \frac{L_\omega}{S_\omega(8L_\omega + 2)} \quad \text{and} \quad \frac{\eta_\omega}{\eta_\theta} < \frac{\mu}{30L_\omega + 5}.
\]
Therefore, we have
\[
\frac{\eta_\omega}{\eta_\theta} < \frac{\mu}{30L_\omega + 5}, \quad \eta_\omega < \eta_\omega \text{ and } \eta_\theta < \eta_\theta,
\]
where
\[
\eta_\omega = \min \left\{ \frac{L_\omega}{S_\omega(8L_\omega + 2)}, \frac{1}{2L_\omega} \right\},
\]
\[
\eta_\theta = \min \left\{ \frac{1}{150\mu}, \frac{1}{100(2\mu + S_\omega)} \right\}.
\]

\[\square\]

**B.5 Proof of Theorem 2**

Let \( k = 1/\min(k_1, k_2) \), and \( \phi = \max\{1, 1/\eta_\omega^2, 1/\eta_\theta^2\} \). Then we have
\[
NJ_N \leq \sum_{t=1}^{N} \phi \cdot E\left(\left\|\omega^{(t+1)} - \omega^{(t)}\right\|^2_2 + \left\|\theta^{(t+1)} - \theta^{(t)}\right\|^2_2\right)
\]
\[
\leq \phi k \left( \sum_{t=1}^{N} (E^{(t)} - E^{(t+1)}) + \sum_{t=1}^{N} \left( \eta_\omega + s\eta_\omega^2 + \frac{s\eta_\omega}{2\mu} \right) E \left\| \xi^{(t)}_\omega \right\|^2_2 + \frac{1}{2\mu} \cdot E \left\| \xi^{(t-1)}_\theta \right\|^2_2 \right)
\]
\[
+ s \cdot (\eta_\omega/2 \cdot E \left\| \xi^{(t-1)}_\theta \right\|^2_2 + 3\eta_\omega \eta_\theta/2 \cdot \left( E \left\| \xi^{(t+1)}_\theta \right\|^2_2 + E \left\| \xi^{(t)}_\theta \right\|^2_2 \right))
\]
\[
\leq \phi k \left( (E^{(1)} - E^{(N)}) + \sum_{t=1}^{N} \left( 2 \max \left\{ \eta_\omega + s\eta_\omega^2 + \frac{s\eta_\omega}{2\mu}, \frac{s\eta_\omega}{2\mu} \right\} \right) E \left\| \xi^{(t)}_\omega \right\|^2_2 \right)
\]
\[
+ 3 \max \left\{ \frac{1}{2\mu}, \frac{3\eta_\omega \eta_\theta}{2} \right\} \left( E \left\| \xi^{(t)}_\theta \right\|^2_2 \right) . \quad (67)
\]

Now set
\[
\nu = \max \left\{ 2 \max \left( \eta_\omega + s\eta_\omega^2 + \frac{s\eta_\omega}{2\mu}, \frac{s\eta_\omega}{2\mu} \right), 3 \max \left( \frac{1}{2\mu}, \frac{3\eta_\omega \eta_\theta}{2} \right) \right\}
\]
and divide both sides of (67) by \( N \), we have
\[
J_N \leq \frac{k\phi(E^{(1)} - E^{(N)})}{N} + k\phi \nu \left( M_G + \frac{M_0}{\eta_\theta} \right). \quad (68)
\]
By definition of \( E \) in (56) and Lemma 10, we have
\[
E^{(N)} \geq F(\omega^{(N)}, \theta^{(N)}) \geq - \left( 2\sqrt{2}\rho_\theta \kappa + \frac{\mu}{2} \kappa^2 + \lambda B_H \right) > -\infty.
\]
Now for any given \( \epsilon > 0 \), we take
\[
q_\theta = \frac{4k\phi M_\theta}{\epsilon}, \quad q_\omega = \frac{4k\phi M_\omega}{\epsilon} \quad \text{and} \quad N = k\phi \frac{2\epsilon}{\epsilon},
\]
and obtain
\[
J_N \leq \epsilon.
\]

27
C Empirical Maximizer Case

Notice that the object function (2) is quadratic in $\theta$, and thus we are able to get the empirical maximizer by simple calculation instead of performing stochastic gradient ascent at each iteration. This allows us to avoid using large batches while achieving the same sample complexity $\tilde{O}(\frac{1}{\epsilon^2})$.

