Journal of Machine Learning Research 23 (2022) 1-91 Submitted 11/21; Revised 10/22; Published 10/22

Global Optimality and Finite Sample Analysis of Softmax Off-Policy Actor Critic under State Distribution Mismatch

Shangtong Zhang SHANGTONG@VIRGINIA.EDU
University of Virginia
85 Engineer's Way, Charlottesville, VA, 22903, United States
Remi Tachet des Combes† REMI.TACHET@MICROSOFT.COM
Microsoft Research Montreal
6795 Rue Marconi, Suite 400, Montreal, Quebec, H2S 3J9, Canada
Romain Laroche† ROMAIN.LAROCHE@MICROSOFT.COM
Microsoft Research Montreal
6795 Rue Marconi, Suite 400, Montreal, Quebec, H2S 3J9, Canada

Editor: John Shawe-Taylor

Abstract

In this paper, we establish the global optimality and convergence rate of an off-policy actor critic algorithm in the tabular setting without using density ratio to correct the discrepancy between the state distribution of the behavior policy and that of the target policy. Our work goes beyond existing works on the optimality of policy gradient methods in that existing works use the exact policy gradient for updating the policy parameters while we use an approximate and stochastic update step. Our update step is not a gradient update because we do not use a density ratio to correct the state distribution, which aligns well with what practitioners do. Our update is approximate because we use a learned critic instead of the true value function. Our update is stochastic because at each step the update is done for only the current state action pair. Moreover, we remove several restrictive assumptions from existing works in our analysis. Central to our work is the finite sample analysis of a generic stochastic approximation algorithm with time-inhomogeneous update operators on time-inhomogeneous Markov chains, based on its uniform contraction properties.

Keywords: off-policy learning, actor-critic, policy gradient, density ratio, distribution mismatch

1. Introduction

Policy gradient methods (Williams, 1992), as well as their actor-critic extensions (Sutton et al., 1999; Konda and Tsitsiklis, 1999), are an important class of Reinforcement Learning (RL, Sutton and Barto 2018) algorithms and have enjoyed great empirical success (Silver et al., 2016; Mnih et al., 2016; Vinyals et al., 2019), which motivates the importance of the theoretical analysis of policy gradient methods. Policy gradient and actor-critic methods are essentially stochastic gradient ascent algorithms and, therefore, expected to converge to stationary points under mild conditions in on-policy settings, where an agent selects actions according to its current policy (Sutton et al., 1999; Konda and Tsitsiklis, 1999; Kumar et al.,

† indicates equal advising.

©2022 Shangtong Zhang, Remi Tachet des Combes, Romain Laroche.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v23/21-1306.html.

2019; Zhang et al., 2020a; Wu et al., 2020; Xu et al., 2020; Qiu et al., 2021). Off-policy learning is a paradigm where an agent learns a policy of interest, referred to as the target policy, but selects actions according to a different policy, referred to as the behavior policy. Compared with on-policy learning, off-policy learning exhibits improved sample efficiency (Lin, 1992; Sutton et al., 2011) and safety (Dulac-Arnold et al., 2019). In off-policy settings, the density ratio, i.e. the ratio between the state distribution of the target policy and that of the behavior policy (Hallak and Mannor, 2017; Gelada and Bellemare, 2019; Liu et al., 2018; Nachum et al., 2019; Zhang et al., 2020b), can be used to correct the state distribution mismatch between the behavior policy and the target policy. Consequently, convergence to stationary points of actor-critic methods in off-policy settings with density ratio has also been established (Liu et al., 2019; Zhang et al., 2020c; Huang and Jiang, 2021; Xu et al., 2021).

The seminal work of Agarwal et al. (2020) goes beyond stationary points by establishing the global optimality of policy gradient methods in the tabular setting. Mei et al. (2020) further provide some missing convergence rates. Both, however, use the exact policy gradient instead of an approximate and stochastic gradient, i.e., they assume the value function and the state distribution of the current policy are known and query the value function for all states at every iteration. Despite the aforementioned limitation, Agarwal et al. (2020) still lay the first step towards understanding the global optimality of policy gradient methods. Their success has also been extended to the off-policy setting by Laroche and Tachet (2021), who, importantly, consider off-policy actor critic methods without correcting the state distribution mismatch. Consequently, the update step they perform is not a gradient. This aligns better with RL practices: to achieve good performance, practitioners usually do not correct the state distribution mismatch with density ratios for large scale RL experiments (Wang et al., 2017; Espeholt et al., 2018; Vinyals et al., 2019; Schmitt et al., 2020; Zahavy et al., 2020). Still, Laroche and Tachet (2021) use exact and expected update steps, instead of approximate and stochastic update steps.

