Generative Adversarial Imitation Learning with Neural Network Parameterization: Global Optimality and Convergence Rate

Yufeng Zhang [1] **Qi Cai** [1] **Zhuoran Yang** [2] **Zhaoran Wang** [1]

Abstract

Generative adversarial imitation learning (GAIL) demonstrates tremendous success in practice, especially when combined with neural networks. Different from reinforcement learning, GAIL learns both policy and reward function from expert (human) demonstration. Despite its empirical success, it remains unclear whether GAIL with neural networks converges to the globally optimal solution. The major difficulty comes from the nonconvex-nonconcave minimax optimization structure. To bridge the gap between practice and theory, we analyze a gradient-based algorithm with alternating updates and establish its sublinear convergence to the globally optimal solution. To the best of our knowledge, our analysis establishes the global optimality and convergence rate of GAIL with neural networks for the first time.

1. Introduction

The goal of imitation learning (IL) is to learn to perform a task based on expert demonstration (Ho & Ermon, 2016). In contrast to reinforcement learning (RL), the agent only has access to the expert trajectories but not the rewards. The most straightforward approach of IL is behavioral cloning (BC) (Pomerleau, 1991). BC treats IL as the supervised learning problem of predicting the actions based on the states. Despite its simplicity, BC suffers from the compounding errors caused by covariate shift (Ross et al., 2011; Ross & Bagnell, 2010). Another approach of IL is inverse reinforcement learning (IRL) (Russell, 1998; Ng & Russell, 2000; Levine & Koltun, 2012; Finn et al., 2016), which jointly learns the reward function and the corresponding optimal policy. IRL formulates IL as a bilevel optimization problem. Specifically, IRL solves an RL subproblem given a reward function at the inner level and searches for the reward function which makes the expert policy optimal at the outer level. However, IRL is computationally inefficient as it requires fully solving an RL subproblem at each iteration of the outer level. Moreover, the desired reward function may be nonunique. To address such issues of IRL, (Ho & Ermon, 2016) propose generative adversarial imitation learning (GAIL), which searches for the optimal policy without fully solving an RL subproblem given a reward function at each iteration. GAIL solves IL via minimax optimization with alternating updates. In particular, GAIL alternates between (i) minimizing the discrepancy in expected cumulative reward between the expert policy and the learned policy and (ii) maximizing such a discrepancy over the reward function class. Such an alternating update scheme mirrors the training of generative adversarial networks (GANs) (Goodfellow et al., 2014; Arjovsky et al., 2017), where the learned policy acts as the generator while the reward function acts as the discriminator.

Incorporated with neural networks, which parameterize the learned policy and the reward function, GAIL achieves significant empirical success in challenging applications, such as natural language processing (Yu et al., 2016), autonomous driving (Kuefler et al., 2017), human behavior modeling (Merel et al., 2017), and robotics (Tai et al., 2018). Despite its empirical success, GAIL with neural networks remains less understood in theory. The major difficulty arises from the following aspects: (i) GAIL involves minimax optimization, while the existing analysis of policy optimization with neural networks (Anthony & Bartlett, 2009; Liu et al., 2019; Bhandari & Russo, 2019; Wang et al., 2019) only focuses on a minimization or maximization problem. (ii) GAIL with neural networks is nonconvex-nonconcave, and therefore, the existing analysis of convex-concave optimization with alternating updates is not applicable (Nesterov, 2013). There is an emerging body of literature (Rafique et al., 2018; Zhang et al., 2019b) that casts nonconvex-nonconcave optimization as bilevel optimization, where the inner level is solved to a high precision as in IRL. However, such analysis is not applicable to GAIL as it involves alternating updates.

[1]Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL 60208, USA [2]Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA. Correspondence to: Zhaoran Wang <zhaoranwang@gmail.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

In this paper, we bridge the gap between practice and theory by establishing the global optimality and convergence of GAIL with neural networks. Specifically, we parameterize the learned policy and the reward function with two-layer neural networks and consider solving GAIL by alternatively updating the learned policy via a step of natural policy gradient (Kakade, 2002; Peters & Schaal, 2008) and the reward function via a step of gradient ascent. In particular, we parameterize the state-action value function (also known as the Q-function) with a two-layer neural network and apply a variant of the temporal difference algorithm (Sutton & Barto, 2018) to solve the policy evaluation subproblem in natural policy gradient. We prove that the learned policy $\bar{\pi}$ converges to the expert policy $\pi_{\rm E}$ at a $1/\sqrt{T}$ rate in the $\mathcal{R}$-distance (Chen et al., 2020), which is defined as $\mathbb{D}_{\mathcal{R}}(\pi_{\rm E}, \bar{\pi}) = \max_{r\in\mathcal{R}} J(\pi_{\rm E}; r) - J(\bar{\pi}; r)$. Here $J(\pi; r)$ is the expected cumulative reward of a policy π given a reward function $r(s, a)$ and $\mathcal{R}$ is the reward function class. The core of our analysis is constructing a potential function that tracks the $\mathcal{R}$-distance. Such a rate of convergence implies that the learned policy $\bar{\pi}$ (approximately) outperforms the expert policy $\pi_{\rm E}$ given any reward function $r \in \mathcal{R}$ within a finite number of iterations T. In other words, the learned policy $\bar{\pi}$ is globally optimal. To the best of our knowledge, our analysis establishes the global optimality and convergence of GAIL with neural networks for the first time. It is worth mentioning that our analysis is straightforwardly applicable to linear and tabular settings, which, however, are not our focus.

Related works. Our work extends an emerging body of literature on RL with neural networks (Xu et al., 2019a; Zhang et al., 2019a; Bhandari & Russo, 2019; Liu et al., 2019; Wang et al., 2019; Agarwal et al., 2019) to IL. This line of research analyzes the global optimality and convergence of policy gradient for solving RL, which is a minimization or maximization problem. In contrast, we analyze GAIL, which is a minimax optimization problem.

Our work is also related to the analysis of apprenticeship learning (Syed et al., 2008) and GAIL (Cai et al., 2019a; Chen et al., 2020). (Syed et al., 2008) analyze the convergence and generalization of apprenticeship learning. They assume the state space to be finite, and thus, do not require function approximation for the policy and the reward function. In contrast, we assume the state space to be infinite and employ function approximation based on neural networks. (Cai et al., 2019a) study the global optimality and convergence of GAIL in the setting of linear-quadratic regulators. In contrast, our analysis handles general MDPs without restrictive assumptions on the transition kernel and the reward function. (Chen et al., 2020) study the convergence and generalization of GAIL in the setting of general MDPs. However, they only establish the convergence to a stationary point. In contrast, we establish the global optimality of GAIL.

Notations. Let $[n] = \{1, \ldots, n\}$ for $n \in \mathbb{N}_+$ and $[m : n] = \{m, m+1, \ldots, n\}$ for $m < n$. Also, let $N(\mu, \Sigma)$ be the Gaussian distribution with mean μ and covariance Σ. We denote by $\mathcal{P}(\mathcal{X})$ the set of all probability measures over the space $\mathcal{X}$. For a function $f : \mathcal{X} \to \mathbb{R}$, a constant $p \geq 1$, and a probability measure $\mu \in \mathcal{P}(\mathcal{X})$, we denote by $\|f\|_{p,\mu} = (\int_{\mathcal{X}} |f(x)|^p \mathrm{d}\mu(x))^{1/p}$ the $L_p(\mu)$ norm of the function f. For any two functions $f, g : \mathcal{X} \to \mathbb{R}$, we denote by $\langle f, g\rangle_{\mathcal{X}} = \int_{\mathcal{X}} f(x) \cdot g(x)\mathrm{d}x$ the inner product on the space $\mathcal{X}$.

2. Background

In this section, we introduce reinforcement learning (RL) and generative adversarial imitation learning (GAIL).

2.1. Reinforcement Learning

We consider a Markov decision process (MDP) $(\mathcal{S}, \mathcal{A}, r, P, \rho, \gamma)$. Here $\mathcal{S} \subseteq \mathbb{R}^{d_1}$ is the state space, $\mathcal{A} \subseteq \mathbb{R}^{d_2}$ is the action space, which is assumed to be finite throughout this paper, $r : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ is the reward function, $P : \mathcal{S}\times\mathcal{A} \to \mathcal{P}(\mathcal{S})$ is the transition kernel, $\rho \in \mathcal{P}(\mathcal{S})$ is the initial state distribution, and $\gamma \in (0, 1)$ is the discount factor. Without loss of generality, we assume that $\mathcal{S}\times\mathcal{A}$ is compact and that $\|(s, a)\|_2 \leq 1$ for any $(s, a) \in \mathcal{S}\times\mathcal{A} \subseteq \mathbb{R}^d$, where $d = d_1 + d_2$. An agent following a policy $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$ interacts with the environment in the following manner. At the state $s_t \in \mathcal{S}$, the agent takes the action $a_t \in \mathcal{A}$ with probability $\pi(a_t \,|\, s_t)$ and receives the reward $r_t = r(s_t, a_t)$. The environment then transits into the next state s_{t+1} with probability $P(s_{t+1} \,|\, s_t, a_t)$. Given a policy π and a reward function $r(s, a)$, we define the state-action value function $Q^\pi_r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ as follows,

$$Q^\pi_r(s, a) = \mathbb{E}_\pi\biggl[(1-\gamma)\cdot\sum_{t=0}^\infty \gamma^t \cdot r(s_t, a_t) \,\bigg|\, s_0 = s, a_0 = a\biggr]. \tag{2.1}$$

Here the expectation $\mathbb{E}_\pi$ is taken with respect to $a_t \sim \pi(\cdot \,|\, s_t)$ and $s_{t+1} \sim P(\cdot \,|\, s_t, a_t)$. Correspondingly, we define the state value function $V^\pi_r : \mathcal{S} \to \mathbb{R}$ and the advantage function $A^\pi_r : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ as follows,

$$V^\pi_r(s) = \mathbb{E}_{a\sim\pi(\cdot \,|\, s)}\bigl[Q^\pi_r(s, a)\bigr],$$
$$A^\pi_r(s, a) = Q^\pi_r(s, a) - V^\pi_r(s).$$

The goal of RL is to maximize the following expected cumulative reward,

$$J(\pi; r) = \mathbb{E}_{s\sim\rho}\bigl[V^\pi_r(s)\bigr]. \tag{2.2}$$

The policy π induces a state visitation measure $d_\pi \in \mathcal{P}(\mathcal{S})$ and a state-action visitation measure $\nu_\pi \in \mathcal{P}(\mathcal{S}\times\mathcal{A})$, which take the forms of

$$d_\pi(s) = (1-\gamma)\cdot \sum_{t=0}^{\infty} \gamma^t \cdot \mathbb{P}\big(s_t = s \,\big|\, s_0 \sim \rho, \pi\big),$$
$$\nu_\pi(s,a) = d_\pi(s)\cdot \pi(a\,|\,s). \tag{2.3}$$

It then holds that $J(\pi; r) = \mathbb{E}_{(s,a)\sim\nu_\pi}[r(s,a)]$. Meanwhile, we assume that the policy π induces a state stationary distribution ϱ_π over $\mathcal{S}$, which satisfies that

$$\varrho_\pi(s) = \mathbb{P}\big(s_{t+1} = s \,\big|\, s_t \sim \rho_\pi, a_t \sim \pi(\cdot\,|\,s_t)\big).$$

We denote by $\rho_\pi(s,a) = \varrho(s)\cdot\pi(a\,|\,s)$ the state-action stationary distribution over $\mathcal{S}\times\mathcal{A}$.

2.2. Generative Adversarial Imitation Learning

The goal of imitation learning (IL) is to learn a policy that outperforms the expert policy $\pi_{\rm E}$ based on the trajectory $\{(s_i^{\rm E}, a_i^{\rm E})\}_{i\in[T_{\rm E}]}$ of $\pi_{\rm E}$. We denote by $\nu_{\rm E} = \nu_{\pi_{\rm E}}$ and $d_{\rm E} = d_{\pi_{\rm E}}$ the state-action and state visitation measures induced by the expert policy, respectively, and assume that the expert trajectory $\{(s_i, a_i)\}_{i\in[T_{\rm E}]}$ is drawn from $\nu_{\rm E}$. To this end, we parameterize the policy and the reward function by π_θ for $\theta\in\mathcal{X}_\Pi$ and $r_\beta(s,a)$ for $\beta\in\mathcal{X}_R$, respectively, and solve the following minimax optimization problem known as GAIL (Ho & Ermon, 2016),

$$\min_{\theta\in\mathcal{X}_\Pi}\max_{\beta\in\mathcal{X}_R} L(\theta,\beta), \tag{2.4}$$
$$\text{where } L(\theta,\beta) = J(\pi_{\rm E}; r_\beta) - J(\pi_\theta; r_\beta) - \lambda\cdot\psi(\beta).$$

Here $J(\pi; r)$ is the expected cumulative reward defined in (2.2), $\psi : \mathcal{X}_R \to \mathbb{R}_+$ is the regularizer, and $\lambda \ge 0$ is the regularization parameter. Given a reward function class $\mathcal{R}$, we define the $\mathcal{R}$-distance between two policies π_1 and π_2 as follows,

$$\mathbb{D}_{\mathcal{R}}(\pi_1,\pi_2) = \max_{r\in\mathcal{R}} J(\pi_1; r) - J(\pi_2; r) \tag{2.5}$$
$$= \max_{r\in\mathcal{R}} \mathbb{E}_{\nu_{\pi_1}}\big[r(s,a)\big] - \mathbb{E}_{\nu_{\pi_2}}\big[r(s,a)\big].$$

When $\mathcal{R}$ is the class of 1-Lipschitz functions, $\mathbb{D}_{\mathcal{R}}(\pi_1,\pi_2)$ is the Wasserstein-1 metric between the state-action visitation measures induced by π_1 and π_2. However, $\mathbb{D}_{\mathcal{R}}(\pi_1,\pi_2)$ is not a metric in general. When $\mathbb{D}_{\mathcal{R}}(\pi_1,\pi_2) \le 0$, the policy π_2 outperforms the policy π_1 for any reward function $r\in\mathcal{R}$. Such a notion of $\mathcal{R}$-distance is used in (Chen et al., 2020). We denote by $\mathcal{R}_\beta = \{r_\beta(s,a)\,|\,\beta\in\mathcal{X}_R\}$ the reward function class parameterized with β. Hence, the optimization problem in (2.4) minimizes the $\mathcal{R}_\beta$-distance $\mathbb{D}_{\mathcal{R}_\beta}(\pi_{\rm E},\pi_\theta)$ (up to the regularizer $\lambda\cdot\psi(\beta)$), which searches for a policy $\bar\pi$ that (approximately) outperforms the expert policy given any reward function $r_\beta\in\mathcal{R}_\beta$.

3. Algorithm

In this section, we introduce an algorithm with alternating updates for GAIL with neural networks, which employs natural policy gradient (NPG) to update the policy π_θ and gradient ascent to update the reward function $r_\beta(s,a)$.

3.1. Parameterization with Neural Networks

We define a two-layer neural network with rectified linear units (ReLU) as follows,

$$u_{W,b}(s,a) = \frac{1}{\sqrt{m}}\sum_{l=1}^{m} b_l\cdot \mathbb{1}\big\{(s,a)^\top[W]_l > 0\big\}\cdot(s,a)^\top[W]_l$$
$$= \sum_{l=1}^{m}\big[\phi_{W,b}(s,a)\big]_l^\top[W]_l. \tag{3.1}$$

Here $m\in\mathbb{N}_+$ is the width of the neural network, $b = (b_1,\ldots,b_m)^\top \in\mathbb{R}^m$ and $W = ([W]_1^\top,\ldots,[W]_m^\top)^\top \in\mathbb{R}^{md}$ are the parameters, and $\phi_{W,b}(s,a) = ([\phi_{W,b}(s,a)]_1^\top,\ldots,[\phi_{W,b}(s,a)]_m^\top)^\top\in\mathbb{R}^{md}$ is called the feature vector in the sequel, where

$$\big[\phi_{W,b}(s,a)\big]_l = \frac{b_l}{\sqrt{m}}\cdot\mathbb{1}\big\{(s,a)^\top[W]_l > 0\big\}(s,a). \tag{3.2}$$

It then holds that $u_{W,b}(s,a) = W^\top\phi_{W,b}(s,a)$. Note that the feature vector $\phi_{W,b}(s,a)$ depends on the parameters W and b. We consider the following random initialization,

$$b_l \overset{\text{i.i.d.}}{\sim} \text{Unif}\big(\{-1,1\}\big),\quad [W_0]_l \overset{\text{i.i.d.}}{\sim} N(0, I_d/d),\quad \forall l\in[m]. \tag{3.3}$$

Throughout the training process, we keep the parameter b fixed while updating W. For notational simplicity, we write $u_{W,b}(s,a)$ as $u_W(s,a)$ and $\phi_{W,b}(s,a)$ as $\phi_W(s,a)$ in the sequel. We denote by $\mathbb{E}_{\text{init}}$ the expectation taken with respect to the random initialization in (3.3). For an absolute constant $B>0$, we define the parameter domain as

$$S_B = \big\{W\in\mathbb{R}^{md}\,\big|\,\|W-W_0\|_2\le B\big\}, \tag{3.4}$$

which is the ball centered at W_0 with the domain radius B.

In the sequel, we consider the following energy-based policy,

$$\pi_\theta(a\,|\,s) = \frac{\exp\big(\tau\cdot u_\theta(s,a)\big)}{\sum_{a'\in\mathcal{A}}\exp\big(\tau\cdot u_\theta(s,a')\big)}, \tag{3.5}$$

where $\tau\ge 0$ is the inverse temperature parameter and $u_\theta(s,a)$ is the energy function parameterized by the neural network defined in (3.1) with $W=\theta$. In the sequel, we call θ the policy parameter. Meanwhile, we parameterize the reward function $r_\beta(s,a)$ as follows,

$$r_\beta(s,a) = (1-\gamma)^{-1}\cdot u_\beta(s,a), \tag{3.6}$$

where $u_\beta(s,a)$ is the neural network defined in (3.1) with $W=\beta$ and γ is the discount factor. Here we use the scaling parameter $(1-\gamma)^{-1}$ to ensure that for any policy π the state-action value function $Q^\pi_{r_\beta}(s,a)$ defined in (2.1) is well approximated by the neural network defined in (3.1). In the sequel, we call β the reward parameter and define the reward function class as

$$\mathcal{R}_\beta = \{r_\beta(s,a)\,|\,\beta\in S_{B_\beta}\},$$

where S_{B_β} is the parameter domain of β defined in (3.4) with domain radius B_β. To facilitate algorithm design, we establish the following proposition, which calculates the explicit expressions of the gradients $\nabla L(\theta,\beta)$ and the Fisher information $\mathcal{I}(\theta)$. Recall that the Fisher information is defined as

$$\mathcal{I}(\theta) = \mathbb{E}_{(s,a)\sim\nu_{\pi_\theta}}\big[\nabla_\theta \log\pi_\theta(s,a)\nabla_\theta\log\pi_\theta(s,a)^\top\big]. \tag{3.7}$$

Proposition 3.1 (Gradients and Fisher Information)**.** We call $\iota_\theta(s,a) = \tau^{-1}\cdot\nabla_\theta\log\pi_\theta(a\,|\,s)$ the temperature-adjusted score function. It holds that

$$\iota_\theta(s,a) = \phi_\theta(s,a) - \mathbb{E}_{a'\sim\pi_\theta(\cdot\,|\,s)}\big[\phi_\theta(s,a')\big]. \tag{3.8}$$

It then holds that

$$\mathcal{I}(\theta) = \tau^2\cdot\mathbb{E}_{(s,a)\sim\nu_{\pi_\theta}}\big[\iota_\theta(s,a)\,\iota_\theta(s,a)^\top\big], \tag{3.9}$$

$$\nabla_\theta L(\theta,\beta) = -\tau\cdot\mathbb{E}_{(s,a)\sim\nu_{\pi_\theta}}\big[Q^{\pi_\theta}_{r_\beta}(s,a)\cdot\iota_\theta(s,a)\big], \tag{3.10}$$

$$\begin{aligned}\nabla_\beta L(\theta,\beta) = {}& (1-\gamma)^{-1}\cdot\mathbb{E}_{(s,a)\sim\nu_{\mathrm{E}}}\big[\phi_\beta(s,a)\big]\\ &- (1-\gamma)^{-1}\cdot\mathbb{E}_{(s,a)\sim\nu_{\pi_\theta}}\big[\phi_\beta(s,a)\big]\\ &- \lambda\cdot\nabla_\beta\psi(\beta),\end{aligned} \tag{3.11}$$

where $Q^{\pi_\theta}_{r_\beta}(s,a)$ is the state-action value function defined in (2.1) with $\pi=\pi_\theta$ and $r=r_\beta$, ν_{π_θ} is the state-action visitation measure defined in (2.3) with $\pi=\pi_\theta$, and $\mathcal{I}(\theta)$ is the Fisher information defined in (3.7).