Given a fixed $\omega(t)$, by definition of $G$ in (25), we can get the optimal $\theta^*(\omega(t))$ in population form:

$$
\theta^*(\omega(t)) = \frac{1}{\mu} [G(\pi_{\omega(t)}) - G(\pi^*)].
$$  \hfill (69)

At the $(t + 1)$-th iteration, we can achieve the closed form of the empirical maximizer by sampling one trajectory randomly and then apply stochastic gradient descent to the outer minimization. More specifically, the updating rule is defined as

$$
\tilde{\theta}(t + 1)(\omega(t)) = \frac{1}{\mu} [\tilde{G}(\pi_{\omega(t)}) - \tilde{G}(\pi^*)],
$$  \hfill (70)

$$
\omega(t + 1) = \omega(t) - \eta \nabla_\omega \tilde{f}_t(\omega(t), \tilde{\theta}(t + 1)),
$$  \hfill (71)

where $\tilde{G}$ is the empirical version of $G$ obtained by sampling, and $\nabla_\omega \tilde{f}$ is the stochastic approximation of $\nabla_\omega F$ at the $(t + 1)$-th iteration.

We define the stationary point as follow.

Definition 4. We call $\omega^*$ an stationary point if $\nabla_\omega F(\omega^*, \theta^*(\omega^*)) = 0$.

Before we proceed with the convergence analysis, we impose the following assumptions on the problem.

Assumption 6. There is some constant $M_G > 0$ s.t. for any $\omega, \tilde{\theta}$ and $\pi$, the following two conditions hold.

Unbiased : $\mathbb{E} \nabla_\omega \tilde{f}_t(\omega, \tilde{\theta}) = \nabla_\omega F(\omega, \tilde{\theta}), \mathbb{E} \tilde{G}(\pi) = \mathbb{E} G(\pi)$.

Gradient bounded : $\mathbb{E} \| \nabla_\omega \tilde{f}_t(\omega, \theta) \|^2_2 \leq M_G$.

Assumption 6 requires the stochastic gradient to be unbiased with bounded second moment.

Corollary 4. Under Assumption 1, 3 and 5, there exists $B_F = \frac{12\rho^2}{\mu} + \lambda H$ such that for any $\omega$, we have $|F(\omega, \theta^*(\omega))| < B_F$.

Proof. By Equation (69), we have $\|\theta^*(\omega)\|_2 \leq \frac{2\sqrt{2} \rho}{\mu}$. Plugging this into $F(\omega, \theta^*(\omega))$, we have for any $\omega$,

$$
|F(\omega, \theta^*(\omega))| = |\mathbb{E}_{\omega_{\omega}} (\theta^*(\omega), g(\psi s, \psi a)) - \mathbb{E}_{\pi_{\omega}} (\theta^*(\omega), g(\psi s, \psi a)) - \lambda H(\pi_{\omega}) - \frac{\mu}{2} \|\theta^*(\omega)\|^2_2 |
$$

$$
\leq 2 \|\theta^*(\omega)\|_2 \cdot \max_{s,a} \|g(\psi s, \psi a)\|_2 + \lambda H(\pi_{\omega}) + \frac{\mu}{2} \|\theta^*(\omega)\|^2_2
$$

$$
\leq \frac{12 \rho^2}{\mu} + \lambda H.
$$

For notational simplicity, We define $\xi_{\omega}^{(t)} = \nabla_\theta F(\omega(t), \tilde{\theta}(t+1)) - \nabla_\theta \tilde{f}_t(\omega(t), \tilde{\theta}(t+1))$ as the i.i.d stochastic noise and $\sigma^2_{\theta} = \theta^*(\omega(t)) - \hat{\theta}(t)$ as the i.i.d error of empirical maximizer. We denote $\mathbb{E}$ as the expectation over $\xi_{\omega}^{(t)} (t \geq 1)$ and $\sigma^2_{\theta} (t \geq 1)$. We measure the sub-stationarity of the algorithm at the iteration $N$ by

$$
I_N = \min_{1 \leq t \leq N} \mathbb{E} \|\nabla_\omega F(\omega(t), \theta^*(\omega(t)))\|^2_2.
$$

Then we state the global convergence of the above mentioned optimization method.
Theorem 3. Suppose Assumptions 1, 3, 5, 6 hold. Given any $\epsilon > 0$, we take $\eta = \sqrt{2(\mu + \lambda B_H)}(L_\omega + S_\mu^2/\mu)M_G$, then we need at most
\[ N = O\left(\frac{(\rho_2^2/\mu + \lambda B_H)(L_\omega + S_\mu^2/\mu)M_G}{\epsilon^2}\right) \]
iterations to have $I_N < \epsilon$.