In this work, we go beyond Agarwal et al. (2020); Laroche and Tachet (2021) by establishing the global optimality and convergence rate of an off-policy actor critic algorithm with approximate and stochastic update steps. Similarly, we study the off-policy actor critic algorithm in the tabular setting with softmax parameterization of the policy. Like Laroche and Tachet (2021), we do not use the density ratio to correct the state distribution mismatch. We, however, use a learned value function (i.e., approximate updates) and perform stochastic updates for both the actor and the critic. Further, we use the KL divergence between a uniformly random policy and the current policy as a regularization with a decaying weight for the actor update. Our off-policy actor critic algorithm, therefore, runs in three timescales: the critic is updated in the fastest timescale; the actor runs in the middle timescale; the weight of regularization decays in the slowest timescale. Besides the advances of using approximate and stochastic update steps,we also remove two restrictive assumptions. The first assumption requires that the initial distribution of the Markov Decision Process (MDP) covers the whole state space, which is crucial to get the desired optimality in Agarwal et al. (2020). The second assumption requires that the optimal policy of the MDP is unique, which is crucial to get the nonasymptotic convergence rate of Laroche and Tachet (2021) for the softmax parameterization. Thanks to the off-policy learning and the decaying KL divergence regularization, we are able to remove those two assumptions in our analysis.

One important ingredient of our convergence results is the finite sample analysis of a generic stochastic approximation algorithm with time-inhomogeneous update operators on time-inhomogeneous Markov chains (Section 3). Similar to Chen et al. (2021), we rely on the use of the generalized Moreau envelope to form a Lyapunov function. Our results, however, extend those of Chen et al. (2021) from time-homogeneous to time-inhomogeneous Markov chains and from time-homogeneous to time-inhomogeneous update operators. Those extensions make our results immediately applicable to the off-policy actor-critic settings (Section 4) and are made possible by establishing a form of uniform contraction of the time-inhomogeneous update operators. Moreover, we demonstrate that our analysis can also be used for analyzing the soft actor-critic (a.k.a. maximum entropy RL, Nachum et al. 2017; Haarnoja et al. 2018) under state distribution mismatch (Section 5).

2. Background

In this paper, calligraphic letters denote sets and we use vectors and functions interchangeably when it does not confuse, e.g., let $f : \mathcal{S} \to \mathbb{R}$ be a function; we also use f to denote the vector in $\mathbb{R}^{|\mathcal{S}|}$ whose s-th element is $f(s)$. All vectors are column. We use $\|\cdot\|$ to denote the standard ℓ_2 norm and $\langle x, y\rangle \doteq x^\top y$ for the inner product in Euclidean spaces. $\|\cdot\|_p$ is the standard ℓ_p norm. For any norm $\|\cdot\|_m$, $\|\cdot\|_m^*$ denotes its dual norm.

We consider an infinite horizon MDP with a finite state space $\mathcal{S}$, a finite action space $\mathcal{A}$, a reward function $r : \mathcal{S} \times \mathcal{A} \to [-r_{max}, r_{max}]$ for some positive scalar r_{max}, a transition kernel $p : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$, a discount factor $\gamma \in [0, 1)$, and an initial distribution $p_0 : \mathcal{S} \to [0, 1]$. At time step 0, an initial state S_0 is sampled according to p_0. At time step t, an agent in state S_t takes an action $A_t \sim \pi(\cdot|S_t)$ according to a policy $\pi : \mathcal{A} \times \mathcal{S} \to [0, 1]$, gets a reward $R_{t+1} \doteq r(S_t, A_t)$, and proceeds to a successor state $S_{t+1} \sim p(\cdot|S_t, A_t)$. The return at time step t is the random variable

$$G_t \doteq \sum_{i=0}^{\infty} \gamma^i R_{t+i+1},$$

which allows us to define state- and action-value functions v_π and q_π as

$$\begin{aligned} v_\pi(s) &\doteq \mathbb{E}[G_t|S_t = s, \pi, p], \\ q_\pi(s, a) &\doteq \mathbb{E}[G_t|S_t = s, A_t = a, \pi, p]. \end{aligned}$$

The performance of the policy π is measured by the expected discounted sum of rewards

$$J(\pi; p_0) \doteq \sum_s p_0(s) v_\pi(s).$$

Prediction and control are two fundamental tasks of RL.