Proof. See §C.1 for a detailed proof. □

Note that the expression of the policy gradient $\nabla_\theta L(\theta,\beta)$ in (3.10) of Proposition 3.1 involves the state-action value function $Q^{\pi_\theta}_{r_\beta}(s,a)$. To this end, we estimate the state-action value function $Q^\pi_r(s,a)$ by $\widehat{Q}_\omega(s,a)$, which is parameterized as follows,

$$\widehat{Q}_\omega(s,a) = u_\omega(s,a). \tag{3.12}$$

Here $u_\omega(s,a)$ is the neural network defined in (3.1) with $W=\omega$. In the sequel, we call ω the value parameter.

3.2. GAIL with Alternating Updates

We employ an actor-critic scheme with alternating updates of the policy and the reward function, which is presented in Algorithm 1. Specifically, we update the policy parameter θ via natural policy gradient and update the reward parameter β via gradient ascent in the actor step, while we estimate the state-action value function $Q^\pi_r(s,a)$ via neural temporal difference (TD) (Cai et al., 2019c) in the critic step.

Actor Step. In the k-th actor step, we update the policy parameter θ and the reward parameter β as follows,

$$\theta_{k+1} = \tau_{k+1}^{-1}\cdot(\tau_k\cdot\theta_k - \eta\cdot\delta_k), \tag{3.13}$$

$$\beta_{k+1} = \mathrm{Proj}_{S_{B_\beta}}\big\{\beta_k + \eta\cdot\widehat{\nabla}_\beta L(\theta_k,\beta_k)\big\}, \tag{3.14}$$

where $\tau_{k+1}=\eta+\tau_k$ and

$$\delta_k \in \mathop{\mathrm{argmin}}_{\delta\in S_{B_\theta}} \big\|\widehat{\mathcal{I}}(\theta_k)\delta - \tau_k\cdot\widehat{\nabla}_\theta L(\theta_k,\beta_k)\big\|_2. \tag{3.15}$$

Here $\eta>0$ is the stepsize, S_{B_θ} and S_{B_β} are the parameter domains of θ and β defined in (3.4) with domain radii B_θ and B_β, respectively, $\mathrm{Proj}_{S_{B_\beta}}:\mathbb{R}^{md}\to S_{B_\beta}$ is the projection operator, τ_k is the inverse temperature parameter of π_{θ_k}, and $\widehat{\mathcal{I}}(\theta_k)$, $\widehat{\nabla}_\theta L(\theta_k,\beta_k)$, $\widehat{\nabla}_\beta L(\theta_k,\beta_k)$ are the estimators of $\mathcal{I}(\theta_k)$, $\nabla_\theta L(\theta_k,\beta_k)$, $\nabla_\beta L(\theta_k,\beta_k)$, respectively, which are defined in the sequel. In (3.13), we update the policy parameter θ_k along the direction δ_k, which approximates the natural policy gradient $\mathcal{I}(\theta)^{-1}\cdot\nabla_\theta L(\theta,\beta)$, and in (3.15) we update the inverse temperature parameter τ_k to ensure that $\theta_{k+1}\in S_{B_\theta}$. Meanwhile, in (3.14), we update the reward parameter β via (projected) gradient ascent. Following from (3.9) and (3.10) of Proposition 3.1, we construct the estimators of $\mathcal{I}(\theta_k)$ and $\nabla_\theta L(\theta_k,\beta_k)$ as follows,

$$\widehat{\mathcal{I}}(\theta_k) = \frac{\tau_k^2}{N}\sum_{i=1}^N \iota_{\theta_k}(s_i,a_i)\,\iota_{\theta_k}(s_i,a_i)^\top, \tag{3.16}$$

$$\widehat{\nabla}_\theta L(\theta_k,\beta_k) = -\frac{\tau_k}{N}\sum_{i=1}^N \widehat{Q}_{\omega_k}(s_i,a_i)\cdot\iota_{\theta_k}(s_i,a_i), \tag{3.17}$$

where $\{(s_i,a_i)\}_{i\in[N]}$ is sampled from the state-action visitation measure $\nu_{\pi_{\theta_k}}$ given θ_k with the batch size N, and $\widehat{Q}_{\omega_k}(s,a)$ is the estimator of $Q^{\pi_{\theta_k}}_{r_{\beta_k}}(s,a)$ computed in the critic step. Meanwhile, following from (3.11) of Proposition 3.1, we construct the estimator of $\nabla_\beta L(\theta_k,\beta_k)$ as follows,

$$\begin{aligned}\widehat{\nabla}_\beta L(\theta,\beta) = {}& \frac{1}{N\cdot(1-\gamma)}\sum_{i=1}^N\big[\phi_{\beta_k}(s_i^{\mathrm{E}},a_i^{\mathrm{E}}) - \phi_{\beta_k}(s_i,a_i)\big]\\ &- \lambda\cdot\nabla_\beta\psi(\beta_k),\end{aligned} \tag{3.18}$$

where $\{(s_i^{\mathrm{E}},a_i^{\mathrm{E}})\}_{i\in[N]}$ is the expert trajectory. For notational simplicity, we write $\pi_k=\pi_{\theta_k}$, $r_k(s,a)=r_{\beta_k}(s,a)$, $d_k=$

d_{π_k} and $\nu_k = \nu_{\pi_k}$ for the k-th step hereafter, where π_θ is the policy, $r_\beta(s,a)$ is the reward function, and d_π, ν_π are the visitation measures defined in (2.3).

Critic Step. Note that the estimator $\widehat{\nabla}_\theta L(\theta,\beta)$ in (3.17) involves the estimator $\widehat{Q}_{\omega_k}(s,a)$ of $Q^{\pi_k}_{r_k}(s,a)$. To this end, we parameterize $\widehat{Q}_\omega(s,a)$ as in (3.12) and adapt neural TD (Cai et al., 2019c), which solves the following minimization problem,

$$\omega_k = \mathop{\mathrm{argmin}}_{\omega \in S_{B_\omega}} \mathbb{E}_{(s,a)\sim\rho_k}\big[\widehat{Q}_\omega(s,a) - \mathcal{T}^{\pi_k}_{r_k}\widehat{Q}_\omega(s,a)\big]^2. \tag{3.19}$$

Here S_{B_ω} is the parameter domain with domain radius B_ω, ρ_k is the state-action stationary distribution induced by π_k, and $\mathcal{T}^{\pi_k}_{r_k}$ is the Bellman operator. Note that the Bellman operator $\mathcal{T}^\pi_r$ is defined as follows,

$$\mathcal{T}^\pi_r Q(s,a) = (1-\gamma)\cdot r(s,a) + \gamma \cdot \mathbb{E}_\pi\big[Q(s',a') \,\big|\, s,a\big],$$

where the expectation is taken with respect to $s' \sim P(\cdot \,|\, s,a)$ and $a' \sim \pi(\cdot \,|\, s')$. In neural TD, we iteratively update the value parameter ω via

$$\begin{aligned}\delta(j) &= Q_{\omega(j)}(s,a) - r(s,a) - \gamma\cdot Q_{\omega(j)}(s',a'),\\ \omega(j+1) &= \mathrm{Proj}_{S_{B_\omega}}\big\{\omega(j) - \alpha\cdot\delta(j)\cdot\nabla_\omega Q_{\omega(j)}(s,a)\big\},\end{aligned} \tag{3.20}$$

where $\delta(j)$ is the Bellman residual, $\alpha > 0$ is the stepsize, (s,a) is sampled from the state-action stationary distribution ρ_k, and $s' \sim P(\cdot \,|\, s,a), a' \sim \pi_k(\cdot \,|\, s')$ are the subsequent state and action. We defer the detailed discussion of neural TD to §B.

To approximately obtain the compatible function approximation (Sutton et al., 2000; Wang et al., 2019), we share the random initialization among the policy π_θ, the reward function $r_\beta(s,a)$, and the state-action value function $\widehat{Q}_\omega(s,a)$. In other words, we set $\theta_0 = \beta_0 = \omega(0) = W_0$ in our algorithm, where W_0 is the random initialization in (3.3). The output of GAIL is the mixed policy $\bar{\pi}$ (Altman, 1999). Specifically, the mixed policy $\bar{\pi}$ of $\pi_0, \ldots, \pi_{T-1}$ is executed by randomly selecting a policy π_k for $k \in [0 : T-1]$ with equal probability before time $t = 0$ and exclusively following π_k thereafter. It then holds for any reward function $r(s,a)$ that

$$J(\bar{\pi}; r) = \frac{1}{T}\sum_{k=0}^{T-1} J(\pi_k; r). \tag{3.21}$$

4. Main Results

In this section, we first present the assumptions for our analysis. Then, we establish the global optimality and convergence of Algorithm 1.

Algorithm 1 GAIL

Input: Expert trajectory $\{(s^{\mathrm{E}}_i, a^{\mathrm{E}}_i)\}_{i\in[T_{\mathrm{E}}]}$, number of iterations T, number of iterations T_{TD} of neural TD, stepsize η, stepsize α of neural TD, batch size N, and domain radii $B_\theta, B_\omega, B_\beta$.

1: **Initialization.** Initialize $b_l \sim \mathrm{Unif}(\{-1,1\})$ and $[W_0]_l \sim N(0, I_d/d)$ for any $l \in [m]$ and set $\tau_0 \leftarrow 0$, $\theta_0 \leftarrow W_0$, and $\beta_0 \leftarrow W_0$.
2: **for** $k = 0, 1, \ldots, T-1$ **do**
3: Update value parameter ω_k via Algorithm 2 with π_k, r_k, W_0, b, T_{TD}, and α as the input.
4: Sample $\{(s_i, a_i)\}_{i=1}^N$ from the state-action visitation measure ν_k, and estimate $\widehat{\mathcal{I}}(\theta_k)$, $\widehat{\nabla}_\theta L(\theta_k, \beta_k)$, and $\widehat{\nabla}_\beta L(\theta_k, \beta_k)$ via (3.16), (3.17), and (3.18), respectively.
5: Solve $\delta_k \leftarrow \mathrm{argmin}_{\delta\in S_\theta}\big\|\widehat{\mathcal{I}}(\theta_k)\cdot\delta - \tau_k\cdot\widehat{\nabla}_\theta L(\theta_k,\beta_k)\big\|_2$ and set $\tau_{k+1} \leftarrow \tau_k + \eta$.
6: Update policy parameter θ via $\theta_{k+1} \leftarrow \tau_{k+1}^{-1}\cdot(\tau_k\cdot\theta_k - \eta\cdot\delta_k)$.
7: Update reward parameter β via $\beta_{k+1} \leftarrow \mathrm{Proj}_{S_{B_\beta}}\{\beta_k + \eta\cdot\widehat{\nabla}_\beta L(\theta_k,\beta_k)\}$.
8: **end for**

Output: Mixed policy $\bar{\pi}$ of $\pi_0, \ldots, \pi_{T-1}$.

4.1. Assumptions

We impose the following assumptions on the stationary distributions $\varrho_k \in \mathcal{P}(\mathcal{S}), \rho_k \in \mathcal{P}(\mathcal{S}\times\mathcal{A})$ and the visitation measures $d_k \in \mathcal{P}(\mathcal{S}), \nu_k \in \mathcal{P}(\mathcal{S}\times\mathcal{A})$.

Assumption 4.1. We assume that the following properties hold.

(a) Let μ be either ρ_k or ν_k. We assume for an absolute constant $c > 0$ and any $y > 0$ and $w \neq 0$ that

$$\mathbb{E}_{(s,a)\sim\mu}\Big[\mathbb{1}\big\{|w^\top(s,a)| \le y\big\}\Big] \le \frac{c\cdot y}{\|w\|_2}.$$

(b) We assume for an absolute constant $C_h > 0$ that

$$\max_{k\in\mathbb{N}}\bigg\{\bigg\|\frac{\mathrm{d}d_{\mathrm{E}}}{\mathrm{d}d_k}\bigg\|_{2,d_k} + \bigg\|\frac{\mathrm{d}\nu_{\mathrm{E}}}{\mathrm{d}\nu_k}\bigg\|_{2,\nu_k}\bigg\} \le C_h,$$

$$\max_{k\in\mathbb{N}}\bigg\{\bigg\|\frac{\mathrm{d}d_{\mathrm{E}}}{\mathrm{d}\varrho_k}\bigg\|_{2,\varrho_k} + \bigg\|\frac{\mathrm{d}\nu_{\mathrm{E}}}{\mathrm{d}\rho_k}\bigg\|_{2,\rho_k}\bigg\} \le C_h.$$

Here $\mathrm{d}d_{\mathrm{E}}/\mathrm{d}d_k$, $\mathrm{d}\nu_{\mathrm{E}}/\mathrm{d}\nu_k$, $\mathrm{d}d_{\mathrm{E}}/\mathrm{d}\varrho_k$, and $\mathrm{d}\nu_{\mathrm{E}}/\mathrm{d}\rho_k$ are the Radon-Nikodym derivatives.

Assumption 4.1 (a) holds when the probability density functions of ρ_k and ν_k are uniformly upper bounded across k. Assumption 4.1 (b) states that the concentrability coefficients are uniformly upper bounded across k, which is commonly used in the analysis of RL (Szepesvári & Munos, 2005; Munos & Szepesvári, 2008; Antos et al., 2008; Farahmand et al., 2010; Scherrer et al., 2015; Farahmand et al., 2016; Lazaric et al., 2016).

For notational simplicity, we write $u_0(s,a) = u_{W_0}(s,a)$ and $\phi_0(s,a) = \phi_{W_0}(s,a)$, where $u_{W_0}(s,a)$ is the neural network defined in (3.1) with $W = W_0$, $\phi_{W_0}(s,a)$ is the feature vector defined in (3.2) with $W = W_0$, and W_0 is the random initialization in (3.3). We impose the following assumptions on the neural network $u_0(s,a)$ and the transition kernel P.

Assumption 4.2. We assume that the following properties hold.

(a) Let $\bar{U} = \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} |u_0(s,a)|$. We assume for absolute constants $M_0 > 0$ and $v > 0$ and any $t > 2M_0$ that

$$\mathbb{E}_{\text{init}}[\bar{U}^2] \le M_0^2, \quad \mathbb{P}(\bar{U} > t) \le \exp(-v \cdot t^2). \quad (4.1)$$

(b) We assume that the transition kernel P belongs to the following class,

$$\widetilde{\mathcal{M}}_{\infty,B_P} = \biggl\{ P(s' \,|\, s,a) = \int \vartheta(s,a;w)^\top \varphi(s';w) \,\mathrm{d}q(w) \,\biggm|\, \sup_w \biggl\| \int \varphi(s;w) \mathrm{d}s \biggr\|_2 \le B_P \biggr\}.$$

Here $B_P > 0$ is an absolute constant, q is the probability density function of $N(0, I_d/d)$, and $\vartheta(s,a;w)$ is defined as $\vartheta(s,a;w) = \mathbb{1}\{w^\top(s,a) > 0\} \cdot (s,a)$.

Assumption 4.2 (b) states that the MDP belongs to (a variant of) the class of linear MDPs (Yang & Wang, 2019a;b; Jin et al., 2019; Cai et al., 2019b). However, our class of transition kernels is infinite-dimensional, and thus, captures a rich class of MDPs. To understand Assumption 4.2 (a), recall that we initialize the neural network with $[W_0]_l \sim N(0, I_d/d)$ and $b_l \sim \text{Unif}(\{-1,1\})$ for any $l \in [m]$. Thus, the neural network $u_0(s,a)$ defined in (3.1) with $W = W_0$ converges to a Gaussian process indexed by $(s,a) \in \mathcal{S}\times\mathcal{A}$ as m goes to infinity. Following from the facts that the maximum of a Gaussian process over a compact index set is sub-Gaussian (van de Geer & Muro, 2014) and that $\mathcal{S}\times\mathcal{A}$ is compact, it is reasonable to assume that $\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} |u_0(s,a)|$ is sub-Gaussian, which further implies the existence of the absolute constants M_0 and v in (4.1) of Assumption 4.2 (a). Moreover, if we assume that m is even and initialize the parameters W_0, b as follows,

$$\begin{cases} [W_0]_l \overset{\text{i.i.d.}}{\sim} N(0, I_d/d), \;\; b_l \sim \text{Unif}\bigl(\{-1,1\}\bigr), \;\; \forall l \le m/2, \\ [W_0]_l = [W_0]_{l-m/2}, \;\; b_l = -b_{l-m/2}, \;\; \forall l > m/2, \end{cases} \quad (4.2)$$

we have that $u_0(s,a) = 0$ for any $(s,a) \in \mathcal{S}\times\mathcal{A}$, which allows us to set $M_0 = 0$ and $v = +\infty$ in Assumption 4.2 (a). Also, it holds that $0 = u_0(s,a) \in \mathcal{R}_\beta$, which implies that $\mathbb{D}_{\mathcal{R}_\beta}(\pi_1, \pi_2) \ge 0$ for any π_1 and π_2. The proof of our results with the random initialization in (4.2) is identical.

Finally, we impose the following assumption on the regularizer $\psi(\beta)$ and the variances of the estimators $\widehat{\mathcal{I}}(\theta)$, $\widehat{\nabla}_\theta L(\theta,\beta)$, and $\widehat{\nabla}_\beta L(\theta,\beta)$ defined in (3.16), (3.17), and (3.18), respectively.

Assumption 4.3. We assume that the following properties hold.

(a) We assume for an absolute constant $\sigma > 0$ that

$$\mathbb{E}_k\Bigl[\Bigl\| \widehat{\mathcal{I}}(\theta_k)W - \mathbb{E}_k\bigl[\widehat{\mathcal{I}}(\theta_k)W\bigr] \Bigr\|_2^2\Bigr] \le \tau_k^4 \sigma^2/N, \quad \forall W \in S_{B_\theta}, \quad (4.3)$$

$$\mathbb{E}_k\Bigl[\Bigl\| \widehat{\nabla}_\theta L(\theta_k,\beta_k) - \mathbb{E}_k\bigl[\widehat{\nabla}_\theta L(\theta_k,\beta_k)\bigr] \Bigr\|_2^2\Bigr] \le \tau_k^2 \sigma^2/N, \quad (4.4)$$

$$\mathbb{E}_k\Bigl[\Bigl\| \widehat{\nabla}_\beta L(\theta_k,\beta_k) - \mathbb{E}_k\bigl[\widehat{\nabla}_\beta L(\theta_k,\beta_k)\bigr] \Bigr\|_2^2\Bigr] \le \sigma^2/N, \quad (4.5)$$

where τ_k is the inverse temperature parameter in (3.5), $N \in \mathbb{N}_+$ is the batch size, and S_{B_θ} is the parameter domain of θ defined in (3.4) with the domain radius B_θ. Here the expectation $\mathbb{E}_k$ is taken with respect to the k-th batch, which is drawn from ν_k given θ_k.