Proof. By employing the first inequality in Lemma 4, we have
\[ F(\omega^{(t+1)}, \theta^*(\omega^{(t)})) - F(\omega^{(t)}, \theta^*(\omega^{(t)})) - \langle \nabla_\omega F(\omega^{(t)}, \theta^*(\omega^{(t)})) , \omega^{(t+1)} - \omega^{(t)} \rangle \]
\[ \leq \frac{L_\omega}{2} \left\| \omega^{(t+1)} - \omega^{(t)} \right\|_2^2. \] (72)
Note that
\[ \mathbb{E}\langle \nabla_\omega F(\omega^{(t)}, \theta^*(\omega^{(t)})) , \omega^{(t+1)} - \omega^{(t)} \rangle \]
\[ = \mathbb{E}\langle \nabla_\omega F(\omega^{(t)}, \theta^*(\omega^{(t)})) , -\eta \omega \langle \nabla_\omega F(\omega^{(t)}, \bar{\theta}^{(t)}) + \xi^{(t)} \rangle \rangle \]
\[ = \mathbb{E}\langle \nabla_\omega F(\omega^{(t)}, \theta^*(\omega^{(t)})) , -\eta \omega \langle \nabla_\omega F(\omega^{(t)}, \theta^*(\omega^{(t)}) - \sigma^t_\theta) + \xi^{(t)} \rangle \rangle \]
\[ = \mathbb{E}\langle \nabla_\omega F(\omega^{(t)}, \theta^*(\omega^{(t)})) , -\eta \omega \langle \nabla_\omega F(\omega^{(t)}, \theta^*(\omega^{(t)})) + \xi^{(t)} \rangle \rangle \]
\[ = -\eta \omega \mathbb{E} \left\| \nabla_\omega F(\omega^{(t)}, \theta^*(\omega^{(t)})) \right\|_2^2, \] (73)
where (i) comes from the unbiased property of $\bar{\theta}^{(t)}$ and the fact that $\nabla_\omega F(\omega, \theta)$ is linear in $\theta$, and
\[ \langle ii \rangle \] comes from the unbiased property of $f_j(\omega, \bar{\theta})$. Now taking the expectation on both sides of (72)
and plugging (73) in, we obtain
\[ \mathbb{E} F(\omega^{(t+1)}, \theta^*(\omega^{(t)})) - F(\omega^{(t)}, \theta^*(\omega^{(t)})) + \eta \omega \mathbb{E} \left\| \nabla_\omega F(\omega^{(t)}, \theta^*(\omega^{(t)})) \right\|_2^2 \leq \frac{L_\omega}{2} \eta \omega^2 M_G. \] (74)
Dividing both sides by $\eta \omega$ and rearranging the terms in (74), we get
\[ \mathbb{E} \left\| \nabla_\omega F(\omega^{(t)}, \theta^*(\omega^{(t)})) \right\|_2^2 \leq \frac{\mathbb{E} F(\omega^{(t)}, \theta^*(\omega^{(t)})) - F(\omega^{(t+1)}, \theta^*(\omega^{(t)}))}{\eta \omega} \]
\[ \leq \frac{\mathbb{E} F(\omega^{(t)}, \theta^*(\omega^{(t)})) - F(\omega^{(t+1)}, \theta^*(\omega^{(t+1)}))}{\eta \omega} + \frac{L_\omega}{2} \eta \omega M_G \]
\[ + \frac{\mathbb{E} F(\omega^{(t+1)}, \theta^*(\omega^{(t+1)})) - F(\omega^{(t+1)}, \theta^*(\omega^{(t)}))}{\eta \omega}. \] (75)
Now consider $F(\omega^{(t+1)}, \theta^*(\omega^{(t+1)})) - F(\omega^{(t+1)}, \theta^*(\omega^{(t)}))$, we have
\[ F(\omega^{(t+1)}, \theta^*(\omega^{(t+1)})) - F(\omega^{(t+1)}, \theta^*(\omega^{(t)})) \]
\[ = \langle G(\pi^{(t+1)}), \theta^*(\omega^{(t+1)})) - \langle G(\pi^{(t+1)}), \theta^*(\omega^{(t)}) \rangle \]
\[ - \frac{\mu}{2} \left( \left\| \theta^*(\omega^{(t+1)})) \right\|_2^2 - \left\| \theta^*(\omega^{(t)}) \right\|_2^2 \right) \]
\[ = \langle \mu (\theta^{(t+1)})^2, \theta^*(\omega^{(t+1)}) - \theta^*(\omega^{(t)})) \]
\[ - \frac{\mu}{2} \left( \left\| \theta^*(\omega^{(t+1)})) + \theta^*(\omega^{(t)}) \right\|_2^2 - \left\| \theta^*(\omega^{(t+1)})) - \theta^*(\omega^{(t)})) \right\|_2^2 \]
\[ = \frac{\mu}{2} \left\| \theta^*(\omega^{(t+1)}) - \theta^*(\omega^{(t)}) \right\|_2^2 \]
\[ = \frac{\mu}{2} \left\| \frac{1}{\mu} (G(\pi^{(t+1)})) - G(\pi^{(t)})) \right\|_2^2 \]
\[ \leq \frac{S_\mu^2}{2\mu} \left\| \omega^{(t+1)} - \omega^{(t)} \right\|_2^2. \] (76)
Taking expectation on both sides of (76) with respect to the noise introduced by SGD, we have
\[ \mathbb{E} F(\omega(t+1), \theta^*(\omega(t+1))) - F(\omega(t+1), \theta^*(\omega(t))) \leq \frac{S_\omega^2}{2\mu} \eta^2 M_G. \]