The goal of prediction is to estimate the values v_π or q_π. Take estimating q_π as an example. Let $q_t \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$ be our estimate for q_π at time t. SARSA (Rummery and

Niranjan, 1994) updates $\{q_t\}$ iteratively as

$$\delta_t \doteq R_{t+1} + \gamma q_t(S_{t+1}, A_{t+1}) - q_t(S_t, A_t),$$
$$q_{t+1}(s,a) \doteq \begin{cases} q_t(s,a) + \alpha_t \delta_t, & (s,a) = (S_t, A_t) \\ q_t(s,a), & (s,a) \neq (S_t, A_t) \end{cases}, \tag{1}$$

where δ_t is called the temporal difference error (Sutton, 1988) and $\{\alpha_t\}$ is a sequence of learning rates. It is proved by Bertsekas and Tsitsiklis (1996) that, under mild conditions, $\{q_t\}$ converges to q_π almost surely. So far we have considered on-policy learning, where the policy of interest is the same as the policy used in action selection. In the off-policy learning setting, the goal is still to estimate q_π. Action selection is, however, done using a different policy μ (i.e., $A_t \sim \mu(\cdot|S_t)$). We refer to π and μ as the target and behavior policy respectively. Off-policy expected SARSA (Asis et al., 2018) updates $\{q_t\}$ iteratively as

$$\delta_t \doteq R_{t+1} + \gamma \sum_{a'} \pi(a'|S_{t+1}) q_t(S_{t+1}, a') - q_t(S_t, A_t),$$
$$q_{t+1}(s,a) \doteq \begin{cases} q_t(s,a) + \alpha_t \delta_t, & (s,a) = (S_t, A_t) \\ q_t(s,a), & (s,a) \neq (S_t, A_t) \end{cases},$$

where the target policy π, instead of the behavior policy μ, is used to compute the temporal difference error.

The goal of control is to find a policy π_* such that $\forall \pi, s$

$$v_\pi(s) \leq v_{\pi_*}(s). \tag{2}$$

One common approach for control is policy gradient. In this paper, we consider a softmax parameterization for the policy π. Letting $\theta \in \mathbb{R}^{|\mathcal{S} \times \mathcal{A}|}$ be the parameters of the policy, We represent it as

$$\pi(a|s) \doteq \frac{\exp(\theta_{s,a})}{\sum_{a'} \exp\big(\theta_{s,a'}\big)},$$

where $\theta_{s,a}$ is the (s,a)-indexed element of θ. Policy gradient methods then update θ iteratively as

$$\theta_{t+1} \doteq \theta_t + \beta_t \nabla_\theta J(\pi_{\theta_t}; p_0). \tag{3}$$

Here $\{\beta_t\}$ is a sequence of learning rates and π_θ emphasizes the dependence of the policy π on its parameter θ. In the rest of the paper, we omit the θ in ∇_θ for simplicity. Agarwal et al. (2020); Mei et al. (2020) prove that when $p_0(s) > 0$ holds for all s and $\{\beta_t\}$ is set properly, the iterates $\{\theta_t\}$ generated by (3) satisfy

$$\lim_{t \to \infty} J(\theta_t; p_0) = J(\pi_*; p_0),$$

confirming the optimality of policy gradient methods in the tabular setting with exact gradients. Mei et al. (2020) also establish a convergence rate for the softmax parameterization.

In practice, we, however, usually do not have access to $\nabla J(\pi_{\theta_t}; p_0)$. Fortunately, the policy gradient theorem (Sutton et al., 1999) asserts that

$$\nabla J(\pi_\theta; p_0) \doteq \frac{1}{1-\gamma} \sum_s d_{\pi_\theta,\gamma,p_0}(s) \sum_a q_{\pi_\theta}(s,a) \nabla \pi_\theta(a|s),$$

where

$$d_{\pi,\gamma,p_0} \doteq (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \Pr(S_t = s | p_0, \pi)$$

is the normalized discounted state occupancy measure. Hence instead of using the gradient update (3), practitioners usually consider the following approximate and stochastic gradient update for the on-policy setting:

$$\theta_{t+1} \doteq \theta_t + \beta_t \gamma^t q_t(S_t, A_t) \nabla \log \pi_{\theta_t}(A_t|S_t), \tag{4}$$

where q_t is updated according to (1). We refer to (4) and (1) as on-policy actor critic, where the actor refers to π_θ and the critic refers to q. Usually α_t is much larger than β_t, i.e., the critic is updated much faster than the actor and the actor is, therefore, quasi-stationary from the perspective of the critic. Consequently, in the limit, we can expect q_t to converge to $q_{\pi_{\theta_t}}$, after which $\gamma^t q_t(S_t, A_t) \nabla \log \pi_{\theta_t}(A_t|S_t)$ becomes an unbiased estimator of $\nabla J(\pi_{\theta_t}; p_0)$ and the actor update becomes the standard stochastic gradient ascent.

In the off-policy setting, at time step t, the action selection is done according to some behavior policy μ_{θ_t}. Here μ_θ does not need to have the same parameterization as π_θ, e.g., μ_θ can be a softmax policy with a different temperature, a mixture of a uniformly random policy and a softmax policy, or a constant policy μ. To account for the difference between π_θ and μ_θ, one must reweight the actor update (4) as

$$\theta_{t+1} \doteq \theta_t + \beta_t \varrho_t \rho_t q_t(S_t, A_t) \nabla \log \pi_{\theta_t}(A_t|S_t), \tag{5}$$

where

$$\rho_t \doteq \frac{\pi_{\theta_t}(A_t|S_t)}{\mu_{\theta_t}(A_t|S_t)}$$

is the importance sampling ratio to correct the discrepancy in action selection and

$$\varrho_t \doteq \frac{d_{\pi_{\theta_t},\gamma,p_0}(s)}{d_t(s)} \quad \text{with} \quad d_t(s) \doteq \Pr(S_t = s | \mu_{\theta_0}, \dots, \mu_{\theta_t})$$

is the density ratio to correct the discrepancy in state distribution. Thanks to ρ_t and ϱ_t, in the limit, (5) is still a stochastic gradient ascent algorithm following the gradient $\nabla J(\pi_{\theta_t}; p_0)$ if q_t converges to $q_{\pi_{\theta_t}}$. Theoretical analysis of variants of (5) includes Liu et al. (2019); Zhang et al. (2020c); Huang and Jiang (2021); Xu et al. (2021). Practitioners, however, usually use only ρ_t but completely ignore ϱ_t, yielding variants of

$$\theta_{t+1} \doteq \theta_t + \beta_t \rho_t q_t(S_t, A_t) \nabla \log \pi_{\theta_t}(A_t|S_t). \tag{6}$$

Clearly, (6) can no longer be regarded as a stochastic gradient ascent algorithm even if q_t converges to $q_{\pi_{\theta_t}}$ because of the missing term ϱ_t used to correct the state distribution. Still, variants of (6) enjoy great empirical success (Wang et al., 2017; Espeholt et al., 2018; Vinyals et al., 2019; Schmitt et al., 2020; Zahavy et al., 2020). To understand the behavior of (6), Laroche and Tachet (2021) study the following update rule:

$$\theta_{t+1} \doteq \theta_t + \beta_t \sum_s d_t(s) \sum_a q_{\pi_{\theta_t}}(s,a) \nabla \pi_{\theta_t}(a|s). \tag{7}$$

Different from (6), where the update step is approximate and stochastic, the update in (7) is exact and expected. Laroche and Tachet (2021) prove that under mild conditions, the iterates $\{\theta_t\}$ generated by (7) satisfy

$$\lim_{t\to\infty} J(\pi_{\theta_t}; p_0) = J(\pi_*; p_0).$$

If we further assume the optimal policy π_* is unique and $\inf_{s,t} d_t(s) > 0$, a nonasymptotic convergence rate of (7) is available.

3. Stochastic Approximation with Time-Inhomogeneous Operators on Time-Inhomogeneous Markov Chains

In this section, we provide finite sample analysis of a generic stochastic approximation algorithm with time-inhomogeneous update operators on time-inhomogeneous Markov chains. The results presented in this section are used in the analysis of critics in the rest of this work and may be of independent interest.