(b) We assume that the regularizer $\psi(\beta)$ in (2.4) is convex and L_ψ-Lipschitz continuous over the compact parameter domain S_{B_β}.

Assumption 4.3 (a) holds when $\widehat{Q}_{\omega_k}(s_i,a_i) \cdot \iota_{\theta_k}(s_i,a_i)$, $\iota_{\theta_k}(s_i,a_i)\iota_{\theta_k}(s_i,a_i)^\top$, and $\phi_{\beta_k}(s_i,a_i)$ have uniformly upper bounded variances across $i \in [m]$ and k, and the Markov chain that generates $\{(s_i,a_i)\}_{i\in[N]}$ mixes sufficiently fast (Zhang et al., 2019a). Similar assumptions are also used in the analysis of policy optimization (Xu et al., 2019a;b). Also, Assumption 4.3 (b) holds for most commonly used regularizers (Ho & Ermon, 2016).

4.2. Global Optimality and Convergence

In this section, we establish the global optimality and convergence of Algorithm 1. The following proposition adapted from (Cai et al., 2019c) characterizes the global optimality and convergence of neural TD, which is presented in Algorithm 2.

Proposition 4.4 (Global Optimality and Convergence of Neural TD)**.** In Algorithm 2, we set $T_{\text{TD}} = \Omega(m)$, $\alpha =$

$\min\{(1-\gamma)/8, m^{-1/2}\}$, and $B_\omega = c \cdot (B_\beta + B_P \cdot (M_0 + B_\beta))$ for an absolute constant $c > 0$. Let π_k, r_k be the input and ω_k be the output of Algorithm 2. Under Assumptions 4.1 and 4.2, it holds for an absolute constant $C_v > 0$ that

$$\mathbb{E}_{\text{init}}\Big[\big\|Q_{\omega_k}(s,a) - Q^{\pi_k}_{r_k}(s,a)\big\|^2_{2,\rho_k}\Big] \\ = \mathcal{O}\big(B_\omega^3 \cdot m^{-1/2} + B_\omega^{5/2} \cdot m^{-1/4} + B_\omega^2 \cdot \exp(-C_v \cdot B_\omega^2)\big). \tag{4.6}$$

Here the expectation $\mathbb{E}_{\text{init}}$ is taken with respect to the random initialization in (3.3).

Proof. See §B.1 for a detailed proof. □

The term $B_\omega^2 \cdot \exp(-C_v \cdot B_\omega^2)$ in (4.6) of Proposition 4.4 characterizes the hardness of estimating the state-action value function $Q^{\pi_k}_{r_k}(s,a)$ by the neural network defined in (3.1), which arises because $\|Q^{\pi_k}_{r_k}(s,a)\|_\infty$ is not uniformly upper bounded across k. Note that if we employ the random initialization in (4.2), we have that $C_v = +\infty$. And consequently, such a term vanishes. We are now ready to establish the global optimality and convergence of Algorithm 1.

Theorem 4.5 (Global Optimality and Convergence of GAIL)**.** We set $\eta = 1/\sqrt{T}$ and $B_\omega = c \cdot (B_\beta + B_P \cdot (M_0 + B_\beta))$ for an absolute constant $c > 0$, and $B_\theta = B_\omega$ in Algorithm 1. Let $\bar\pi$ be the output of Algorithm 1. Under Assumptions 4.1-4.3, it holds that

$$\mathbb{E}\big[\mathbb{D}_{\mathcal{R}_\beta}(\pi_{\mathrm{E}}, \bar\pi)\big] \le \underbrace{\frac{(1-\gamma)^{-1}\cdot\log|\mathcal{A}| + 13\bar B^2 + M_0^2 + 8}{\sqrt{T}}}_{\text{(i)}} \\ + \underbrace{2\lambda \cdot L_\psi \cdot \bar B}_{\text{(ii)}} + \underbrace{\frac{1}{T}\sum_{k=0}^{T-1}\varepsilon_k}_{\text{(iii)}}. \tag{4.7}$$

Here $\bar B = \max\{B_\theta, B_\omega, B_\beta\}$, $\mathbb{D}_{\mathcal{R}_\beta}$ is the $\mathcal{R}_\beta$-distance defined in (2.5) with $\mathcal{R}_\beta = \{r_\beta(s,a)\,|\,\beta \in S_{B_\beta}\}$, the expectation is taken with respect to the random initialization in (3.3) and the T batches, and the error term ε_k satisfies that

$$\varepsilon_k = \underbrace{2\sqrt{2}\cdot C_h \cdot \bar B \cdot \sigma \cdot N^{-1/2}}_{\text{(iii.a)}} + \underbrace{\epsilon_{Q,k}}_{\text{(iii.b)}} \\ + \underbrace{\mathcal{O}(k\cdot \bar B^{3/2}\cdot m^{-1/4} + \bar B^{5/4}\cdot m^{-1/8})}_{\text{(iii.c)}}, \tag{4.8}$$

where C_h is defined in Assumption 4.1, L_ψ and σ are defined in Assumption 4.3, and $\epsilon_{Q,k} = \mathcal{O}(B_\omega^3 \cdot m^{-1/2} + B_\omega^{5/2} \cdot m^{-1/4} + B_\omega^2 \cdot \exp(-C_v \cdot B_\omega^2))$ is the error induced by neural TD (Algorithm 2).

Proof. See §5 for a detailed proof. □

The optimality gap in (4.7) of Theorem 4.5 is measured by the expected $\mathcal{R}_\beta$-distance $\mathbb{D}_{\mathcal{R}_\beta}(\pi_{\mathrm{E}}, \bar\pi)$ between the expert policy π_{E} and the learned policy $\bar\pi$. Thus, by showing that the optimality gap is upper bounded by $\mathcal{O}(1/\sqrt{T})$, we prove that $\bar\pi$ (approximately) outperforms the expert policy π_{E} in expectation when the number of iterations T goes to infinity. As shown in (4.7) of Theorem 4.5, the optimality gap is upper bounded by the sum of the three terms. Term (i) corresponds to the $1/\sqrt{T}$ rate of convergence of Algorithm 1. Term (ii) corresponds to the bias induced by the regularizer $\lambda \cdot \psi(\beta)$ in the objective function $L(\theta, \beta)$ defined in (2.4). Term (iii) is upper bounded by the sum of the three terms in (4.8) of Theorem 4.5. In detail, term (iii.a) corresponds to the error induced by the variances of $\widehat{\mathcal{I}}(\theta)$, $\widehat\nabla_\theta L(\theta,\beta)$, and $\widehat\nabla_\beta L(\theta,\beta)$ defined in (4.3), (4.4), and (4.5) of Assumption 4.3, which vanishes as the batch size N in Algorithm 1 goes to infinity. Term (iii.b) is the error of estimating $Q^\pi_r(s,a)$ by $\widehat Q_\omega(s,a)$ using neural TD (Algorithm 2). As shown in Proposition 4.4, the estimation error $\epsilon_{Q,k}$ vanishes as m and B_ω go to infinity. Term (iii.c) corresponds to the linearization error of the neural network defined in (3.1), which is characterized in Lemma A.2. Following from Theorem 4.5, it holds for $B_\omega = \Omega((C_v^{-1}\cdot\log T)^{1/2})$, $m = \Omega(\bar B^{10}\cdot T^6)$, and $N = \Omega(\bar B^2 \cdot T \cdot \sigma^2)$ that $\mathbb{E}\big[\mathbb{D}_{\mathcal{R}_\beta}(\pi_{\mathrm{E}}, \bar\pi)\big] = \mathcal{O}(T^{-1/2} + \lambda)$, which implies the $1/\sqrt{T}$ rate of convergence of Algorithm 1 (up to the bias induced by the regularizer).

5. Proof of Main Results

In this section, we present the proof of Theorem 4.5, which establishes the global optimality and convergence of Algorithm 1. For notational simplicity, we write $\pi^s(a) = \pi(a\,|\,s)$ for any policy π, state $s \in \mathcal{S}$, and action $a \in \mathcal{A}$. For any policies π_1, π_2 and distribution μ over $\mathcal{S}$, we denote the expected Kullback-Leibler (KL) divergence by KL^μ, which is defined as $\mathrm{KL}^\mu(\pi_1\,\|\,\pi_2) = \mathbb{E}_{s\sim\mu}[\mathrm{KL}(\pi_1^s\,\|\,\pi_2^s)]$. For any visitation measures $d_\pi \in \mathcal{P}(\mathcal{S})$ and $\nu_\pi \in \mathcal{P}(\mathcal{S}\times\mathcal{A})$, we denote by $\mathbb{E}_{d_\pi}$ and $\mathbb{E}_{\nu_\pi}$ the expectations taken with respect to $s \sim d_\pi$ and $(s,a)\sim\nu_\pi$, respectively.

Following from the property of the mixed policy $\bar\pi$ in (3.21), we have that

$$\mathbb{E}\big[\mathbb{D}_{\mathcal{R}_\beta}(\pi_{\mathrm{E}}, \bar\pi)\big] = \mathbb{E}\big[\max_{\beta'\in S_{B_\beta}} J(\pi_{\mathrm{E}}; r_{\beta'}) - J(\bar\pi; r_{\beta'})\big] \\ = \mathbb{E}\Big[\max_{\beta'\in S_{B_\beta}} \frac{1}{T}\sum_{k=0}^{T-1} J(\pi_{\mathrm{E}}; r_{\beta'}) - J(\pi_k; r_{\beta'})\Big]. \tag{5.1}$$

We now upper bound the optimality gap in (5.1) by upper bounding the following difference of expected cumulative rewards,

$$J(\pi_{\mathrm{E}}; r_{\beta'}) - J(\pi_k; r_{\beta'}) = \underbrace{J(\pi_{\mathrm{E}}; r_k) - J(\pi_k; r_k)}_{\text{(i)}} + \underbrace{L(\theta_k, \beta') - L(\theta_k, \beta_k)}_{\text{(ii)}} + \underbrace{\lambda \cdot \big(\psi(\beta') - \psi(\beta_k)\big)}_{\text{(iii)}}, \quad (5.2)$$

where $\beta' \in S_{B_\beta}$ is chosen arbitrarily and $L(\theta, \beta)$ is the objective function defined in (2.4). Following from Assumption 4.3 and the fact that $\beta_k, \beta' \in S_{B_\beta}$, we have that

$$\lambda\big(\psi(\beta') - \psi(\beta_k)\big) \leq \lambda L_\psi \|\beta' - \beta_k\|_2 \leq 2\lambda L_\psi B_\beta, \quad (5.3)$$

which upper bounds term (iii) of (5.2). It remains to upper bound terms (i) and (ii) of (5.2), which hinges on the one-point convexity of $J(\pi; r)$ with respect to π and the (approximate) convexity of $L(\theta, \beta)$ with respect to β.

Upper bound of term (i) in (5.2). In what follows, we upper bound term (i) of (5.2). We first introduce the following cost difference lemma (Kakade & Langford, 2002), which corresponds to the one-point convexity of $J(\pi; r)$ with respect to π. Recall that $d_{\mathrm{E}} \in \mathcal{P}(\mathcal{S})$ is the state visitation measure induced by the expert policy π_{E}.

Lemma 5.1 (Cost Difference Lemma, Lemma 6.1 in (Kakade & Langford, 2002))**.** For any policy π and reward function $r(s, a)$, it holds that

$$J(\pi_{\mathrm{E}}; r) - J(\pi; r) = (1-\gamma)^{-1} \mathbb{E}_{d_{\mathrm{E}}}\Big[\big\langle Q_r^\pi(s, \cdot), \pi_{\mathrm{E}}^s - \pi^s \big\rangle_{\mathcal{A}}\Big], \quad (5.4)$$

where γ is the discount factor.

Furthermore, we establish the following lemma, which upper bounds the right-hand side of (5.4) in Lemma 5.1.

Lemma 5.2. Under Assumptions 4.1-4.3, we have that

$$\mathbb{E}_{d_{\mathrm{E}}}\Big[\big\langle Q_{r_k}^{\pi_k}(s, \cdot), \pi_{\mathrm{E}}^s - \pi_k^s \big\rangle_{\mathcal{A}}\Big] = \eta^{-1} \cdot \mathrm{KL}^{d_{\mathrm{E}}}(\pi_{\mathrm{E}} \,\|\, \pi_k) - \eta^{-1} \cdot \mathrm{KL}^{d_{\mathrm{E}}}(\pi_{\mathrm{E}} \,\|\, \pi_{k+1}) + \Delta_k^{\text{(i)}},$$

where

$$\mathbb{E}\big[|\Delta_k^{\text{(i)}}|\big] = 2\sqrt{2} \cdot C_h \cdot B_\theta^{1/2} \cdot \sigma^{1/2} \cdot N^{-1/4} + \eta \cdot (M_0^2 + 9B_\theta^2) + \epsilon_{Q,k} + \mathcal{O}(\eta^{-1} \cdot \tau_{k+1} \cdot B_\theta^{3/2} \cdot m^{-1/4} + B_\theta^{5/4} \cdot m^{-1/8}). \quad (5.5)$$

Here M_0 is defined in Assumption 4.2, σ is defined in Assumption 4.3, N is the batch size in (3.16)-(3.18), and $\epsilon_{Q,k} = \mathcal{O}(B_\omega^3 \cdot m^{-1/2} + B_\omega^{5/2} \cdot m^{-1/4} + B_\omega^2 \cdot \exp(-C_v \cdot B_\omega^2))$ for an absolute constant $C_v > 0$, which depends on the absolute constant v in Assumption 4.2.

Proof. See §C.2 for a detailed proof. □

Combining Lemmas 5.1 and 5.2, we have that

$$J(\pi_{\mathrm{E}}; r_k) - J(\pi_k; r_k) \leq \frac{\mathrm{KL}^{d_{\mathrm{E}}}(\pi_{\mathrm{E}} \,\|\, \pi_k) - \mathrm{KL}^{d_{\mathrm{E}}}(\pi_{\mathrm{E}} \,\|\, \pi_{k+1}) + \eta \cdot \Delta_k^{\text{(i)}}}{\eta \cdot (1-\gamma)}, \quad (5.6)$$

which upper bounds term (i) of (5.2). Here $\Delta_k^{\text{(i)}}$ is upper bounded in (5.5) of Lemma 5.2.

Upper bound of term (ii) in (5.2). In what follows, we upper bound term (ii) of (5.2). We first establish the following lemma, which characterizes the (approximate) convexity of $L(\theta, \beta)$ with respect to β.

Lemma 5.3. Under Assumption 4.1, it holds for any $\beta' \in S_{B_\beta}$ that

$$\mathbb{E}_{\text{init}}\big[L(\theta_k, \beta') - L(\theta_k, \beta_k)\big] = \mathbb{E}_{\text{init}}\big[\nabla_\beta L(\theta_k, \beta_k)^\top (\beta' - \beta_k)\big] + \mathcal{O}(B_\beta^{3/2} \cdot m^{-1/4}).$$

Proof. See §C.3 for a detailed proof. □

The term $\mathcal{O}(B_\beta^{3/2} \cdot m^{-1/4})$ in Lemma 5.3 arises from the linearization error of the neural network, which is characterized in Lemma A.2. Based on Lemma 5.3, we establish the following lemma, which upper bounds term (ii) of (5.2).

Lemma 5.4. Under Assumptions 4.1 and 4.3, we have that

$$L(\theta_k, \beta') - L(\theta_k, \beta_k) \leq \eta^{-1} \cdot \|\beta_k - \beta'\|_2^2 - \eta^{-1} \cdot \|\beta_{k+1} - \beta'\|_2^2 + \Delta_k^{\text{(ii)}},$$

where

$$\mathbb{E}\big[|\Delta_k^{\text{(ii)}}|\big] = \eta \cdot \Big(2\big(2(1-\gamma)^{-1} + \lambda \cdot L_\psi\big)^2 + \sigma^2 \cdot N^{-1}\Big) + 2B_\beta \cdot \sigma \cdot N^{-1/2} + \mathcal{O}(B_\beta^{3/2} \cdot m^{-1/4}). \quad (5.7)$$

Proof. See §C.4 for a detailed proof. □

By Lemma 5.4, we have that

$$L(\theta_k, \beta') - L(\theta_k, \beta_k) \leq \Delta_k^{\text{(ii)}} + \eta^{-1} \cdot \big(\|\beta_k - \beta'\|_2^2 - \|\beta_{k+1} - \beta'\|_2^2\big), \quad (5.8)$$

which upper bounds term (ii) of (5.2). Here $\Delta_k^{\text{(ii)}}$ is upper bounded in (5.7) of Lemma 5.4.

Plugging (5.3), (5.6), and (5.8) into (5.2), we obtain that

$$J(\pi_{\mathrm{E}}; r_{\beta'}) - J(\pi_k; r_{\beta'}) \leq \frac{\mathrm{KL}^{d_{\mathrm{E}}}(\pi_{\mathrm{E}} \,\|\, \pi_k) - \mathrm{KL}^{d_{\mathrm{E}}}(\pi_{\mathrm{E}} \,\|\, \pi_{k+1})}{\eta \cdot (1-\gamma)} + \eta^{-1}\|\beta_k - \beta'\|_2^2 - \eta^{-1}\|\beta_{k+1} - \beta'\|_2^2 + 2\lambda \cdot L_\psi \cdot B_\beta + \Delta_k. \quad (5.9)$$

Here $\Delta_k = \Delta_k^{(\mathrm{i})} + \Delta_k^{(\mathrm{ii})}$, where $\Delta_k^{(\mathrm{i})}$ and $\Delta_k^{(\mathrm{ii})}$ are upper bounded in (5.5) and (5.7) of Lemmas 5.2 and 5.4, respectively. Note that the upper bound of Δ_k does not depend on θ and β. Upon telescoping (5.9) with respect to k, we obtain that

$$J(\pi_{\mathrm{E}}; r_{\beta'}) - J(\bar{\pi}; r_{\beta'}) = \frac{1}{T}\sum_{k=0}^{T-1}\big[J(\pi_{\mathrm{E}}; r_{\beta'}) - J(\pi_k; r_{\beta'})\big]$$
$$\leq \frac{(1-\gamma)^{-1}\cdot \mathrm{KL}^{d_{\mathrm{E}}}(\pi_{\mathrm{E}}\,\|\,\pi_0) + \|\beta_0 - \beta'\|_2^2}{\eta\cdot T}$$
$$+ 2\lambda\cdot L_\psi \cdot B_\beta + \frac{1}{T}\sum_{k=0}^{T-1}|\Delta_k|. \qquad (5.10)$$

Following from the fact that $\tau_0 = 0$ and the parameterization of π_θ in (3.5), it holds that π_0^s is the uniform distribution over $\mathcal{A}$ for any $s \in \mathcal{S}$. Thus, we have $\mathrm{KL}^{d_{\mathrm{E}}}(\pi_{\mathrm{E}}\,\|\,\pi_0) \leq \log|\mathcal{A}|$. Meanwhile, following from the fact that $\beta' \in S_{B_\beta}$, it holds that $\|\beta' - \beta_0\|_2 \leq B_\beta$. Finally, by setting $\eta = T^{-1/2}$, $\tau_k = k\cdot\eta$, and $\bar{B} = \max\{B_\theta, B_\beta\}$ in (5.10), we have that

$$\mathbb{E}\big[\mathbb{D}_{\mathcal{R}_\beta}(\pi_{\mathrm{E}}, \bar{\pi})\big] = \mathbb{E}\big[\max_{\beta'\in S_{B_\beta}} J(\pi_{\mathrm{E}}; r_{\beta'}) - J(\bar{\pi}; r_{\beta'})\big]$$
$$\leq \frac{(1-\gamma)^{-1}\cdot\log|\mathcal{A}| + 4B_\beta^2}{\eta\cdot T} + 2\lambda\cdot L_\psi\cdot B_\beta$$
$$+ \frac{\mathbb{E}\big[\max_{\beta'}\sum_{k=0}^{T-1}|\Delta_k|\big]}{T}$$
$$= \frac{(1-\gamma)^{-1}\cdot\log|\mathcal{A}| + 13\bar{B}^2 + M_0^2 + 8}{\sqrt{T}} + 2\lambda\cdot L_\psi\cdot\bar{B}$$
$$+ \frac{\sum_{k=0}^{T-1}\varepsilon_k}{T}.$$

Here ε_k is upper bounded as follows,

$$\varepsilon_k = 2\sqrt{2}\cdot C_h\cdot\bar{B}\cdot\sigma\cdot N^{-1/2} + \epsilon_{Q,k}$$
$$+ \mathcal{O}(k\cdot\bar{B}^{3/2}\cdot m^{-1/4} + \bar{B}^{5/4}\cdot m^{-1/8}),$$

where $\epsilon_{Q,k} = \mathcal{O}(B_\omega^3\cdot m^{-1/2} + B_\omega^{5/2}\cdot m^{-1/4} + B_\omega^2\cdot\exp(-C_v\cdot B_\omega^2))$ for an absolute constant $C_v > 0$. Thus, we complete the proof of Theorem 4.5.