Summing the equation (75) up, we have
\[
\sum_{i=1}^{N} \mathbb{E} \left\| \nabla_\omega F(\omega(t), \theta^*(\omega(t))) \right\|_2^2 \\
\leq \frac{1}{\eta_\omega} \sum_{i=1}^{N} \mathbb{E} F(\omega(t), \theta^*(\omega(t))) - F(\omega(t+1), \theta^*(\omega(t))) + \frac{L_\omega}{2} N \eta_\omega M_G \\
\leq \frac{1}{\eta_\omega} \sum_{i=1}^{N} \mathbb{E} F(\omega(t), \theta^*(\omega(t))) - F(\omega(t+1), \theta^*(\omega(t))) \\
+ \frac{1}{\eta_\omega} \sum_{i=1}^{N} \mathbb{E} F(\omega(t+1), \theta^*(\omega(t))) - F(\omega(t+1), \theta^*(\omega(t+1))) + \frac{L_\omega}{2} N \eta_\omega M_G \\
\leq \frac{1}{\eta_\omega} \sum_{i=1}^{N} \mathbb{E} F(\omega(t), \theta^*(\omega(t))) - F(\omega(t+1), \theta^*(\omega(t+1))) \\
+ \frac{1}{\eta_\omega} \sum_{i=1}^{N} S_\omega^2 \eta_\omega^2 M_G + \frac{L_\omega}{2} N \eta_\omega M_G.
\]

Dividing both sides of the above equation by \( N \), we get
\[
\min_{1 \leq t \leq N} \mathbb{E} \left\| \nabla_\omega F(\omega(t), \theta^*(\omega(t))) \right\|_2^2 \leq \frac{1}{N \eta_\omega} \frac{\mathbb{E} F(\omega(1), \theta^*(\omega(1))) - \mathbb{E} F(\omega(N+1), \theta^*(\omega(N+1)))}{N \eta_\omega} + \frac{L_\omega}{2} + \frac{S_\omega^2}{2\mu} \eta_\omega M_G.
\]

Since \( |\mathbb{E} F(\omega(1), \theta^*(\omega(1))) - \mathbb{E} F(\omega(N+1), \theta^*(\omega(N+1)))| \leq 2B_F \). Take \( \eta_\omega = \sqrt{\frac{2B_F}{(L_\omega/2 + S_\omega^2/2)M_G N}} \), then we have
\[
\min_{1 \leq t \leq N} \mathbb{E} \left\| \nabla_\omega F(\omega(t), \theta^*(\omega(t))) \right\|_2^2 \leq 4 \sqrt{\frac{B_F (L_\omega + S_\omega^2/\mu) M_G}{N}}
\]

where \( B_F = \frac{12\rho^2}{\mu} + \lambda M_H \). This implies that when \( \eta_\omega = \frac{\epsilon}{\sqrt{2(2L_\omega + S_\omega^2/\mu)M_G}} \), we need at most
\[
N = O \left( \frac{(\rho^2/\mu + \lambda M_H)(L_\omega + S_\omega^2/\mu) M_G}{\epsilon^2} \right)
\]
such that \( I_N < \epsilon \). \( \square \)