To motivate this part, consider using off-policy expected SARSA to update the critic in off-policy actor critic. We have

$$\begin{aligned}
\delta_t \doteq & R_{t+1} + \gamma \sum_{a'} \pi_{\theta_t}(a'|S_{t+1}) q_t(S_{t+1}, a') - q_t(S_t, A_t), \\
q_{t+1}(s,a) \doteq & \begin{cases} q_t(s,a) + \alpha_t \delta_t, & (s,a) = (S_t, A_t) \\ q_t(s,a), & (s,a) \neq (S_t, A_t) \end{cases}.
\end{aligned}$$

Equivalently, we can rewrite the above update in a more compact form as

$$q_{t+1} = q_t + \alpha_t \left(F_{\theta_t}(q_t, S_t, A_t, S_{t+1}) - q_t \right), \tag{8}$$

where

$$\begin{aligned}
F_\theta(q, s_0, a_0, s_1)[s,a] \doteq & \mathbb{I}_{(s_0,a_0)=(s,a)} \delta_\theta(q, s_0, a_0, s_1) + q(s,a), \\
\delta_\theta(q, s_0, a_0, s_1) \doteq & r(s_0, a_0) + \gamma \sum_{a_1} \pi_\theta(a_1|s_1) q(s_1, a_1) - q(s_0, a_0).
\end{aligned}$$

Here, $\mathbb{I}_{statement}(\cdot)$ is the indicator function whose value is 1 if the statement is true, and 0 otherwise. The update (8) motivates us to study a generic stochastic approximation algorithm in the form of

$$w_{t+1} \doteq w_t + \alpha_t (F_{\theta_t}(w_t, Y_t) - w_t + \epsilon_t). \tag{9}$$

Here $\left\{w_t \in \mathbb{R}^K\right\}$ are the iterates generated by the stochastic approximation algorithm, $\{Y_t\}$ is a sequence of random variables evolving in a finite space $\mathcal{Y}$, $\left\{\theta_t \in \mathbb{R}^L\right\}$ is another sequence of random variables controlling the transition of $\{Y_t\}$, F_θ is a function from $\mathbb{R}^K \times \mathcal{Y}$ to $\mathbb{R}^K$ parameterized by θ, and $\left\{\epsilon_t \in \mathbb{R}^K\right\}$ is a sequence of random noise. The analysis of critics in this paper only requires $\epsilon_t \equiv 0$. Nevertheless, we consider a generic noise process $\{\epsilon_t\}$ for generality.

The results in this section extend Theorem 2.1 of Chen et al. (2021) in two aspects. First, the operator F_θ changes every time step due to the change of θ, while Chen et al. (2021) consider a fixed operator F. Second, the random process $\{Y_t\}$ evolves according to time-varying dynamics controlled by $\{\theta_t\}$, while Chen et al. (2021) assume $\{Y_t\}$ is a Markov chain with fixed dynamics. The introduction of $\{\theta_t\}$ makes our results immediately applicable to the analysis of actor-critic algorithms. We now state our assumptions. It is worth reiterating that all the $\{\theta_t\}$ below refers to the random sequence used in the update (9).

Assumption 3.1 *(Time-inhomogeneous Markov chain) There exists a family of parameterized transition matrices $\Lambda_P \doteq \left\{P_\theta \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{Y}|} | \theta \in \mathbb{R}^L\right\}$ such that*

$$\Pr(Y_{t+1} = y) = P_{\theta_{t+1}}(Y_t, y).$$

Assumption 3.2 *(Uniform ergodicity) Let $\bar{\Lambda}_P$ be the closure of Λ_P. For any $P \in \bar{\Lambda}_P$, the chain induced by P is ergodic. We use d_θ to denote the invariant distribution of the chain induced by P_θ.*