References

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. Optimality and approximation with policy gradient methods in Markov decision processes. *arXiv preprint arXiv:1908.00261*, 2019.

Altman, E. *Constrained Markov decision processes*, volume 7. CRC Press, 1999.

Anthony, M. and Bartlett, P. L. *Neural network learning: Theoretical foundations*. Cambridge University Press, 2009.

Antos, A., Szepesvári, C., and Munos, R. Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129, 2008.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017.

Bhandari, J. and Russo, D. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019.

Cai, Q., Hong, M., Chen, Y., and Wang, Z. On the global convergence of imitation learning: A case for linear quadratic regulator. *arXiv preprint arXiv:1901.03674*, 2019a.

Cai, Q., Yang, Z., Jin, C., and Wang, Z. Provably efficient exploration in policy optimization. *arXiv preprint arXiv:1912.05830*, 2019b.

Cai, Q., Yang, Z., Lee, J. D., and Wang, Z. Neural temporal-difference learning converges to global optima. *arXiv preprint arXiv:1905.10027*, 2019c.

Chen, M., Wang, Y., Liu, T., Yang, Z., Li, X., Wang, Z., and Zhao, T. On computation and generalization of generative adversarial imitation learning. *arXiv preprint arXiv:2001.02792*, 2020.

Farahmand, A.-m., Szepesvári, C., and Munos, R. Error propagation for approximate policy and value iteration. In *Advances in Neural Information Processing Systems*, pp. 568–576, 2010.

Farahmand, A.-m., Ghavamzadeh, M., Szepesvári, C., and Mannor, S. Regularized policy iteration with nonparametric function spaces. *The Journal of Machine Learning Research*, 17(1):4809–4874, 2016.

Finn, C., Levine, S., and Abbeel, P. Guided cost learning: Deep inverse optimal control via policy optimization. In *International Conference on Machine Learning*, pp. 49–58, 2016.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

Ho, J. and Ermon, S. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*, pp. 4565–4573, 2016.

Hofmann, T., Schölkopf, B., and Smola, A. J. Kernel methods in machine learning. *The Annals of Statistics*, pp. 1171–1220, 2008.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. *arXiv preprint arXiv:1907.05388*, 2019.

Kakade, S. and Langford, J. Approximately optimal approximate reinforcement learning. In *International Conference on Machine Learning*, volume 2, pp. 267–274, 2002.

Kakade, S. M. A natural policy gradient. In *Advances in Neural Information Processing Systems*, pp. 1531–1538, 2002.

Kuefler, A., Morton, J., Wheeler, T., and Kochenderfer, M. Imitating driver behavior with generative adversarial networks. In *IEEE Intelligent Vehicles Symposium*, pp. 204–211. IEEE, 2017.

Lazaric, A., Ghavamzadeh, M., and Munos, R. Analysis of classification-based policy iteration algorithms. *The Journal of Machine Learning Research*, 17(1):583–612, 2016.

Levine, S. and Koltun, V. Continuous inverse optimal control with locally optimal examples. *arXiv preprint arXiv:1206.4617*, 2012.

Liu, B., Cai, Q., Yang, Z., and Wang, Z. Neural proximal/trust region policy optimization attains globally optimal policy. *arXiv preprint arXiv:1906.10306*, 2019.

Merel, J., Tassa, Y., TB, D., Srinivasan, S., Lemmon, J., Wang, Z., Wayne, G., and Heess, N. Learning human behaviors from motion capture by adversarial imitation. *arXiv preprint arXiv:1707.02201*, 2017.

Munos, R. and Szepesvári, C. Finite-time bounds for fitted value iteration. *The Journal of Machine Learning Research*, 9(May):815–857, 2008.

Nesterov, Y. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Ng, A. Y. and Russell, S. J. Algorithms for inverse reinforcement learning. In *International Conference on Machine Learning*, pp. 663–670, 2000.

Peters, J. and Schaal, S. Natural actor-critic. *Neurocomputing*, 71(7-9):1180–1190, 2008.

Pomerleau, D. A. Efficient training of artificial neural networks for autonomous navigation. *Neural Computation*, 3(1):88–97, 1991.

Rafique, H., Liu, M., Lin, Q., and Yang, T. Non-convex min-max optimization: Provable algorithms and applications in machine learning. *arXiv:1810.02060*, 2018.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, pp. 1177–1184, 2008.

Rahimi, A. and Recht, B. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. *Advances in Neural Information Processing Systems*, pp. 1313–1320, 2009.

Ross, S. and Bagnell, D. Efficient reductions for imitation learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 661–668, 2010.

Ross, S., Gordon, G., and Bagnell, D. A reduction of imitation learning and structured prediction to no-regret online learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 627–635, 2011.

Russell, S. Learning agents for uncertain environments. In *Conference on Learning Theory*, pp. 101–103, 1998.

Scherrer, B., Ghavamzadeh, M., Gabillon, V., Lesner, B., and Geist, M. Approximate modified policy iteration and its application to the game of Tetris. *The Journal of Machine Learning Research*, 16:1629–1676, 2015.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT Press, 2018.

Sutton, R. S., McAllester, D. A., Singh, S. P., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, pp. 1057–1063, 2000.

Syed, U., Bowling, M., and Schapire, R. E. Apprenticeship learning using linear programming. In *International Conference on Machine Learning*, pp. 1032–1039, 2008.

Szepesvári, C. and Munos, R. Finite time bounds for sampling based fitted value iteration. In *International Conference on Machine Learning*, pp. 880–887. ACM, 2005.

Tai, L., Zhang, J., Liu, M., and Burgard, W. Socially compliant navigation through raw depth inputs with generative adversarial imitation learning. In *IEEE International Conference on Robotics and Automation*, pp. 1111–1117. IEEE, 2018.

van de Geer, S. and Muro, A. On higher order isotropy conditions and lower bounds for sparse quadratic forms. *Electronic Journal of Statistics*, 8(2):3031–3061, 2014. doi: 10.1214/15-EJS983.

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

Xu, P., Gao, F., and Gu, Q. An improved convergence analysis of stochastic variance-reduced policy gradient. *arXiv preprint arXiv:1905.12615*, 2019a.

Xu, P., Gao, F., and Gu, Q. Sample efficient policy gradient methods with recursive variance reduction. *arXiv preprint arXiv:1909.08610*, 2019b.

Yang, L. and Wang, M. Sample-optimal parametric Q-learning using linearly additive features. In *International Conference on Machine Learning*, pp. 6995–7004, 2019a.

Yang, L. F. and Wang, M. Reinforcement leaning in feature space: Matrix bandit, kernels, and regret bound. *arXiv preprint arXiv:1905.10389*, 2019b.

Yu, L., Zhang, W., Wang, J., and Yu, Y. SeqGAN: Sequence generative adversarial nets with policy gradient. *arXiv preprint arXiv:1609.05473*, 2016.

Zhang, K., Koppel, A., Zhu, H., and Başar, T. Global convergence of policy gradient methods to (almost) locally optimal policies. *arXiv preprint arXiv:1906.08383*, 2019a.

Zhang, K., Yang, Z., and Baar, T. Policy optimization provably converges to Nash equilibria in zero-sum linear quadratic games. *arXiv preprint arXiv:1906.00729*, 2019b.

A. Neural Networks

In what follows, we present the properties of the neural network defined in (3.1). First, we define the following function class.

Definition A.1 (Function Class)**.** For $B > 0$ and $m \in \mathbb{N}_+$, we define

$$\mathcal{F}_{B,m} = \big\{ W^\top \phi_0(s,a) \,\big|\, W \in \mathbb{R}^{md},\ \|W - W_0\|_2 \le B \big\},$$

where $\phi_0(s,a)$ is the feature vector defined in (3.2) with $W = W_0$.

As shown in (Rahimi & Recht, 2008), the feature $\phi_0(s,a)$ induces a reproducing kernel Hilbert space (RKHS), namely $\mathcal{H}$. When m goes to infinity, $\mathcal{F}_{B,m}$ approximates a ball in $\mathcal{H}$, which captures a rich class of functions (Hofmann et al., 2008; Rahimi & Recht, 2008). Furthermore, we obtain the following lemma from (Cai et al., 2019c), which characterizes the linearization error of the neural network defined in (3.1).

Lemma A.2 (Linearization Error, Lemma 5.1 in (Cai et al., 2019c))**.** Under Assumption 4.1, it holds for any $W, W_1, W_2 \in S_B$ that,

$$\mathbb{E}_{\text{init}}\Big[\big\|W^\top \phi_{W_1}(s,a) - W^\top \phi_{W_2}(s,a)\big\|_{2,\mu}^2\Big] = \mathcal{O}(B^3 \cdot m^{-1/2}),$$
$$\mathbb{E}_{\text{init}}\Big[\big\|W^\top \phi_{W_1}(s,a) - W^\top \phi_{W_2}(s,a)\big\|_{1,\mu}\Big] = \mathcal{O}(B^{3/2} \cdot m^{-1/4}),$$

where $\phi_W(s,a)$ is the feature vector defined in (3.2) and $\mu \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$ is a distribution that satisfies Assumption 4.1.

Proof. See §A.1 for a detailed proof. □

Following from Lemma A.2, the function class $\mathcal{F}_{B,m}$ defined in Definition A.1 is a first-order approximation of the class of the neural networks defined in (3.1). Meanwhile, we establish the following lemma to characterize the sub-Gaussian property of the neural network defined in (3.1).

Lemma A.3. Under Assumption 4.2, for any $W, W' \in S_B$, it holds that $\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} |W^\top \phi_{W'}(s,a)|$ is sub-Gaussian, where the randomness comes from the random initialization W_0 in the definition of S_B in (3.4). Moreover, it holds that

$$\mathbb{E}_{\text{init}}\Big[\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \big|W^\top \phi_{W'}(s,a)\big|^2\Big] \le 2M_0^2 + 18B^2$$

and that

$$\mathbb{P}\Big(\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \big|W^\top \phi_{W'}(s,a)\big| > t\Big) \le \exp(-v \cdot t^2/2), \quad \forall t > 2M_0 + 6B.$$

Proof. See §A.2 for a detailed proof. □

A.1. Proof of Lemma A.2

Proof. We consider any $W, W' \in S_B$. By the definition of $\phi_W(s,a)$ in (3.2) and the triangle inequality, we have that

$$\begin{aligned}&\big|W^\top \phi_{W'}(s,a) - W^\top \phi_0(s,a)\big| \\ &\quad \le \frac{1}{\sqrt{m}} \sum_{l=1}^m \big|[W]_l^\top (s,a)\big| \cdot \Big|\mathbb{1}\big\{(s,a)^\top [W']_l > 0\big\} - \mathbb{1}\big\{(s,a)^\top [W_0]_l > 0\big\}\Big|.\end{aligned} \tag{A.1}$$

We now upper bound the right-hand side of (A.1). For the term $|[W]_l^\top (s,a)|$ in (A.1), we have that

$$\begin{aligned}\big|[W]_l^\top (s,a)\big| &\le \big|[W_0]_l^\top (s,a)\big| + \Big|\big([W]_l - [W_0]_l\big)^\top (s,a)\Big| \\ &\le \big|[W_0]_l^\top (s,a)\big| + \big\|[W]_l - [W_0]_l\big\|_2,\end{aligned} \tag{A.2}$$

where the first inequality follows from the triangle inequality and the second inequality follows from the Cauchy-Schwartz inequality and the fact that $\|(s,a)\|_2 \le 1$. To upper bound the term $|\mathbb{1}\{(s,a)^\top [W']_l > 0\} - \mathbb{1}\{(s,a)^\top [W_0]_l > 0\}|$ on the right-hand side of (A.1), note that $\mathbb{1}\{(s,a)^\top [W']_l > 0\} \ne \mathbb{1}\{(s,a)^\top [W_0]_l > 0\}$ implies that

$$\big|[W_0]_l^\top (s,a)\big| \le \big|[W']_l^\top (s,a) - [W_0]_l^\top (s,a)\big| \le \big\|[W']_l - [W_0]_l\big\|_2.$$

Thus, we have that

$$\Big|\mathbb{1}\big\{(s,a)^\top [W']_l > 0\big\} - \mathbb{1}\big\{(s,a)^\top [W_0]_l > 0\big\}\Big| \le \mathbb{1}\Big\{\big|(s,a)^\top [W_0]_l\big| \le \big\|[W']_l - [W_0]_l\big\|_2\Big\}. \qquad \text{(A.3)}$$

Plugging (A.2) and (A.3) into (A.1), we have that

$$\begin{aligned}
&\big|W^\top \phi_{W'}(s,a) - W^\top \phi_0(s,a)\big| \\
&\le \frac{1}{\sqrt{m}} \sum_{l=1}^m \mathbb{1}\Big\{\big|(s,a)^\top [W_0]_l\big| \le \big\|[W']_l - [W_0]_l\big\|_2\Big\} \cdot \Big(\big|(s,a)^\top [W_0]_l\big| + \big\|[W]_l - [W_0]_l\big\|_2\Big) \\
&\le \frac{1}{\sqrt{m}} \sum_{l=1}^m \mathbb{1}\Big\{\big|(s,a)^\top [W_0]_l\big| \le \big\|[W']_l - [W_0]_l\big\|_2\Big\} \cdot \Big(\big\|[W']_l - [W_0]_l\big\|_2 + \big\|[W]_l - [W_0]_l\big\|_2\Big).
\end{aligned}$$

By the fact that $W, W' \in S_B$, we obtain that

$$\big|W^\top \phi_{W'}(s,a) - W^\top \phi_0(s,a)\big|^2 \le \frac{4B^2}{m} \sum_{l=1}^m \mathbb{1}\Big\{\big|(s,a)^\top [W_0]_l\big| \le \big\|[W']_l - [W_0]_l\big\|_2\Big\}.$$

By setting $y = \|[W']_l - [W_0]_l\|_2$ in Assumption 4.1, we have that

$$\big\|W^\top \phi_{W'}(s,a) - W^\top \phi_0(s,a)\big\|_{2,\mu}^2 \le \frac{8B^2}{m} \sum_{l=1}^m \frac{c \cdot \big\|[W']_l - [W_0]_l\big\|_2}{\big\|[W_0]_l\big\|_2}.$$

Taking the expectation with respect to the random initialization in (3.3) and using the Cauchy-Schwartz inequality, we have that

$$\begin{aligned}
&\mathbb{E}_{\text{init}}\Big[\big\|W^\top \phi_{W'}(s,a) - W^\top \phi_0(s,a)\big\|_{2,\mu}^2\Big] \\
&\le \mathbb{E}_{\text{init}}\bigg[\frac{8cB^2}{m}\Big(\sum_{l=1}^m \big\|[W']_l - [W_0]_l\big\|_2^2\Big)^{1/2} \cdot \Big(\sum_{l=1}^m 1/\big\|[W_0]_l\big\|_2^2\Big)^{1/2}\bigg] \\
&\le \frac{8cB^3}{m} \mathbb{E}_{\text{init}}\bigg[\Big(\sum_{l=1}^m 1/\big\|[W_0]_l\big\|_2^2\Big)^{1/2}\bigg] \\
&\le \frac{8cB^3}{\sqrt{m}} \Big(\mathbb{E}_{w \sim N(0, I_d/d)}\big[1/\|w\|_2^2\big]\Big)^{1/2} \\
&= \mathcal{O}(B^3 \cdot m^{-1/2}),
\end{aligned}$$

where the second inequality follows from the fact that $\|W' - W_0\|_2 \le B$, the third inequality follows from Jensen's inequality, and the last inequality follows from Assumption 4.1 and Lemma A.2. Thus, for any $W, W_1, W_2 \in S_B$, we have that

$$\begin{aligned}
&\mathbb{E}_{\text{init}}\Big[\big\|W^\top \phi_{W_1}(s,a) - W^\top \phi_{W_2}(s,a)\big\|_{2,\mu}^2\Big] \\
&\le 2\mathbb{E}_{\text{init}}\Big[\big\|W^\top \phi_{W_1}(s,a) - W^\top \phi_0(s,a)\big\|_{2,\mu}^2\Big] + 2\mathbb{E}_{\text{init}}\Big[\big\|W^\top \phi_{W_2}(s,a) - W^\top \phi_0(s,a)\big\|_{2,\mu}^2\Big] \\
&= \mathcal{O}(B^3 \cdot m^{-1/2}).
\end{aligned}$$

Moreover, following from the Cauchy-Schwartz inequality, we have that $\|\cdot\|_{1,\mu} \le \|\cdot\|_{2,\mu}$. Thus, by Jensen's inequality, we have that

$$\begin{aligned}
&\mathbb{E}_{\text{init}}\Big[\big\|W^\top \phi_{W_1}(s,a) - W^\top \phi_{W_2}(s,a)\big\|_{1,\mu}\Big] \\
&\le \mathbb{E}_{\text{init}}\Big[\big\|W^\top \phi_{W_1}(s,a) - W^\top \phi_{W_2}(s,a)\big\|_{2,\mu}\Big] \\
&= \mathcal{O}(B^{3/2} \cdot m^{-1/4}),
\end{aligned}$$

which completes the proof of Lemma A.2. □

A.2. Proof of Lemma A.3

In what follows, we present the proof of Lemma A.3.

Proof. Recall that we write $u_W(s,a) = W^\top \phi_W(s,a)$ and $u_0(s,a) = u_{W_0}(s,a)$. Then, we have

$$\begin{aligned} \big|W^\top \phi_{W'}(s,a)\big| &\le \big|u_0(s,a)\big| + \big|(W-W')^\top \phi_{W'}(s,a)\big| + \big|u_{W'}(s,a) - u_0(s,a)\big| \\ &\le \big|u_0(s,a)\big| + \|W-W'\|_2 \cdot \big\|\phi_{W'}(s,a)\big\|_2 + \big|u_{W'}(s,a) - u_0(s,a)\big|, \end{aligned} \tag{A.4}$$

where the last inequality follows from the Cauchy-Schwartz inequality. It suffices to upper bound the three terms on the right-hand side of (A.4). Note that we have $W, W' \in S_B$ and $\|\phi_{W'}(s,a)\|_2 \le 1$. We have that

$$\|W-W'\|_2 \cdot \big\|\phi_{W'}(s,a)\big\|_2 \le 2B. \tag{A.5}$$

It remains to upper bound the term $|u_{W'}(s,a) - u_0(s,a)|$ in (A.4). Note that $u_W(s,a)$ is almost everywhere differentiable with respect to W. Also, it holds that $\nabla_W u_W(s,a) = \phi_W(s,a)$. Thus, following from the mean-value theorem and the Cauchy-Schwartz inequality, we have that

$$\big|u_{W'}(s,a) - u_0(s,a)\big| \le \sup_{W\in S_B} \big\|\phi_W(s,a)\big\|_2 \cdot \|W' - W_0\|_2 \le B, \tag{A.6}$$

where the second inequality follows from the fact that $\|\phi_W(s,a)\|_2 \le 1$ and $W' \in S_B$. Plugging (A.5) and (A.6) into (A.4), we have that

$$\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \big|W^\top \phi_{W'}(s,a)\big| \le \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \big|u_0(s,a)\big| + 3B.$$

Following from Assumption 4.2, we have that $\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} |W^\top \phi_{W'}(s,a)|$ is sub-Gaussian. Furthermore, it holds that

$$\mathbb{E}_{\text{init}}\Big[\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \big|W^\top \phi_{W'}(s,a)\big|^2\Big] \le 2\mathbb{E}_{\text{init}}\Big[\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \big|u_0(s,a)\big|^2\Big] + 18B^2 \le 2M_0^2 + 18B^2$$

and that

$$\begin{aligned} \mathbb{P}\Big(\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \big|W^\top \phi_{W'}(s,a)\big| > t\Big) &\le \mathbb{P}\Big(\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \big|u_0(s,a)\big| + 3B > t\Big) \\ &\le \exp\big(-v\cdot (t-3B)^2\big) \le \exp\big(-v\cdot t^2/2\big) \end{aligned}$$

for $t > 2M_0 + 6B$. Thus, we complete the proof of Lemma A.3. □

B. Neural Temporal Difference

In this section, we introduce neural TD (Cai et al., 2019c), which computes ω_k in Algorithm 1. Specifically, neural TD solves the optimization problem in (3.19) via the update in (3.20), which is presented in Algorithm 2.

B.1. Proof of Proposition 4.4

Proof. We obtain the following proposition from (Cai et al., 2019c), which characterizes the convergence of Algorithm 2.

Proposition B.1 (Proposition 4.7 in (Cai et al., 2019c))**.** We set $\alpha = \min\{(1-\gamma)/8, T_{\text{TD}}^{-1/2}\}$ in Algorithm 2. Let $Q_{\bar\omega}(s,a)$ be the state-action value function associated with the output $\bar\omega$. Under Assumption 4.1, it holds for any policy π and reward function $r(s,a)$ that

$$\begin{aligned} \mathbb{E}_{\text{init}}\Big[\big\|Q_{\bar\omega}(s,a) - Q_r^\pi(s,a)\big\|_{2,\rho_\pi}^2\Big] &= 2\mathbb{E}_{\text{init}}\Big[\big\|\text{Proj}_{\mathcal{F}_{B_\omega,m}} Q_r^\pi(s,a) - Q_r^\pi(s,a)\big\|_{2,\rho_\pi}^2\Big] \\ &\quad + \mathcal{O}(B_\omega^2 \cdot T_{\text{TD}}^{-1/2} + B_\omega^3 \cdot m^{-1/2} + B_\omega^{5/2}\cdot m^{-1/4}), \end{aligned} \tag{B.1}$$

where $\mathcal{F}_{B_\omega,m}$ is defined in Definition A.1.

Algorithm 2 Neural TD

Input: Policy π, reward function r, initialization W_0, b, number of iterations T_{TD} of neural TD, and stepsize α of neural TD.
1: **Initialization.** Set $S_{B_\omega} \leftarrow \{W \in \mathbb{R}^{md} \,|\, \|W - W_0\|_2 \le B_\omega\}$ and $\omega(0) \leftarrow W_0$.
2: **for** $j = 0, \ldots, T_{\mathrm{TD}} - 1$ **do**
3: Sample (s, a, s', a'), where $(s, a) \sim \rho_\pi$, $s' \sim P(\cdot \,|\, s, a)$, and $a' \sim \pi(\cdot \,|\, s')$.
4: Compute the Bellman residual $\delta(j) = Q_{\omega(j)}(s, a) - (1-\gamma) \cdot r(s, a) - \gamma \cdot Q_{\omega(j)}(s', a')$.
5: Update ω via $\omega(j+1) \leftarrow \mathrm{Proj}_{S_{B_\omega}}\big\{\omega(j) - \alpha \cdot \delta(j) \cdot \phi_{\omega(j)}(s, a)\big\}$.
6: **end for**
Output: Output $\bar\omega = T^{-1} \sum_{t=0}^{T_{\mathrm{TD}}-1} \omega(j)$.

Recall that we denote by $\phi_0(s, a)$ the feature vector corresponding to the random initialization in (3.3). We establish the following lemma to upper bound the bias $\mathbb{E}_{\mathrm{init}}[\|\mathrm{Proj}_{\mathcal{F}_{B_\omega, m}} Q_r^\pi(s, a) - Q_r^\pi(s, a)\|_{2,\rho_\pi}^2]$ in (B.1) of Proposition B.1 when the reward function $r(s, a)$ belongs to the reward function class $\mathcal{R}_\beta$.

Lemma B.2. We consider any reward function $r_\beta(s, a) \in \mathcal{R}_\beta$ and policy π. Under Assumptions 4.1 and 4.2, it holds for $B_\omega > B_\beta + (1-\gamma)^{-1} \cdot \gamma \cdot B_P \cdot (2M_0 + 3B_\beta)$ and an absolute constant $C_v = (2 \cdot \gamma^2 \cdot B_P^2)^{-1} \cdot (1-\gamma)^2 \cdot v$ that

$$\mathbb{E}_{\mathrm{init}}\Big[\big\|\mathrm{Proj}_{\mathcal{F}_{B_\omega, m}} Q_{r_\beta}^\pi(s, a) - Q_{r_\beta}^\pi(s, a)\big\|_{2,\rho_\pi}^2\Big] = \mathcal{O}\big(B_\beta^3 \cdot m^{-1/2} + B_\omega^2 \cdot m^{-1} + B_\omega^2 \cdot \exp(-C_v \cdot B_\omega^2)\big).$$

Proof. See §B.2 for a detailed proof. □

Combining Proposition B.1 and Lemma B.2, for $B_\omega > B_\beta + (1-\gamma)^{-1} \cdot \gamma \cdot B_P \cdot (2M_0 + 3B_\beta)$, we have for any π that

$$\mathbb{E}_{\mathrm{init}}\Big[\big\|Q_{\bar\omega}(s, a) - Q_{r_\beta}^\pi(s, a)\big\|_{2,\rho_\pi}^2\Big] = \mathcal{O}\big(B_\omega^2 \cdot T_{\mathrm{TD}}^{-1/2} + B_\omega^3 \cdot m^{-1/2} + B_\omega^{5/2} \cdot m^{-1/4} + B_\omega^2 \cdot \exp(-C_v \cdot B_\omega^2)\big).$$

Finally, by setting $T_{\mathrm{TD}} = \Omega(m)$, we have that

$$\mathbb{E}_{\mathrm{init}}\Big[\big\|Q_{\bar\omega}(s, a) - Q_{r_\beta}^\pi(s, a)\big\|_{2,\rho_\pi}^2\Big] = \mathcal{O}\big(B_\omega^3 \cdot m^{-1/2} + B_\omega^{5/2} \cdot m^{-1/4} + B_\omega^2 \cdot \exp(-C_v \cdot B_\omega^2)\big),$$

which completes the proof of Proposition 4.4. □

B.2. Proof of Lemma B.2

Proof. For notational simplicity, we write $\vartheta(s, a; w) = \mathbb{1}\{|w^\top(s, a)| > 0\} \cdot (s, a)$. Under Assumption 4.2, we have that

$$P(s' \,|\, s, a) = \int \vartheta(s, a; w)^\top \varphi(s'; w) \mathrm{d}q(w), \quad \text{where } \sup_w \bigg\|\int \varphi(s; w) \mathrm{d}s\bigg\|_2 \le B_P. \tag{B.2}$$

Thus, since $r_\beta = (1-\gamma)^{-1} \cdot u_\beta(s, a)$, we have that

$$\begin{aligned} Q_{r_\beta}^\pi(s, a) &= (1-\gamma) \cdot r_\beta(s, a) + \gamma \cdot \int_{\mathcal{S}} P(s' \,|\, s, a) \cdot V_{r_\beta}^\pi(s') \mathrm{d}s' \\ &= u_\beta(s, a) + \int_{\mathcal{S}} \gamma \cdot V_{r_\beta}^\pi(s') \cdot \int \vartheta(s, a; w)^\top \varphi(s'; w) \mathrm{d}q(w) \mathrm{d}s' \\ &= u_\beta(s, a) + \int \vartheta(s, a; w)^\top \bigg(\gamma \cdot \int_{\mathcal{S}} \varphi(s'; w) V_{r_\beta}^\pi(s') \mathrm{d}s'\bigg) \mathrm{d}q(w), \end{aligned}$$

where the second equality follows from (B.2) and the last equality follows from Fubini's theorem. In the sequel, we define

$$\alpha(w) = \gamma \cdot \int_{\mathcal{S}} \varphi(s'; w) V_{r_\beta}^\pi(s') \mathrm{d}s'. \tag{B.3}$$

Note that $\alpha(w) \in \mathbb{R}^d$. Then, we have that

$$Q^{\pi}_{r_\beta}(s,a) = u_\beta(s,a) + \int \vartheta(s,a;w)^\top \alpha(w) \mathrm{d}q(w).$$

To prove Lemma B.2, we first approximate $Q^{\pi}_{r_\beta}(s,a)$ by

$$\bar{Q}(s,a) = u_\beta(s,a) + \int \vartheta(s,a;w)^\top \bar{\alpha}(w) \mathrm{d}q(w), \tag{B.4}$$

where $\bar{\alpha}(w) = \alpha(w) \cdot \mathbb{1}\{\|\alpha(w)\|_2 \le K\}$ for an absolute constant $K > 0$ specified later. Then, it holds for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ that

$$\begin{aligned} \big|\bar{Q}(s,a) - Q^{\pi}_{r_\beta}(s,a)\big| &\le \int \Big|\vartheta(s,a;w)^\top \big(\bar{\alpha}(w) - \alpha(w)\big)\Big| \mathrm{d}q(w) \\ &\le \int \big\|\vartheta(s,a;w)\big\|_2 \cdot \big\|\bar{\alpha}(w) - \alpha(w)\big\|_2 \mathrm{d}q(w) \\ &\le \sup_w \big\|\bar{\alpha}(w) - \alpha(w)\big\|_2, \end{aligned}$$

where the second inequality follows from the Cauchy-Schwartz inequality and the last inequality follows from the fact that $\|\vartheta(s,a;w)\|_2 \le 1$. Thus, we have that

$$\big\|\bar{Q}(s,a) - Q^{\pi}_{r_\beta}(s,a)\big\|_{2,\rho_\pi} \le \big\|\bar{Q}(s,a) - Q^{\pi}_{r_\beta}(s,a)\big\|_{\infty} \le \sup_w \big\|\bar{\alpha}(w) - \alpha(w)\big\|_2. \tag{B.5}$$

We now upper bound the right-hand side of (B.5). To this end, we show that $\sup_w \|\alpha(w)\|_2$ is sub-Gaussian in the sequel. By the definition of $\alpha(w)$ in (B.3), we have that

$$\begin{aligned} \sup_w \big\|\alpha(w)\big\|_2 &= \gamma \cdot \bigg\| \int_{\mathcal{S}} \varphi(s';w) V^{\pi}_{r_\beta}(s') \mathrm{d}s' \bigg\|_2 \\ &\le \gamma \cdot \sup_{s' \in \mathcal{S}} \big|V^{\pi}_{r_\beta}(s')\big| \cdot \sup_w \bigg\| \int_{\mathcal{S}} \varphi(s';w) \mathrm{d}s' \bigg\|_2 \\ &\le \gamma \cdot B_P \cdot \sup_{s' \in \mathcal{S}} \big|V^{\pi}_{r_\beta}(s')\big| \\ &\le \gamma \cdot (1-\gamma)^{-1} \cdot B_P \cdot \sup_{(s,a) \in \mathcal{S} \times \mathcal{A}} \big|u_\beta(s,a)\big|, \end{aligned} \tag{B.6}$$

where the second inequality follows from Assumption 4.2 and the third inequality follows from the fact that $V^{\pi}_{r_\beta}(s) = \mathbb{E}_{(s',a') \sim \nu_\pi(s)}[r_\beta(s',a')]$. Here we denote by $\nu_\pi(s)$ the state-action visitation measure starting from the state s and following the policy π. Following from Lemma A.3, we have that $\sup_w \|\alpha(w)\|_2$ is sub-Gaussian. By Lemma A.3 and (B.6), it holds for $t > (1-\gamma)^{-1} \cdot \gamma \cdot B_P \cdot (2M_0 + 3B_\beta)$ that

$$\begin{aligned} \mathbb{P}\Big(\sup_w \big\|\alpha(w)\big\|_2 > t\Big) &\le \mathbb{P}\Big(\gamma \cdot (1-\gamma)^{-1} \cdot B_P \cdot \sup_{(s,a) \in \mathcal{S} \times \mathcal{A}} \big|u_\beta(s,a)\big| > t\Big) \\ &\le \exp\bigg(-\frac{v \cdot (1-\gamma)^2 \cdot t^2}{2\gamma^2 \cdot B_P^2}\bigg). \end{aligned} \tag{B.7}$$

Let the absolute constant K satisfy that $K > (1-\gamma)^{-1} \cdot \gamma \cdot B_P \cdot (2M_0 + 3B_\beta)$ in (B.7). For notational simplicity, we write $C_v = (2 \cdot \gamma^2 \cdot B_P^2)^{-1} \cdot v \cdot (1-\gamma)^2$. By the fact that $\|\bar{\alpha}(w) - \alpha(w)\|_2 = \|\alpha(w)\|_2 \cdot \mathbb{1}\{\|\alpha(w)\|_2 > K\}$, we have that

$$\sup_w \big\|\bar{\alpha}(w) - \alpha(w)\big\|_2 \le \sup_w \big\|\alpha(w)\big\|_2 \cdot \mathbb{1}\Big\{\sup_w \|\alpha(w)\|_2 > K\Big\}.$$

Following from (B.5) and (B.7), we have that

$$\begin{aligned}
&\mathbb{E}_{\text{init}}\Big[\big\|\bar{Q}(s,a)-Q^{\pi}_{r_\beta}(s,a)\big\|_{2,\rho_\pi}\Big]\\
&\quad\le \mathbb{E}\Big[\sup_w\big\|\alpha(w)\big\|_2\cdot\mathbb{1}\Big\{\sup_w\|\alpha(w)\|_2>K\Big\}\Big]\\
&\quad\le \int_0^K t\cdot\mathbb{P}\Big(\sup_w\|\alpha(w)\|_2>K\Big)\mathrm{d}t+\int_K^\infty t\cdot\mathbb{P}\Big(\sup_w\|\alpha(w)\|_2>t\Big)\mathrm{d}t\\
&\quad=\mathcal{O}\big(K^2\cdot\exp(-C_v\cdot K^2)\big).
\end{aligned} \tag{B.8}$$

We now construct $\widehat{Q}(s,a)\in\mathcal{F}_{K,m}$, which approximates $\bar{Q}(s,a)$ defined in (B.4). We define

$$f(s,a)=\int\vartheta(s,a;w)^\top\bar{\alpha}(w)\mathrm{d}q(\omega).$$

Then, we have that $\bar{Q}(s,a)=u_\beta(s,a)+f(s,a)$. Note that $f(s,a)$ belongs to the following function class,

$$\widetilde{\mathcal{F}}_{K,\infty}=\Big\{\int\vartheta(s,a;w)^\top\alpha(w)\mathrm{d}q(\omega)\ \Big|\ \sup_w\big\|\alpha(w)\big\|_2\le K\Big\}.$$

We now show that $f(s,a)$ is well approximated by the following function class,

$$\widetilde{\mathcal{F}}_{K,m}=\Big\{W^\top\phi_0(s,a)=\frac{1}{\sqrt{m}}\sum_{l=1}^m[W]_l^\top\vartheta\big(s,a;[W]_l\big)\ \Big|\ \sup_l\big\|[W]_l\big\|_2\le K/\sqrt{m}\Big\},$$

where $\phi_0(s,a)$ is the feature vector corresponding to the random initialization. We obtain the following lemma from (Rahimi & Recht, 2009), which characterizes the approximation error of $\widetilde{\mathcal{F}}_{K,\infty}$ by $\widetilde{\mathcal{F}}_{K,m}$.

Lemma B.3 (Lemma 1 in (Rahimi & Recht, 2009))**.** For any $f(s,a)\in\widetilde{\mathcal{F}}_{K,\infty}$, it holds with probability at least $1-\delta$ that

$$\big\|\text{Proj}_{\widetilde{\mathcal{F}}_{K,m}}f(s,a)-f(s,a)\big\|_{2,\mu}\le K\cdot m^{-1/2}\cdot\big(1+\sqrt{2\log(1/\delta)}\big),$$

where $\mu\in\mathcal{P}(\mathcal{S}\times\mathcal{A})$.

Lemma B.3 implies that there exists $\widehat{f}(s,a)\in\widetilde{\mathcal{F}}_{K,m}$ such that

$$\begin{aligned}
\mathbb{E}_{\text{init}}\Big[\big\|\widehat{f}(s,a)-f(s,a)\big\|_{2,\rho_\pi}^2\Big]&=\int_0^\infty\mathbb{P}\Big(\big\|\widehat{f}(s,a)-f(s,a)\big\|_{2,\rho_\pi}^2>y\Big)\mathrm{d}y\\
&\le\int_0^\infty y\cdot\exp\big(-1/2\cdot(\sqrt{my}/K-1)^2\big)=\mathcal{O}(K^2/m).
\end{aligned} \tag{B.9}$$

By the fact that $\widehat{f}(s,a)\in\widetilde{\mathcal{F}}_{K,m}$ and the definition of $\mathcal{F}_{K,m}$ in Definition A.1, we have that $\widehat{f}(s,a)\in\mathcal{F}_{K,m}-u_0(s,a)$. Let

$$\widehat{Q}(s,a)=\beta^\top\phi_0(s,a)+\widehat{f}(s,a)=(\beta+W_f)^\top\phi_0(s,a).$$

We then have that $\widehat{Q}(s,a)\in\mathcal{F}_{B_\beta+K,m}$ and that

$$\begin{aligned}
\mathbb{E}_{\text{init}}\Big[\big\|\bar{Q}(s,a)-\widehat{Q}(s,a)\big\|_{2,\rho_k}^2\Big]&\le2\mathbb{E}_{\text{init}}\Big[\big\|u_\beta(s,a)-\beta^\top\phi_0(s,a)\big\|_{2,\rho_\pi}^2\Big]+2\mathbb{E}_{\text{init}}\Big[\big\|\widehat{f}(s,a)-f(s,a)\big\|_{2,\rho_\pi}^2\Big]\\
&=\mathcal{O}(B_\beta^3\cdot m^{-1/2}+K^2\cdot m^{-1}),
\end{aligned} \tag{B.10}$$

where the last inequality follows from Assumption 4.1, Lemma A.2, and (B.9).