Assumption 3.1 prescribes that the random process $\{Y_t\}$ is a time-inhomogeneous Markov chain. It is worth mentioning that Assumption 3.1 does not prescribe how the transition matrices depend on $\{\theta_t\}$. It does not restrict $\{\theta_t\}$ to be deterministic either. An exemplary parameterization we use in the context of off-policy actor critic will be shown later in (13). Assumption 3.2 prescribes the ergodicity of the Markov chains we consider and was also previously used in the analysis of RL algorithms both in the on-policy (Marbach and Tsitsiklis, 2001) and off-policy settings (Zhang et al., 2021). We will show later that Assumption 3.2 is easy to fulfill in our off-policy actor critic setting. Assumption 3.2 implicitly claims that all the matrices in $\bar{\Lambda}_P$ are stochastic matrices. This is indeed trivial to prove. Pick any $P_\infty \in \bar{\Lambda}_P$. Since $\bar{\Lambda}_P$ is the closure of Λ_P, there must exist a sequence $\{P_n\}$ such that $P_n \in \Lambda_P$ and $\lim_{n\to\infty} P_n = P_\infty$. It is then easy to see that $P_\infty(y, y') \in [0, 1]$ and

$$\sum_{y'} P_\infty(y, y') = \sum_{y'} \lim_{n\to\infty} P_n(y, y') = \lim_{n\to\infty} \sum_{y'} P_n(y, y') = 1.$$

In other words, P_∞ is a stochastic matrix. One important consequence of Assumption 3.2 is uniform mixing.

Lemma 1 *(Uniform ergodicity implies uniform mixing) Let Assumption 3.2 hold. Then, there exist constants $C_0 > 0$ and $\tau \in (0, 1)$, independent of θ, such that for any $n > 0$,*

$$\sup_{y,\theta} \sum_{y'} \left| P_\theta^n(y, y') - d_\theta(y') \right| \le C_0 \tau^n. \tag{10}$$

The proof of Lemma 1 is provided in Section A.1. The result in Lemma 1 is referred to as uniform mixing since it demonstrates that for any θ, the chain induced by P_θ mixes geometrically fast, with a common rate τ. For a specific θ, the existence of a θ-dependent mixing rate τ_θ is a well-known result when the chain is ergodic, see, e.g., Theorem 4.9 of Levin and Peres (2017). In Lemma 1, we further conclude to the existence of a θ-independent rate. The ergodicity on the closure $\bar{\Lambda}_P$ is key to our proof. If we make ergodicity assumption only on Λ_P, it might be possible to find a sequence $\{\theta_t\}$ such that the corresponding rates $\{\tau_{\theta_t}\}$ converges to 1. We remark that (10) usually appears as a technical assumption directly in many existing works concerning time-inhomogeneous Markov chains, see, e.g., Zou et al. (2019); Wu et al. (2020). In this paper, we prove that (10) is a consequence of Assumption 3.2, with the help of the extreme value theorem exploiting the compactness of $\bar{\Lambda}_P$. We will show in the next section that Assumption 3.2 can easily be fulfilled.

Assumption 3.3 *(Uniform contraction) For any $\theta \in \mathbb{R}^L$, define $F_\theta : \mathbb{R}^K \to \mathbb{R}^K$ as*

$$\bar{F}_\theta(w) \doteq \sum_{y \in \mathcal{Y}} d_\theta(y) F_\theta(w, y).$$

Then, there exists a constant $\kappa \in (0, 1)$ and a norm $\|\cdot\|_c$ such that for all θ, w, w',

$$\left\|\bar{F}_\theta(w) - \bar{F}_\theta(w')\right\|_c \le \kappa \left\|w - w'\right\|_c.$$

We use w^*_θ to denote the unique fixed point of $\bar{F}_\theta$.

The existence and uniqueness of w^*_θ follows from the Banach fixed point theorem. Assumption 3.3 is another major development beyond Chen et al. (2021). The fact that both $\|\cdot\|_c$ and κ are independent of θ makes it possible to design a Lyapunov function for our time-inhomogeneous Markov chain. We will show later that our critic updates indeed satisfy this uniform contraction assumption.

Assumption 3.4 *(Continuity and boundedness) There exist positive constants $L_F, L'_F, L''_F, U_F, U'_F, U''_F, L_w, U_w, L_P$ such that for any w, w', y, y' and any time step t, k, almost surely,*

(i). $\|F_{\theta_t}(w, y) - F_{\theta_t}(w', y)\|_c \le L_F \|w - w'\|_c$

(ii). $\|F_{\theta_t}(w, y) - F_{\theta_k}(w, y)\|_c \le L'_F \|\theta_t - \theta_k\|_c (\|w\|_c + U'_F)$

(iii). $\|F_{\theta_t}(0, y)\|_c \le U_F$

(iv). $\left\|\bar{F}_{\theta_t}(w) - \bar{F}_{\theta_k}(w)\right\|_c \le L''_F \|\theta_t - \theta_k\|_c (\|w\|_c + U''_F)$