Finally, we set $B_\omega=K+B_\beta>B_\beta+(1-\gamma)^{-1}\cdot\gamma\cdot B_P\cdot(2M_0+3B_\beta)$. Combining (B.8) and (B.10), we have that

$$\begin{aligned}
\mathbb{E}_{\text{init}}\Big[\big\|Q^\pi_{r_\beta}(s,a)-\widehat{Q}(s,a)\big\|_{2,\rho_k}^2\Big]&\le2\mathbb{E}_{\text{init}}\Big[\big\|\bar{Q}(s,a)-\widehat{Q}(s,a)\big\|_{2,\rho_k}^2\Big]+2\mathbb{E}_{\text{init}}\Big[\big\|\bar{Q}(s,a)-Q^\pi_{r_\beta}(s,a)\big\|_{2,\rho_k}^2\Big]\\
&=\mathcal{O}\big(B_\beta^3\cdot m^{-1/2}+B_\omega^2\cdot m^{-1}+B_\omega^2\cdot\exp(-C_v\cdot B_\omega^2)\big),
\end{aligned}$$

where $\widehat{Q}(s,a)\in\mathcal{F}_{B_\omega,m}$. Thus, we complete the proof of Lemma B.2. □

C. Proofs of Auxiliary Results

In what follows, we present the proofs of the lemmas in §3-5.

C.1. Proof of Proposition 3.1

Proof. By the definition of the neural network in (3.1), we have for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ that $\nabla_W u_W(s,a) = \phi_W(s,a)$ almost everywhere. We first calculate $\nabla_\theta L(\theta, \beta)$. Following from the policy gradient theorem (Sutton & Barto, 2018) and the definition of $L(\theta,\beta)$ in (2.4), we have that

$$\begin{aligned}\nabla_\theta L(\theta,\beta) &= -\nabla_\theta J(\pi_\theta; r_\beta) \\ &= -\mathbb{E}_{\nu_{\pi_\theta}}\big[Q^{\pi_\theta}_{r_\beta}(s,a) \cdot \nabla_\theta \log \pi_\theta(a \,|\, s)\big]. \end{aligned} \tag{C.1}$$

Following from the parameterization of π_θ in (3.5) and the definition of $\iota_\theta(s,a)$ in (3.8) of Proposition 3.1, we have that

$$\begin{aligned}\nabla_\theta \log \pi_\theta(a \,|\, s) &= \tau \cdot \phi_\theta(s,a) - \frac{\sum_{a' \in \mathcal{A}} \tau \cdot \exp\big(\tau \cdot \theta^\top \phi_\theta(s,a')\big) \cdot \phi_\theta(s,a')}{\sum_{a' \in \mathcal{A}} \exp\big(\tau \cdot \theta^\top \phi_\theta(s,a')\big)} \\ &= \tau \cdot \Big(\phi_\theta(s,a) - \mathbb{E}_{a' \sim \pi_\theta(\cdot \,|\, s)}\big[\phi_\theta(s,a')\big]\Big) = \tau \cdot \iota_\theta(s,a). \end{aligned} \tag{C.2}$$

Plugging (C.2) into (C.1), we have that

$$\nabla_\theta L(\theta,\beta) = -\tau \cdot \mathbb{E}_{\nu_{\pi_\theta}}\big[Q^{\pi_\theta}_{r_\beta}(s,a) \cdot \iota_\theta(s,a)\big].$$

It remains to calculate $\mathcal{I}(\theta)$ and $\nabla_\beta L(\theta,\beta)$. By (C.2) and the definition of $\mathcal{I}(\theta)$ in (3.7), it holds that

$$\begin{aligned}\mathcal{I}(\theta) &= \mathbb{E}_{\nu_{\pi_\theta}}\big[\nabla \log \pi_\theta(a \,|\, s) \nabla \log \pi_\theta(a \,|\, s)^\top\big] \\ &= \tau^2 \cdot \mathbb{E}_{\nu_{\pi_\theta}}\big[\iota_\theta(s,a)\iota_\theta(s,a)^\top\big]. \end{aligned}$$

By the definition of the objective function $L(\theta,\beta)$ in (2.4), it holds that

$$\begin{aligned}\nabla_\beta L(\theta,\beta) &= \nabla_\beta J(\pi_{\mathrm{E}}; r_\beta) - \nabla_\beta J(\pi_\theta; r_\beta) - \lambda \cdot \nabla_\beta \psi(\beta) \\ &= \mathbb{E}_{\nu_{\mathrm{E}}}\big[\nabla_\beta r_\beta(s,a)\big] - \mathbb{E}_{\nu_{\pi_\theta}}\big[\nabla_\beta r_\beta(s,a)\big] - \lambda \cdot \nabla_\beta \psi(\beta) \\ &= (1-\gamma)^{-1} \cdot \mathbb{E}_{\nu_{\mathrm{E}}}\big[\phi_\beta(s,a)\big] - (1-\gamma)^{-1} \cdot \mathbb{E}_{\nu_{\pi_\theta}}\big[\phi_\beta(s,a)\big] - \lambda \cdot \nabla_\beta \psi(\beta). \end{aligned}$$

Thus, we complete the proof of Proposition 3.1. □

C.2. Proof of Lemma 5.2

Proof. The proof of Lemma 5.2 is similar to that of Lemmas 5.4 and 5.5 in (Wang et al., 2019). By direct calculation, we have that

$$\eta \cdot \mathbb{E}_{d_{\mathrm{E}}}\Big[\big\langle Q^{\pi_k}_{r_k}(s,\cdot), \pi^s_{\mathrm{E}} - \pi^s_k \big\rangle_{\mathcal{A}}\Big] = \mathrm{KL}^{d_{\mathrm{E}}}(\pi_{\mathrm{E}} \,\|\, \pi_k) - \mathrm{KL}^{d_{\mathrm{E}}}(\pi_{\mathrm{E}} \,\|\, \pi_{k+1}) + \eta \cdot \Delta^{(\mathrm{i})}_k,$$

where $\Delta^{(\mathrm{i})}_k$ takes the form of

$$\begin{aligned}\Delta^{(\mathrm{i})}_k &= \eta^{-1} \cdot \Big\{\mathbb{E}_{d_{\mathrm{E}}}\Big[\big\langle \log(\pi^s_{k+1}/\pi^s_k) - \eta \cdot Q^{\pi_k}_{r_k}(s,\cdot), \pi^s_{\mathrm{E}} - \pi^s_k \big\rangle_{\mathcal{A}} + \big\langle \log(\pi^s_{k+1}/\pi^s_k), \pi^s_k - \pi^s_{k+1}\big\rangle_{\mathcal{A}}\Big] - \mathrm{KL}^{d_{\mathrm{E}}}(\pi^s_{k+1} \,\|\, \pi^s_k)\Big\} \\ &= \underbrace{\eta^{-1} \cdot \mathbb{E}_{d_{\mathrm{E}}}\Big[\big\langle \log(\pi^s_{k+1}/\pi^s_k) - \eta \cdot \widehat{Q}_{\omega_k}(s,\cdot), \pi^s_{\mathrm{E}} - \pi^s_k \big\rangle_{\mathcal{A}}\Big]}_{\text{(i.a)}} + \underbrace{\mathbb{E}_{d_{\mathrm{E}}}\Big[\big\langle \widehat{Q}_{\omega_k}(s,\cdot) - Q^{\pi_k}_{r_k}(s,\cdot), \pi^s_{\mathrm{E}} - \pi^s_k\big\rangle_{\mathcal{A}}\Big]}_{\text{(i.b)}} \\ &\quad + \underbrace{\eta^{-1} \cdot \mathbb{E}_{d_{\mathrm{E}}}\Big[\big\langle \log(\pi^s_{k+1}/\pi^s_k), \pi^s_k - \pi^s_{k+1}\big\rangle_{\mathcal{A}} - \mathrm{KL}(\pi^s_{k+1} \,\|\, \pi^s_k)\Big]}_{\text{(i.c)}}. \end{aligned} \tag{C.3}$$

The following lemmas upper bound $\Delta^{(\mathrm{i})}_k$ by upper bounding terms (i.a), (i.b), and (i.c) on the right-hand side of (C.3), respectively. Note that the expectation $\mathbb{E}_{\mathrm{init}, d_{\mathrm{E}}}$ is taken with respect to the random initialization in (3.3) and $s \sim d_{\mathrm{E}}$.

Lemma C.1 (Upper Bound of Term (i.a) in (C.3))**.** Under Assumptions 4.1 and 4.3, we have that

$$\mathbb{E}_{\text{init},d_{\text{E}}}\Big[\Big|\big\langle \log(\pi^s_{k+1}/\pi^s_k) - \eta\cdot \widehat{Q}_{\omega_k}(s,\cdot), \pi^s_{\text{E}} - \pi^s_k\big\rangle_{\mathcal{A}}\Big|\Big]$$
$$= \eta\cdot 2\sqrt{2}\cdot C_h\cdot B_\theta^{1/2}\cdot \sigma^{1/2}\cdot N^{-1/4} + \mathcal{O}(\tau_{k+1}\cdot B_\theta^{3/2}\cdot m^{-1/4} + \eta\cdot B_\theta^{5/4}\cdot m^{-1/8}),$$

where C_h is defined in Assumption 4.1 and σ is defined in Assumption 4.3.

Proof. See §D.1 for a detailed proof. ☐

Lemma C.2 (Upper Bound of Term (i.b) in (C.3))**.** Under Assumption 4.1, we have that

$$\mathbb{E}_{\text{init},d_{\text{E}}}\Big[\big\langle \widehat{Q}_{\omega_k}(s,\cdot) - Q^{\pi_k}_{r_k}(s,\cdot), \pi^s_{\text{E}} - \pi^s_k\big\rangle_{\mathcal{A}}\Big] \le C_h\cdot \epsilon_{Q,k},$$

where $\epsilon_{Q,k}$ takes the form of

$$\epsilon_{Q,k} = \mathbb{E}_{\text{init}}\Big[\big\|Q^{\pi_k}_{r_k}(s,a) - \widehat{Q}_{\omega_k}(s,a)\big\|_{2,\rho_k}\Big]. \tag{C.4}$$

Proof. See §D.2 for a detailed proof. ☐

Lemma C.3 (Upper Bound of Term (i.c) in (C.3))**.** Under Assumptions 4.1 and 4.2, we have that

$$\mathbb{E}_{\text{init},d_{\text{E}}}\Big[\Big|\big\langle \log(\pi^s_{k+1}/\pi^s_k), \pi^s_k - \pi^s_{k+1}\big\rangle_{\mathcal{A}}\Big| - \text{KL}(\pi^s_{k+1}\,\|\,\pi^s_k)\Big]$$
$$= \eta^2\cdot (M_0^2 + 9B_\theta^2) + \mathcal{O}(\tau_{k+1}\cdot B_\theta^{3/2}\cdot m^{-1/4}),$$

where M_0 is defined in Assumption 4.2.

Proof. See §D.3 for a detailed proof. ☐

Finally, by Lemmas C.1-C.3, under Assumptions 4.2 and 4.3, we obtain from (C.3) that

$$\mathbb{E}_{\text{init}}\big[|\Delta_k^{(\text{i})}|\big] = 2\sqrt{2}C_h\cdot B_\theta^{1/2}\cdot\sigma^{1/2}\cdot N^{-1/4} + C_h\cdot\epsilon_{Q,k} + \eta\cdot(M_0^2 + 9B_\theta^2)$$
$$+ \mathcal{O}(\eta^{-1}\cdot\tau_{k+1}\cdot B_\theta^{3/2}\cdot m^{-1/4} + B_\theta^{5/4}\cdot m^{-1/8}).$$

Here M_0 is defined in Assumption 4.2, τ_{k+1} is the inverse temperature parameter of π_{k+1} defined in (3.5), σ is defined in Assumption 4.3, and $\epsilon_{Q,k}$ is defined in (C.4) of Lemma C.2. Following from Proposition 4.4, we have that

$$C_h\cdot\epsilon_{Q,k} = \mathcal{O}\big(B_\omega^3\cdot m^{-1/2} + B_\omega^{5/2}\cdot m^{-1/4} + B_\omega^2\cdot\exp(-C_v\cdot B_\omega^2)\big).$$

Thus, we complete the proof of Lemma 5.2. ☐

C.3. Proof of Lemma 5.3

Proof. We consider a fixed $\beta'\in S_{B_\beta}$. For notational simplicity, we write $r' = r_{\beta'}(s,a)$, $r_k = r_k(s,a)$ and $\phi_\beta = \phi_\beta(s,a)$. By the parameterization of $r_\beta(s,a)$ in (3.6), we have that

$$\begin{aligned}
L(\theta_k,\beta') - L(\theta_k,\beta_k) &= \langle r' - r_k, \nu_{\text{E}} - \nu_k\rangle_{\mathcal{S}\times\mathcal{A}} + \lambda\cdot\psi(\beta_k) - \lambda\cdot\psi(\beta')\\
&= (1-\gamma)^{-1}\cdot\Big(\big\langle\phi_{\beta_k}^\top(\beta'-\beta_k), \nu_{\text{E}} - \nu_k\big\rangle_{\mathcal{S}\times\mathcal{A}} + \langle\phi_{\beta'}^\top\beta' - \phi_{\beta_k}^\top\beta', \nu_{\text{E}} - \nu_k\rangle_{\mathcal{S}\times\mathcal{A}}\Big) + \lambda\cdot\big(\psi(\beta) - \psi(\beta')\big)\\
&\le (\beta'-\beta_k)^\top\nabla_\beta L(\theta_k,\beta_k) + (1-\gamma)^{-1}\cdot\big(\|\phi_{\beta_k}^\top\beta' - \phi_{\beta'}^\top\beta'\|_{1,\nu_k} + \|\phi_{\beta_k}^\top\beta' - \phi_{\beta'}^\top\beta'\|_{1,\nu_{\text{E}}}\big),
\end{aligned} \tag{C.5}$$

where the last inequality follows from (3.10) of Proposition 3.1. Then, we have that

$$\begin{aligned}
&\mathbb{E}_{\text{init}}\big[L(\theta_k,\beta') - L(\theta_k,\beta_k)\big]\\
&\quad\le \mathbb{E}_{\text{init}}\Big[(\beta'-\beta_k)^\top\nabla_\beta L(\theta_k,\beta_k) + (1-\gamma)^{-1}\cdot\big(\|\phi_{\beta_k}^\top\beta' - \phi_{\beta'}^\top\beta'\|_{1,\nu_k} + \|\phi_{\beta_k}^\top\beta' - \phi_{\beta'}^\top\beta'\|_{1,\nu_{\text{E}}}\big)\Big]\\
&\quad\le \mathbb{E}_{\text{init}}\big[(\beta'-\beta_k)^\top\nabla_\beta L(\theta_k,\beta_k)\big] + \mathcal{O}(B_\beta^{3/2}\cdot m^{-1/4}),
\end{aligned}$$

where the last inequality follows from Assumption 4.1, Lemma A.2, and the fact that $\beta',\beta_k\in S_{B_\beta}$. Thus, we complete the proof of Lemma 5.3. ☐

C.4. Proof of Lemma 5.4

Proof. By the update of β_k in (3.14), it holds for any $\beta' \in S_{B_\beta}$ that

$$\big(\beta_k + \eta \cdot \widehat{\nabla}_\beta L(\theta_k, \beta_k) - \beta_{k+1}\big)^\top (\beta' - \beta_{k+1}) \le 0,$$

which further implies that

$$\begin{aligned}\eta \cdot (\beta' - \beta_k)^\top \nabla_\beta L(\theta_k, \beta_k) &\le \|\beta_k - \beta'\|_2^2 - \|\beta_{k+1} - \beta'\|_2^2 + \|\beta_{k+1} - \beta_k\|_2^2 \\ &\quad + \eta \cdot \Big((\beta_{k+1} - \beta_k)^\top \widehat{\nabla}_\beta L(\theta_k, \beta_k) + (\beta_k - \beta')^\top \big(\widehat{\nabla}_\beta L(\theta_k, \beta_k) - \nabla_\beta L(\theta_k, \beta_k)\big)\Big).\end{aligned} \tag{C.6}$$

Combining (C.5) and (C.6), we have that

$$\eta \cdot \big(L(\theta_k, \beta_k) - L(\theta_k, \beta')\big) \le \|\beta_k - \beta'\|_2^2 - \|\beta_{k+1} - \beta'\|_2^2 + \eta \cdot \Delta_k^{\text{(ii)}},$$

where $\Delta_k^{\text{(ii)}}$ takes the form of

$$\begin{aligned}\Delta_k^{\text{(ii)}} &= \underbrace{(\beta_{k+1} - \beta_k)^\top \widehat{\nabla}_\beta L(\theta_k, \beta_k)}_{\text{(ii.a)}} + \underbrace{(\beta_k - \beta')^\top \big(\widehat{\nabla}_\beta L(\theta_k, \beta_k) - \nabla_\beta L(\theta_k, \beta_k)\big)}_{\text{(ii.b)}} \\ &\quad + \underbrace{(1-\gamma)^{-1} \cdot \big(\|\phi_{\beta_k}^\top \beta' - \phi_{\beta'}^\top \beta'\|_{2,\nu_k} + \|\phi_{\beta_k}^\top \beta' - \phi_{\beta'}^\top \beta'\|_{2,\nu_{\mathrm{E}}}\big)}_{\text{(ii.c)}} + \underbrace{\eta^{-1} \cdot \|\beta_{k+1} - \beta_k\|_2^2}_{\text{(ii,d)}}\end{aligned} \tag{C.7}$$

We now upper bound terms (ii.a), (ii.b), and (ii.c) on the right-hand side of (C.7). Following from Assumption 4.1 and Lemma A.2, we have that

$$\mathbb{E}_{\text{init}}\big[\|\phi_{\beta_k}^\top \beta' - \phi_{\beta'}^\top \beta'\|_{2,\nu_k} + \|\phi_{\beta_k}^\top \beta' - \phi_{\beta'}^\top \beta'\|_{2,\nu_{\mathrm{E}}}\big] = \mathcal{O}(B_\beta^{3/2} \cdot m^{-1/4}), \tag{C.8}$$

which upper bounds term (ii.c) of (C.7). For term (ii.d) of (C.7), we have that

$$\eta^{-1} \cdot \|\beta_{k+1} - \beta_k\|^2 \le \eta \cdot \big\|\widehat{\nabla}_\beta L(\theta_k, \beta_k)\big\|^2 \le \eta \cdot \big(2(1-\gamma)^{-1} + \lambda \cdot L_\psi\big)^2, \tag{C.9}$$

where the first inequality follows from the property of projection and the update in (3.14), and the second inequality follows from (3.18), $\|\phi_W(s,a)\| \le 1$, and Assumption 4.3 (b). For term (ii.b) of (C.7), we have that

$$\begin{aligned}&\mathbb{E}\Big[\Big|(\beta_k - \beta')^\top \big(\widehat{\nabla}_\beta L(\theta_k, \beta_k) - \nabla_\beta L(\theta_k, \beta_k)\big)\Big|\Big] \\ &\quad \le \mathbb{E}\Big[\big\|\widehat{\nabla}_\beta L(\theta_k, \beta_k) - \nabla_\beta L(\theta_k, \beta_k)\big\|_2 \cdot \|\beta' - \beta_k\|_2\Big] \le 2B_\beta \cdot \mathbb{E}\big[\|\xi_k'\|_2\big] \le 2B_\beta \cdot (\sigma^2/N)^{1/2},\end{aligned} \tag{C.10}$$

where we write $\xi_k' = \widehat{\nabla}_\beta L(\theta_k, \beta_k) - \nabla_\beta L(\theta_k, \beta_k)$. Here the first inequality follows from the Cauchy-Schwartz inequality, the second inequality follows from the fact that $\beta_k, \beta' \in S_{B_\beta}$, and the last inequality follows from Assumption 4.3. To upper bound term (ii.a) in (C.7), we have that

$$\begin{aligned}&\mathbb{E}\Big[\big|(\beta_{k+1} - \beta_k)^\top \widehat{\nabla}_\beta L(\theta_k, \beta_k)\big|\Big] \\ &\quad \le \mathbb{E}\Big[\big\|\widehat{\nabla}_\beta L(\theta_k, \beta_k)\big\|_2 \cdot \|\beta_{k+1} - \beta_k\|_2\Big] \le \eta \cdot \mathbb{E}\Big[\big\|\widehat{\nabla}_\beta L(\theta_k, \beta_k)\big\|_2^2\Big] = 2\eta \cdot \Big(\big\|\nabla_\beta L(\theta_k, \beta_k)\big\|_2^2 + \mathbb{E}\big[\|\xi_k'\|_2^2\big]\Big),\end{aligned} \tag{C.11}$$

where the first inequality follows from the Cauchy-Schwartz inequality and the second inequality follows from the update of β in (3.14). Furthermore, we have

$$\begin{aligned}\big\|\nabla_\beta L(\theta_k, \beta_k)\big\|_2^2 &= \Big\|(1-\gamma)^{-1} \cdot \mathbb{E}_{\nu_k}\big[\phi_{\beta_k}(s,a)\big] - (1-\gamma)^{-1} \cdot \mathbb{E}_{\nu_{\mathrm{E}}}\big[\phi_{\beta_k}(s,a)\big] + \lambda \cdot \nabla_\beta \psi(\beta_k)\Big\|_2^2 \\ &\le \Big((1-\gamma)^{-1} \cdot \mathbb{E}_{\nu_k}\Big[\big\|\phi_{\beta_k}(s,a)\big\|_2\Big] + (1-\gamma)^{-1} \cdot \mathbb{E}_{\nu_{\mathrm{E}}}\Big[\big\|\phi_{\beta_k}(s,a)\big\|_2\Big] + \lambda \cdot \big\|\nabla_\beta \psi(\beta_k)\big\|_2\Big)^2 \\ &\le \big(2(1-\gamma)^{-1} + \lambda \cdot L_\psi\big)^2,\end{aligned} \tag{C.12}$$

where the first inequality follows from Jensen's inequality and the second inequality follows from the fact that $\|\phi_W(s,a)\|_2 \le 1$ and the Lipschitz continuity of $\psi(\beta)$ in Assumption 4.3. By plugging (C.12) into (C.11), we have that

$$\begin{aligned}\mathbb{E}\Big[\big|\widehat{\nabla}_\beta L(\theta_k,\beta_k)^\top(\beta_k-\beta_{k+1})\big|\Big] &\le \eta\cdot\Big(\big(2(1-\gamma)^{-1}+\lambda\cdot L_\psi\big)^2+\mathbb{E}\big[\|\xi'_k\|_2^2\big]\Big)\\ &\le \eta\cdot\Big(\big(2(1-\gamma)^{-1}+\lambda\cdot L_\psi\big)^2+\sigma^2/N\Big), \end{aligned} \tag{C.13}$$

where the last inequality follows from Assumption 4.3. Finally, by plugging (C.8), (C.9), (C.10), and (C.13) into (C.7), we have that

$$\mathbb{E}_{\text{init}}\big[|\Delta_k^{(\text{ii})}|\big] = \eta\cdot\Big(2\big(2(1-\gamma)^{-1}+\lambda\cdot L_\psi\big)^2+\sigma^2\cdot N^{-1}\Big)+2B_\beta\cdot\sigma\cdot N^{-1/2}+\mathcal{O}(B_\beta^{3/2}\cdot m^{-1/4}).$$

Thus, we complete the proof of Lemma 5.4. □

D. Proofs of Supporting Lemmas

In what follows, we present the proofs of the lemmas in §C.

D.1. Proof of Lemma C.1

Proof. It holds for any policies π,π' that

$$\big\langle D(s),\pi^s-(\pi')^s\big\rangle_{\mathcal{A}}=0, \tag{D.1}$$

where $D(s)$ only depends on the state s. Thus, we have that

$$\begin{aligned}&\big\langle \log(\pi_{k+1}^s/\pi_k^s)-\eta\cdot\widehat{Q}_{\omega_k}(s,\cdot),\pi_{\mathrm{E}}^s-\pi_k^s\big\rangle_{\mathcal{A}}\\ &\quad=\big\langle \tau_{k+1}\cdot\phi_{\theta_{k+1}}(s,\cdot)^\top\theta_{k+1}-\tau_k\cdot\phi_{\theta_k}(s,\cdot)^\top\theta_k-\eta\cdot\phi_{\omega_k}(s,\cdot)^\top\omega_k,\pi_{\mathrm{E}}^s-\pi_k^s\big\rangle_{\mathcal{A}}\\ &\quad=\big\langle \tau_{k+1}\cdot\iota_{\theta_{k+1}}(s,\cdot)^\top\theta_{k+1}-\tau_k\cdot\iota_{\theta_k}(s,\cdot)^\top\theta_k-\eta\cdot\iota_{\omega_k}(s,\cdot)^\top\omega_k,\pi_{\mathrm{E}}^s-\pi_k^s\big\rangle_{\mathcal{A}},\end{aligned}$$

where the first inequality follows from the parameterization of π_θ and $\widehat{Q}_\omega$ in (3.5) and (3.12), respectively, and the second equality follows from the definition of the temperature-adjusted score function $\iota_\theta(s,a)$ in (3.8) of Proposition 3.1. Here, with a slight abuse of the notation, we define

$$\iota_{\omega_k}(s,a)=\phi_{\omega_k}(s,a)-\mathbb{E}_{a'\sim\pi_k(\cdot\,|\,s)}\big[\phi_{\omega_k}(s,a')\big]. \tag{D.2}$$

Then, following from (D.1) and the update $\tau_{k+1}\cdot\theta_{k+1}=\tau_k\cdot\theta_k-\eta\cdot\delta_k$ in (3.13), we have that

$$\begin{aligned}&\big\langle \log(\pi_{k+1}^s/\pi_k^s)-\eta\cdot\widehat{Q}_{\omega_k}(s,\cdot),\pi_{\mathrm{E}}^s-\pi_k^s\big\rangle_{\mathcal{A}}\\ &\quad=\big\langle \tau_{k+1}\cdot\iota_{\theta_{k+1}}(s,\cdot)^\top\theta_{k+1}-\tau_k\cdot\iota_{\theta_k}(s,\cdot)^\top\theta_k-\eta\cdot\iota_{\omega_k}(s,\cdot)^\top\omega_k,\pi_{\mathrm{E}}^s-\pi_k^s\big\rangle_{\mathcal{A}}\\ &\quad=\underbrace{\tau_{k+1}\cdot\big\langle\iota_{\theta_{k+1}}(s,\cdot)^\top\theta_{k+1}-\iota_{\theta_k}(s,\cdot)^\top\theta_{k+1},\pi_{\mathrm{E}}^s-\pi_k^s\big\rangle_{\mathcal{A}}}_{\text{(i)}}-\underbrace{\eta\cdot\big\langle\iota_{\theta_k}(s,\cdot)^\top\delta_k+\iota_{\omega_k}(s,\cdot)^\top\omega_k,\pi_{\mathrm{E}}^s-\pi_k^s\big\rangle_{\mathcal{A}}}_{\text{(ii)}}.\end{aligned} \tag{D.3}$$

In what follows, we upper bound terms (i) and (ii) on the right-hand side of (D.3).

Upper bound of term (i) in (D.3)**.** Following from (3.8) of Proposition 3.1 and (D.1) we have that

$$\begin{aligned}&\Big|\big\langle\iota_{\theta_{k+1}}(s,\cdot)^\top\theta_{k+1}-\iota_{\theta_k}(s,\cdot)^\top\theta_{k+1},\pi_{\mathrm{E}}^s-\pi_k^s\big\rangle_{\mathcal{A}}\Big|\\ &\quad=\Big|\big\langle\phi_{\theta_{k+1}}(s,\cdot)^\top\theta_{k+1}-\phi_{\theta_k}(s,\cdot)^\top\theta_{k+1},\pi_{\mathrm{E}}^s-\pi_k^s\big\rangle_{\mathcal{A}}\Big|\\ &\quad\le\big\|\phi_{\theta_{k+1}}(s,\cdot)^\top\theta_{k+1}-\phi_{\theta_k}(s,\cdot)^\top\theta_{k+1}\big\|_{1,\pi_{\mathrm{E}}^s}+\big\|\phi_{\theta_{k+1}}(s,\cdot)^\top\theta_{k+1}-\phi_{\theta_k}(s,\cdot)^\top\theta_{k+1}\big\|_{1,\pi_k^s},\end{aligned} \tag{D.4}$$

where the inequality follows from the triangle inequality. Following from Assumption 4.1 and Lemma A.2, we have that

$$\mathbb{E}_{\text{init},d_{\mathrm{E}}}\Big[\big\|\phi_{\theta_{k+1}}(s,\cdot)^\top\theta_{k+1}-\phi_{\theta_k}(s,\cdot)^\top\theta_{k+1}\big\|_{1,\pi_{\mathrm{E}}^s}\Big]=\mathcal{O}(B_\theta^{3/2}\cdot m^{-1/4}). \tag{D.5}$$

Furthermore, following from Assumption 4.1, Lemma A.2, and the Cauchy-Schwartz inequality, we have that

$$\begin{aligned}
&\mathbb{E}_{\text{init},d_{\text{E}}}\Big[\big\|\phi_{\theta_{k+1}}(s,\cdot)^\top\theta_{k+1}-\phi_{\theta_k}(s,\cdot)^\top\theta_{k+1}\big\|_{1,\pi_k^s}\Big]\\
&\quad=\mathbb{E}_{\text{init},d_k}\Big[\big\|\phi_{\theta_{k+1}}(s,\cdot)^\top\theta_{k+1}-\phi_{\theta_k}(s,\cdot)^\top\theta_{k+1}\big\|_{1,\pi_k^s}\cdot\frac{\mathrm{d}d_{\text{E}}}{\mathrm{d}d_k}\Big]\\
&\quad\le\big\|\phi_{\theta_{k+1}}(s,a)^\top\theta_{k+1}-\phi_{\theta_k}(s,a)^\top\theta_{k+1}\big\|_{2,\nu_k}\cdot\Big\|\frac{\mathrm{d}d_{\text{E}}}{\mathrm{d}d_k}\Big\|_{2,d_k}\\
&\quad=\mathcal{O}(B_\theta^{3/2}\cdot m^{-1/4}).
\end{aligned}\tag{D.6}$$

Here the expectation $\mathbb{E}_{\text{init},d_k}$ is taken with respect to the random initialization in (3.3) and $s\sim d_k$. Thus, plugging (D.5) and (D.6) into (D.4), we obtain for term (i) of (D.3) that

$$\mathbb{E}_{\text{init},d_{\text{E}}}\Big[\Big|\big\langle\iota_{\theta_{k+1}}(s,\cdot)^\top\theta_{k+1}-\iota_{\theta_k}(s,\cdot)^\top\theta_{k+1},\pi_{\text{E}}^s-\pi_k^s\big\rangle_{\mathcal{A}}\Big|\Big]=\mathcal{O}(B_\theta^{3/2}\cdot m^{-1/4}).\tag{D.7}$$

Upper bound of term (ii) in (D.3). Following from the Cauchy-Schwartz inequality, we have that

$$\begin{aligned}
\mathbb{E}_{d_{\text{E}}}\Big[\Big|\big\langle\iota_{\theta_k}(s,\cdot)^\top\delta_k+\iota_{\omega_k}(s,\cdot)^\top\omega_k,\pi_{\text{E}}^s\big\rangle_{\mathcal{A}}\Big|\Big]&\le\int_{\mathcal{S}\times\mathcal{A}}\big|\iota_{\theta_k}(s,a)^\top\delta_k+\iota_{\omega_k}(s,a)^\top\omega_k\big|\mathrm{d}\nu_{\text{E}}(s,a)\\
&\le\Big\|\frac{\mathrm{d}\nu_{\text{E}}}{\mathrm{d}\nu_k}\Big\|_{2,\nu_k}\cdot\big\|\iota_{\theta_k}(s,a)^\top\delta_k+\iota_{\omega_k}(s,a)^\top\omega_k\big\|_{2,\nu_k}.
\end{aligned}\tag{D.8}$$

Similarly, we have that

$$\begin{aligned}
\mathbb{E}_{d_{\text{E}}}\Big[\Big|\big\langle\iota_{\theta_k}(s,\cdot)^\top\delta_k+\iota_{\omega_k}(s,\cdot)^\top\omega_k,\pi_k^s\big\rangle_{\mathcal{A}}\Big|\Big]&\le\int_{\mathcal{S}\times\mathcal{A}}\Big|\big\langle\iota_{\theta_k}(s,\cdot)^\top\delta_k+\iota_{\omega_k}(s,\cdot)^\top\omega_k,\pi_k^s\big\rangle_{\mathcal{A}}\Big|\mathrm{d}\pi_k^s(a)\mathrm{d}d_{\text{E}}(s)\\
&=\int_{\mathcal{S}\times\mathcal{A}}\Big|\big\langle\iota_{\theta_k}(s,\cdot)^\top\delta_k+\iota_{\omega_k}(s,\cdot)^\top\omega_k,\pi_k^s\big\rangle_{\mathcal{A}}\Big|\cdot\frac{\mathrm{d}d_{\text{E}}}{\mathrm{d}d_k}(s)\mathrm{d}\nu_k(s,a)\\
&\le\Big\|\frac{\mathrm{d}d_{\text{E}}}{\mathrm{d}d_k}\Big\|_{2,d_k}\cdot\big\|\iota_{\theta_k}(s,a)^\top\delta_k+\iota_{\omega_k}(s,a)^\top\omega_k\big\|_{2,\nu_k},
\end{aligned}\tag{D.9}$$

where the last inequality follows from the Cauchy-Schwartz inequality. Combining (D.8) and (D.9), we obtain for term (ii) of (D.3) that

$$\begin{aligned}
&\mathbb{E}_{d_{\text{E}}}\Big[\Big|\big\langle\iota_{\theta_k}(s,\cdot)^\top\delta_k+\iota_{\omega_k}(s,\cdot)^\top\omega_k,\pi_{\text{E}}^s-\pi_k^s\big\rangle_{\mathcal{A}}\Big|\Big]\\
&\quad\le\bigg(\Big\|\frac{\mathrm{d}\nu_{\text{E}}}{\mathrm{d}\nu_k}\Big\|_{2,\nu_k}+\Big\|\frac{\mathrm{d}d_{\text{E}}}{\mathrm{d}d_k}\Big\|_{2,d_k}\bigg)\cdot\big\|\iota_{\theta_k}(s,a)^\top\delta_k+\iota_{\omega_k}(s,a)^\top\omega_k\big\|_{2,\nu_k}\\
&\quad\le C_h\cdot\big\|\iota_{\theta_k}(s,a)^\top\delta_k+\iota_{\omega_k}(s,a)^\top\omega_k\big\|_{2,\nu_k},
\end{aligned}\tag{D.10}$$

where the last inequality follows from Assumption 4.1. To upper bound term (ii) of (D.3), it suffices to upper bound the right-hand side of (D.10). For notational simplicity, we write $\iota_{\theta_k}=\iota_{\theta_k}(s,a)$, $\iota_{\omega_k}=\iota_{\omega_k}(s,a)$, and $\phi_{\omega_k}=\phi_{\omega_k}(s,a)$. By the triangle inequality, we have that

$$\begin{aligned}
\|\delta_k^\top\iota_{\theta_k}+\omega_k^\top\iota_{\omega_k}\|_{2,\nu_k}&=\Big(\mathbb{E}_{\nu_k}\big[(\delta_k^\top\iota_{\theta_k}+\omega_k^\top\iota_{\omega_k})\cdot(\delta_k^\top\iota_{\theta_k}+\omega_k^\top\iota_{\omega_k})\big]\Big)^{1/2}\\
&\le\underbrace{\Big|(\delta_k-\omega_k)^\top\mathbb{E}_{\nu_k}\big[\iota_{\theta_k}(\delta_k^\top\iota_{\theta_k}+\omega_k^\top\iota_{\omega_k})\big]\Big|^{1/2}}_{\text{(ii.a)}}+\underbrace{\Big|\mathbb{E}_{\nu_k}\big[\omega_k^\top(\iota_{\theta_k}-\iota_{\omega_k})\cdot(\delta_k^\top\iota_{\theta_k}+\omega_k^\top\iota_{\omega_k})\big]\Big|^{1/2}}_{\text{(ii.b)}}.
\end{aligned}\tag{D.11}$$

We now upper bound the two terms (ii.a) and (ii.b) on the right-hand side of (D.11). For term (ii.a) of (D.11), following from (3.9) of Proposition 3.1, we have that

$$\mathcal{I}(\theta_k)=\tau_k^2\cdot\mathbb{E}_{\nu_k}[\iota_{\theta_k}\iota_{\theta_k}^\top].\tag{D.12}$$

Recall that the expectation $\mathbb{E}_k$ is taken with respect to the k-th batch. Following from the definition of $\widehat{\nabla}_\theta L(\theta_k, \beta_k)$ in (3.17), we have that

$$\begin{aligned}\mathbb{E}_k\big[\widehat{\nabla}_\theta L(\theta_k, \beta_k)\big] &= -\tau_k \cdot \mathbb{E}_{\nu_k}[\omega_k^\top \phi_{\omega_k} \cdot \iota_{\theta_k}] \\ &= -\tau_k \cdot \mathbb{E}_{\nu_k}[\omega_k^\top \iota_{\omega_k} \cdot \iota_{\theta_k}] - \tau_k \cdot w_k^\top \mathbb{E}_{a'\sim\pi_k^s}\big[\phi_{\omega_k}(s, a')\big] \cdot \mathbb{E}_{\nu_k}[\iota_{\theta_k}] \\ &= -\tau_k \cdot \mathbb{E}_{\nu_k}[\omega_k^\top \iota_{\omega_k} \cdot \iota_{\theta_k}], \end{aligned} \tag{D.13}$$

where the first equality follows from the fact that $\widehat{Q}_{\omega_k}(s,a) = \omega_k^\top \phi_{\omega_k}(s,a)$, while the second and third equalities follow from the definition of $\iota_{\omega_k}(s,a)$ in (D.2). Following from (D.12) and (D.13), we have that

$$\begin{aligned}\Big|(\delta_k - \omega_k)^\top \mathbb{E}_{\nu_k}\big[\iota_{\theta_k}(\delta_k^\top \iota_{\theta_k} + \omega_k^\top \iota_{\omega_k})\big]\Big| &= \tau_k^{-2} \cdot \Big|(\delta_k - \omega_k)^\top \Big(\mathcal{I}(\theta_k)\delta_k - \tau_k \cdot \mathbb{E}_k\big[\widehat{\nabla}_\theta L(\theta, \beta)\big]\Big)\Big| \\ &\le 2B_\theta \cdot \tau_k^{-2} \cdot \Big\|\mathcal{I}(\theta_k)\delta_k - \tau_k \cdot \mathbb{E}_k\big[\widehat{\nabla}_\theta L(\theta, \beta)\big]\Big\|_2. \end{aligned} \tag{D.14}$$

Here the last inequality follows from the Cauchy-Schwartz inequality and the fact that $\|\omega_k - \delta_k\|_2 \le 2B_\theta$ as $\omega_k, \delta_k \in S_{B_\theta}$. For notational simplicity, we define the following error terms,

$$\xi_k^{(1)} = \widehat{\mathcal{I}}(\theta_k)\delta_k - \mathcal{I}(\theta_k)\delta_k, \tag{D.15}$$

$$\xi_k^{(2)} = \widehat{\nabla}_\theta L(\theta_k, \beta_k) - \mathbb{E}_k\big[\widehat{\nabla}_\theta L(\theta_k, \beta_k)\big]. \tag{D.16}$$