(v). $\left\|w^*_{\theta_t} - w^*_{\theta_k}\right\|_c \le L_w \|\theta_t - \theta_k\|_c$

(vi). $\sup_t \left\|w^*_{\theta_t}\right\|_c \le U_w$

(vii). $|P_{\theta_t}(y, y') - P_{\theta_k}(y, y')| \le L_P \|\theta_t - \theta_k\|_c$

Assumption 3.5 *(Noise) Let $\mathcal{F}_t$ be the σ-algebra generated by $\{(w_i, Y_i, \epsilon_i, \theta_i)\}_{0 \leq i \leq t-1} \cup \{w_t, \theta_t\}$, we have*

(i). $\mathbb{E}\left[\epsilon_t \mid \mathcal{F}_t\right] = 0, \forall t$

(ii). There exist positive constants U_ϵ, U'_ϵ such that $\forall t$, $\left\|\epsilon_t\right\|_c \leq U_\epsilon \left\|w_t\right\|_c + U'_\epsilon$

Assumptions 3.4 and 3.5 are natural extensions of the counterparts in Chen et al. (2021) from time-homogeneous to time-inhomogeneous Markov chains and from time-homogeneous to time-inhomogeneous operators.

Assumption 3.6 *(Two timescales) The learning rate $\{\alpha_t\}$ has the form*

$$\alpha_t \doteq \frac{\alpha}{(t+t_0)^{\epsilon_\alpha}},$$

where $\epsilon_\alpha \in (0.5, 1), \alpha > 0, t_0 > 0$ are constants to be tuned. Define another sequence $\{\beta_t\}$ such that

$$\beta_t \doteq \frac{\beta}{(t+t_0)^{\epsilon_\beta}},$$

where $\epsilon_\beta \in (\epsilon_\alpha, 1], \beta \in (0, \alpha)$ are constants to be tuned. Then there exists a constant $L_\theta > 0$ such that $\forall t$, almost surely,

$$\left\|\theta_{t+1} - \theta_t\right\|_c \leq \beta_t L_\theta. \tag{11}$$

Assumption 3.6 ensures that the iterates $\{w_t\}$ evolve sufficiently faster than the change in the dynamics of the chain (i.e., the change of $\{\theta_t\}$). In the off-policy actor critic setting we consider in next section, $\{\alpha_t\}$ and $\{\beta_t\}$ are the learning rates for the critic and the actor respectively. Though Assumption 3.6 explicitly prescribes the form of the sequences $\{\alpha_t\}$ and $\{\beta_t\}$, those are indeed only one of many possible forms (one could e.g., use different t_0 for $\{\alpha_t\}$ and $\{\beta_t\}$), we consider these particular forms to ease presentation. We remark that condition in (11) is also used in Konda (2002), which gives the asymptotic convergence analysis of the canonical on-policy actor critic with linear function approximation. We are now ready to state our main results.

Theorem 2 *Let Assumptions 3.1 - 3.6 hold. For any*

$$\epsilon_w \in (0, \min\left\{2(\epsilon_\beta - \epsilon_\alpha), \epsilon_\alpha\right\}),$$

if t_0 is sufficiently large, then $\forall t$,

$$\mathbb{E}\left[\left\|w_t - w^*_{\theta_t}\right\|_c^2\right] = \mathcal{O}\left(\frac{1}{(t+t_0)^{\epsilon_w}}\right).$$

See Section A.2 for the proof of Theorem 2 and the constants hidden by $\mathcal{O}(\cdot)$. In particular, we clearly document t_0's dependencies. One could alternatively set t_0 to 0, then the convergence rate in Theorem 2 applies only for sufficiently large t. When both the Markov chain and the update operator are time-homogeneous, Chen et al. (2021) demonstrate a convergence rate $\mathcal{O}\left(\frac{1}{t^{\epsilon_\alpha}}\right)$. When $2(\epsilon_\beta - \epsilon_\alpha) > \epsilon_\alpha$ holds, our convergence rate of $\mathcal{O}\left(\frac{1}{t^{\epsilon_w}}\right)$ can be arbitrarily close to $\mathcal{O}\left(\frac{1}{t^{\epsilon_\alpha}}\right)$.