Then, we have for term (ii.a) in (D.11) that

$$\begin{aligned}&\mathbb{E}_{\text{init}}\bigg[\Big|(\delta_k - \omega_k)^\top \mathbb{E}_{\nu_k}\big[\iota_{\theta_k}(\delta_k^\top \iota_{\theta_k} + \omega_k^\top \iota_{\omega_k})\big]\Big|^{1/2}\bigg] \qquad\qquad (\text{D.17}) \\ &\le (2B_\theta)^{1/2} \cdot \tau_k^{-1} \cdot \mathbb{E}_{\text{init}}\bigg[\Big\|\mathcal{I}(\theta_k)\delta_k - \tau_k \cdot \mathbb{E}_k\big[\widehat{\nabla}_\theta L(\theta, \beta)\big]\Big\|_2^{1/2}\bigg] \\ &\le (2B_\theta)^{1/2} \cdot \tau_k^{-1} \cdot \mathbb{E}_{\text{init}}\bigg[\Big(\big\|\widehat{\mathcal{I}}(\theta_k)\delta_k - \tau_k \cdot \widehat{\nabla}_\theta L(\theta, \beta)\big\|_2 + \|\xi_k^{(1)}\|_2 + \tau_k \cdot \|\xi_k^{(2)}\|_2\Big)^{1/2}\bigg] \\ &\le (2B_\theta)^{1/2} \cdot \tau_k^{-1} \cdot \bigg(\mathbb{E}_{\text{init}}\Big[\big\|\widehat{\mathcal{I}}(\theta_k)\delta_k - \tau_k \cdot \widehat{\nabla}_\theta L(\theta, \beta)\big\|_2\Big] + \mathbb{E}_{\text{init}}\big[\|\xi_k^{(1)}\|_2 + \tau_k \cdot \|\xi_k^{(2)}\|_2\big]\bigg)^{1/2}, \end{aligned}$$

where the first inequality follows from (D.14), the second inequality follows from the triangle inequality, and the last inequality follows from Jensen's inequality. Similarly to (D.15), we define the following error term,

$$\xi_k^{(3)} = \widehat{\mathcal{I}}(\theta_k)\omega_k - \mathcal{I}(\theta_k)\omega_k. \tag{D.18}$$

We now upper bound the right-hand side of (D.17). Recall the definition of δ_k in (3.15). We have that

$$\begin{aligned}\big\|\widehat{\mathcal{I}}(\theta_k)\delta_k - \tau_k \cdot \widehat{\nabla}_\theta L(\theta_k, \beta_k)\big\|_2 &\le \big\|\widehat{\mathcal{I}}(\theta_k)\omega_k - \tau_k \cdot \widehat{\nabla}_\theta L(\theta_k, \beta_k)\big\|_2 \qquad (\text{D.19}) \\ &\le \Big\|\mathcal{I}(\theta_k)\omega_k - \tau_k \cdot \mathbb{E}_k\big[\widehat{\nabla}_\theta L(\theta_k, \beta_k)\big]\Big\|_2 + \|\xi_k^{(1)}\|_2 + \tau_k \cdot \|\xi_k^{(2)}\|_2. \end{aligned}$$

Following from (D.12), (D.13), and Jensen's inequality, we have that

$$\begin{aligned}\Big\|\mathcal{I}(\theta_k)\omega_k - \tau_k \cdot \mathbb{E}_k\big[\widehat{\nabla}_\theta L(\theta_k, \beta_k)\big]\Big\|_2 &= \tau_k^2 \cdot \Big\|\mathbb{E}_{\nu_k}\big[\iota_{\theta_k} \cdot \omega_k^\top(\iota_{\theta_k} - \iota_{\omega_k})\big]\Big\|_2 \\ &\le \tau_k^2 \cdot \mathbb{E}_{\nu_k}\Big[\|\iota_{\theta_k}\|_2 \cdot \big|\omega_k^\top(\iota_{\theta_k} - \iota_{\omega_k})\big|\Big] \\ &\le 2\tau_k^2 \cdot \big\|\omega_k^\top(\iota_{\theta_k} - \iota_{\omega_k})\big\|_{1,\nu_k}, \end{aligned}$$

where the last inequality follows from the fact that $\|\iota_\theta\|_2 \le 2$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$. Following from Assumption 4.1 and Lemma A.2, we have that

$$\begin{aligned}\mathbb{E}_{\text{init}}\bigg[\Big\|\mathcal{I}(\theta_k)\omega_k - \tau_k \cdot \mathbb{E}_k\big[\widehat{\nabla}_\theta L(\theta_k, \beta_k)\big]\Big\|_2\bigg] &\le \mathbb{E}_{\text{init}}\Big[2\tau_k^2 \cdot \big\|\omega_k^\top(\iota_{\theta_k} - \iota_{\omega_k})\big\|_{1,\nu_k}\Big] \\ &= \mathcal{O}(\tau_k^2 \cdot B_\theta^{3/2} \cdot m^{-1/4}). \end{aligned} \tag{D.20}$$

Plugging (D.19) and (D.20) into (D.17), we have that

$$\begin{aligned}
&\mathbb{E}_{\text{init}}\Big[\Big|(\delta_k-\omega_k)^\top\mathbb{E}_{\nu_k}\big[\iota_{\theta_k}(\delta_k^\top\iota_{\theta_k}+\omega_k^\top\iota_{\omega_k})\big]\Big|^{1/2}\Big]\\
&\quad=(2B_\theta)^{1/2}\cdot\tau_k^{-1}\cdot\Big(\mathcal{O}(2\tau_k^2\cdot B_\theta^{3/2}\cdot m^{-1/4})+\mathbb{E}_{\text{init}}\big[\|\xi_k^{(1)}\|_2+2\tau_k\cdot\|\xi_k^{(2)}\|_2+\|\xi_k^{(3)}\|_2\big]\Big)^{1/2}\\
&\quad=\mathcal{O}(\tau_k\cdot B_\theta^{5/4}\cdot m^{-1/4})+(2B_\theta)^{1/2}\cdot\tau_k^{-1}\cdot\Big(\mathbb{E}_{\text{init}}\big[\|\xi_k^{(1)}\|_2+2\tau_k\cdot\|\xi_k^{(2)}\|_2+\|\xi_k^{(3)}\|_2\big]\Big)^{1/2}\\
&\quad\le\mathcal{O}(\tau_k\cdot B_\theta^{5/4}\cdot m^{-1/4})+2\sqrt{2}B_\theta^{1/2}\cdot(\sigma^2/N)^{1/4},
\end{aligned}\tag{D.21}$$

where the last inequality follows from Assumption 4.3. We now upper bound term (ii.a) of (D.11). We have that

$$\begin{aligned}
&\mathbb{E}_{\text{init}}\Big[\Big|\mathbb{E}_{\nu_k}\big[\omega_k^\top(\iota_{\theta_k}-\iota_{\omega_k})\cdot(\delta_k^\top\iota_{\theta_k}+\omega_k^\top\iota_{\omega_k})\big]\Big|^{1/2}\Big]\\
&\quad\le\mathbb{E}_{\text{init},\nu_k}\Big[\big|\omega_k^\top(\iota_{\theta_k}-\iota_{\omega_k})\cdot(\delta_k^\top\iota_{\theta_k}+\omega_k^\top\iota_{\omega_k})\big|\Big]^{1/2}\\
&\quad\le\mathbb{E}_{\text{init}}\Big[\big\|\omega_k^\top(\iota_{\theta_k}-\iota_{\omega_k})\big\|_{2,\nu_k}\Big]^{1/2}\cdot\mathbb{E}_{\text{init}}\big[\|\delta_k^\top\iota_{\theta_k}+\omega_k^\top\iota_{\omega_k}\|_{2,\nu_k}\big]^{1/2},
\end{aligned}\tag{D.22}$$

where the expectation $\mathbb{E}_{\text{init},\nu_k}$ is taken with respect to the random initialization in (3.3) and $(s,a)\sim\nu_k$, the first inequality follows from Jensen's inequality, and the second inequality follows from the Cauchy-Schwartz inequality. Following from Assumption 4.1 and Lemma A.2, we have that

$$\mathbb{E}_{\text{init}}\Big[\big\|\omega_k^\top(\iota_{\theta_k}-\iota_{\omega_k})\big\|_{2,\nu_k}\Big]=\mathcal{O}(B_\theta^{3/2}\cdot m^{-1/4}).\tag{D.23}$$

To upper bound the right-hand side of (D.22), it remains to upper bound the term $\mathbb{E}_{\text{init}}[\|\delta_k^\top\iota_{\theta_k}+\omega_k^\top\iota_{\omega_k}\|_{2,\nu_k}]$. We have that

$$\mathbb{E}_{\text{init}}\big[\|\delta_k^\top\iota_{\theta_k}+\omega_k^\top\iota_{\omega_k}\|_{2,\nu_k}\big]\le\mathbb{E}_{\text{init}}\big[\|\delta_k\|_2\cdot\|\iota_{\theta_k}\|_2\big]+\mathbb{E}_{\text{init}}\big[\|\omega_k\|_2\cdot\|\iota_{\omega_k}\|_2\big]=\mathcal{O}(B_\theta),\tag{D.24}$$

where the inequality follows from the Cauchy-Schwartz inequality and the equality follows from the facts that $\|\iota_{\theta_k}\|_2\le 2$, $\|\iota_{\omega_k}\|_2\le 2$, and $\delta_k,\omega_k\in S_{B_\theta}$. Plugging (D.23) and (D.24) into (D.22), we have that

$$\mathbb{E}_{\text{init}}\Big[\Big|\mathbb{E}_{\nu_k}\big[\omega_k^\top(\iota_{\theta_k}-\iota_{\omega_k})\cdot(\delta_k^\top\iota_{\theta_k}+\omega_k^\top\iota_{\omega_k})\big]\Big|^{1/2}\Big]=\mathcal{O}(B_\theta^{5/4}\cdot m^{-1/8}),\tag{D.25}$$

which upper bounds term (ii.b) of (D.11). Plugging (D.21) and (D.25) into (D.11), following from (D.10), we have that

$$\begin{aligned}
&\mathbb{E}_{\text{init},d_{\text{E}}}\Big[\Big|\big\langle\iota_{\theta_k}(s,\cdot)^\top\delta_k-\iota_{\omega_k}(s,\cdot)^\top\omega_k,\pi_{\text{E}}^s-\pi_k^s\big\rangle_{\mathcal{A}}\Big|\Big]\\
&\quad=\eta\cdot C_h\cdot\big(\mathcal{O}(B_\theta^{5/4}\cdot m^{-1/8})+2\sqrt{2}B_\theta^{1/2}\cdot(\sigma^2/N)^{1/4}\big),
\end{aligned}\tag{D.26}$$

which upper bounds term (ii) of (D.3).

Finally, plugging (D.7) and (D.26) into (D.3), we have that

$$\begin{aligned}
&\mathbb{E}_{\text{init},d_{\text{E}}}\Big[\Big|\big\langle\log(\pi_{k+1}^s/\pi_k^s)-\eta\cdot\widehat{Q}_{\omega_k}(s,\cdot),\pi_{\text{E}}^s-\pi_k^s\big\rangle_{\mathcal{A}}\Big|\Big]\\
&\quad=\eta\cdot C_h\cdot 2\sqrt{2}B_\theta^{1/2}\cdot(\sigma^2/N)^{1/4}+\mathcal{O}(\tau_{k+1}\cdot B_\theta^{3/2}\cdot m^{-1/4}+\eta\cdot B_\theta^{5/4}\cdot m^{-1/8}),
\end{aligned}$$

where $\xi_k^{(1)}$, $\xi_k^{(2)}$, and $\xi_k^{(3)}$ are defined in (D.15), (D.16), and (D.18), respectively, and C_h is defined in Assumption 4.1. Thus, we complete the proof of Lemma C.1. □

D.2. Proof of Lemma C.2

Proof. For notational simplicity, for any $(s,a) \in \mathcal{S}\times\mathcal{A}$, we denote by $\Delta_{Q,k}(s,a) = \widehat{Q}_{\omega_k}(s,a) - Q^{\pi_k}_{r_k}(s,a)$ the error of estimating $Q^{\pi_k}_{r_k}(s,a)$ by $\widehat{Q}_{\omega_k}(s,a)$. Then, we have that

$$
\begin{aligned}
&\mathbb{E}_{d_{\rm E}}\Bigl[\bigl|\langle \Delta_{Q,k}(s,\cdot), \pi^s_{\rm E} - \pi^s_k\rangle_{\mathcal{A}}\bigr|\Bigr] \\
&\le \int_{\mathcal{S}\times\mathcal{A}} \bigl|\Delta_{Q,k}(s,a)\bigr| {\rm d}\pi^s_{\rm E}(a){\rm d}d_{\rm E}(s) + \int_{\mathcal{S}\times\mathcal{A}} \bigl|\Delta_{Q,k}(s,a)\bigr| {\rm d}\pi^s_k(a){\rm d}d_{\rm E}(s) \\
&= \int_{\mathcal{S}\times\mathcal{A}} \bigl|\Delta_{Q,k}(s,a)\bigr| \cdot \frac{{\rm d}\nu_{\rm E}}{{\rm d}\rho_k}(s,a){\rm d}\rho_k(s,a) + \int_{\mathcal{S}\times\mathcal{A}} \bigl|\Delta_{Q,k}(s,a)\bigr| \cdot \frac{{\rm d}d_{\rm E}}{{\rm d}\varrho_k}(s){\rm d}\rho_k(s,a) \\
&\le C_h \cdot \|\Delta_{Q,k}\|_{2,\rho_k},
\end{aligned}
$$

where the last inequality follows from the Cauchy-Schwartz inequality and Assumption 4.1. Thus, we complete the proof of Lemma C.2. ☐

D.3. Proof of Lemma C.3

Proof. Following from (D.1) and the parameterization of π_θ in (3.5), we have that

$$
\begin{aligned}
&\bigl\langle \log(\pi^s_{k+1}/\pi^s_k), \pi^s_k - \pi^s_{k+1}\bigr\rangle_{\mathcal{A}} \\
&\quad = \bigl\langle \tau_{k+1}\cdot\theta_{k+1}^\top \phi_{\theta_{k+1}}(s,\cdot) - \tau_k\cdot\theta_k^\top\phi_{\theta_k}(s,\cdot), \pi^s_k - \pi^s_{k+1}\bigr\rangle_{\mathcal{A}} \\
&\quad = \bigl\langle (\tau_{k+1}\cdot\theta_{k+1} - \tau_k\cdot\theta_k)^\top\phi_{\theta_k}(s,\cdot), \pi^s_k - \pi^s_{k+1}\bigr\rangle_{\mathcal{A}} + \tau_{k+1}\cdot\Bigl\langle \theta_{k+1}^\top\bigl(\phi_{\theta_{k+1}}(s,\cdot) - \phi_{\theta_k}(s,\cdot)\bigr), \pi^s_k - \pi^s_{k+1}\Bigr\rangle_{\mathcal{A}}.
\end{aligned} \tag{D.27}
$$

We now upper bound the two terms on the right-hand side of (D.27). For the first term on the right-hand side of (D.27), recall that we define δ_k in (3.15). Thus, we have that

$$
\bigl|(\tau_{k+1}\cdot\theta_{k+1} - \tau_k\cdot\theta_k)^\top\phi_{\theta_k}(s,a)\bigr| = \eta\cdot\bigl|\delta_k^\top\phi_{\theta_k}(s,a)\bigr|. \tag{D.28}
$$

Following from (D.28) and Hölder's inequality, we have for any $s\in\mathcal{S}$ that

$$
\begin{aligned}
&\Bigl|\bigl\langle (\tau_{k+1}\cdot\theta_{k+1} - \tau_k\cdot\theta_k)^\top\phi_{\theta_k}(s,\cdot), \pi^s_k - \pi^s_{k+1}\bigr\rangle_{\mathcal{A}}\Bigr| \\
&\quad \le \bigl\|\delta_k^\top\phi_{\theta_k}(s,\cdot)\bigr\|_\infty \cdot \|\pi^s_k - \pi^s_{k+1}\|_1.
\end{aligned}
$$

Then, following from Pinsker's inequality, we have that

$$
\begin{aligned}
&\Bigl|\bigl\langle (\tau_{k+1}\cdot\theta_{k+1} - \tau_k\cdot\theta_k)^\top\phi_{\theta_k}(s,\cdot), \pi^s_k - \pi^s_{k+1}\bigr\rangle_{\mathcal{A}}\Bigr| - {\rm KL}(\pi^s_{k+1}\,\|\,\pi^s_k) \\
&\quad \le \eta\cdot\bigl\|\delta_k^\top\phi_{\theta_k}(s,\cdot)\bigr\|_\infty \cdot \|\pi^s_k - \pi^s_{k+1}\|_1 - 1/2\cdot\|\pi^s_k - \pi^s_{k+1}\|_1^2 \\
&\quad \le 1/2\cdot\eta^2\cdot\bigl\|\delta_k^\top\phi_{\theta_k}(s,\cdot)\bigr\|_\infty^2.
\end{aligned} \tag{D.29}
$$

By the update of θ_k in (3.13) and the definition of δ_k in (3.15), we have that $\theta_k, \delta_k \in S_{B_\theta}$. Thus, by Lemma A.3, we have that

$$
\mathbb{E}_{\rm init}\Bigl[\bigl\|\delta_k^\top\phi_{\theta_k}(s,\cdot)\bigr\|_\infty^2\Bigr] \le 2M_0 + 18B_\theta^2. \tag{D.30}
$$

Plugging (D.30) into (D.29), we have that

$$
\Bigl|\bigl\langle (\tau_{k+1}\cdot\theta_{k+1} - \tau_k\cdot\theta_k)^\top\phi_{\theta_k}(s,\cdot), \pi^s_k - \pi^s_{k+1}\bigr\rangle_{\mathcal{A}}\Bigr| - {\rm KL}(\pi^s_{k+1}\,\|\,\pi^s_k) \le \eta^2\cdot(M_0^2 + 9B_\theta^2). \tag{D.31}
$$

For the second term on the right-hand side of (D.27), following from Assumption 4.1 and Lemma A.2, we have

$$
\begin{aligned}
&\mathbb{E}_{{\rm init}, d_{\rm E}}\biggl[\Bigl|\Bigl\langle \theta_{k+1}^\top\bigl(\phi_{\theta_{k+1}}(s,\cdot) - \phi_{\theta_k}(s,\cdot)\bigr), \pi^s_k - \pi^s_{k+1}\Bigr\rangle_{\mathcal{A}}\Bigr|\biggr] \\
&\quad \le \mathbb{E}_{{\rm init}, d_{\rm E}}\Bigl[\bigl\|\theta_{k+1}^\top\bigl(\phi_{\theta_{k+1}}(s,\cdot) - \phi_{\theta_k}(s,\cdot)\bigr)\bigr\|_{1,\pi^s_k}\Bigr] + \mathbb{E}_{{\rm init}, d_{\rm E}}\Bigl[\bigl\|\theta_{k+1}^\top\bigl(\phi_{\theta_{k+1}}(s,\cdot) - \phi_{\theta_k}(s,\cdot)\bigr)\bigr\|_{1,\pi^s_{k+1}}\Bigr] \\
&\quad = \mathcal{O}(B_\theta^{3/2}\cdot m^{-1/4}).
\end{aligned} \tag{D.32}
$$

Finally, plugging (D.31) and (D.32) into (D.27), we have that

$$\mathbb{E}_{\text{init},d_{\text{E}}}\Big[\Big|\big\langle \log(\pi_{k+1}^s/\pi_k^s), \pi_k^s - \pi_{k+1}^s \big\rangle_{\mathcal{A}}\Big| - \text{KL}(\pi_{k+1}^s \,\|\, \pi_k^s)\Big]$$
$$= \eta^2 \cdot (M_0^2 + 9B_\theta^2) + \mathcal{O}(\tau_{k+1} \cdot B_\theta^{3/2} \cdot m^{-1/4}),$$

which completes the proof of Lemma C.3. □