4. Off-Policy Actor Critic with Decaying KL Regularization

We analyze the optimality of an off-policy actor critic algorithm without correction of the state distribution mismatch (Algorithm 1). Our analysis provides, to some extent, a theoretical justification for the practice of ignoring this correction.

Algorithm 1: Off-Policy Actor-Critic with Decaying KL Regularization

$S_0 \sim p_0(\cdot)$
$t \leftarrow 0$
while *True* **do**
  Sample $A_t \sim \mu_{\theta_t}(\cdot|S_t)$
  Execute A_t, get R_{t+1}, S_{t+1}
  $\delta_t \leftarrow R_{t+1} + \gamma \sum_{a'} \pi_{\theta_t}(a'|S_{t+1}) q_t(S_{t+1}, a') - q_t(S_t, A_t)$
  $q_{t+1}(s,a) \leftarrow \begin{cases} q_t(s,a) + \alpha_t \delta_t, & (s,a) = (S_t, A_t) \\ q_t(s,a), & \text{otherwise} \end{cases}$
  $\rho_t \leftarrow \frac{\pi_{\theta_t}(A_t|S_t)}{\mu_{\theta_t}(A_t|S_t)}$
  $\theta_{t+1} \leftarrow \theta_t + \beta_t \big(\rho_t \nabla_\theta \log \pi_{\theta_t}(A_t|S_t) \Pi(q_t(S_t, A_t)) - \lambda_t \nabla_\theta \text{KL}\left(\mathcal{U}_\mathcal{A} || \pi_{\theta_t}(\cdot|S_t)\right)\big)$
  $t \leftarrow t + 1$
end

In Algorithm 1, the target policy π_θ is a softmax policy. At each time step t, we sample an action A_t according to the behavior policy μ_{θ_t}. Importantly, though the behavior policy is also solely determined by θ, the parameterization of μ_θ can be arbitrarily different from π_θ. After obtaining the reward R_{t+1} and the successor state S_{t+1}, we update the critic with off-policy expected SARSA, where π_{θ_t} is used as the target policy for bootstrapping. We then update the actor similarly to (6) without correcting the state distribution mismatch. The update to θ_t in Algorithm 1 is different from (6) in two aspects. First, we use a projection $\Pi : \mathbb{R} \to \mathbb{R}$ in case the critic becomes too large:

$$\Pi(x) \doteq \begin{cases} x, & |x| \leq C_\Pi \\ C_\Pi \frac{x}{|x|}, & \text{otherwise} \end{cases},$$

where $C_\Pi \doteq \frac{r_{max}}{1-\gamma}$. Second, we use the KL divergence between a uniformly randomly distribution $\mathcal{U}_\mathcal{A}$ and the current policy $\pi_{\theta_t}(\cdot|S_t)$ as regularization, with a decaying weight λ_t. The KL divergence is introduced to ensure that the target policy π_θ is sufficiently explorative such that there are no bad stationary points (c.f. Theorem 5.2 of Agarwal et al. (2020)). In practice, the entropy of the policy is often used to regularize the policy update (Williams and Peng, 1991; Mnih et al., 2016). Here we use the KL divergence instead of the entropy mainly for technical consideration. We refer the reader to Remark 5.2 of Agarwal et al. (2020) for more discussion about this choice. The decaying weight λ_t is introduced to ensure that, in the limit, the target policy π_θ can still converge to a deterministic policy, which is a necessary condition for optimality.

Algorithm 1 runs in three timescales. The critic runs in the fastest timescale such that it can provide accurate signal for the actor update, which runs in the middle timescale. It is then expected that the actor would converge to stationary points whose suboptimality is controlled by λ_t, which decays in the slowest timescale. Finally, as λ_t diminishes, the suboptimality of the actor decays to 0. To achieve this three timescale setting, we make the following assumptions.

Assumption 4.1 *(Three timescales) The learning rates* $\{\alpha_t\}, \{\beta_t\}$ *and the weights of KL regularization* $\{\lambda_t\}$ *have the forms*

$$\alpha_t \doteq \frac{\alpha}{(t+t_0)^{\epsilon_\alpha}}, \, \beta_t \doteq \frac{\beta}{(t+t_0)^{\epsilon_\beta}}, \, \lambda_t \doteq \frac{\lambda}{(t+t_0)^{\epsilon_\lambda}},$$

where $0.5 < \epsilon_\alpha < \epsilon_\beta \leq 1, \epsilon_\lambda > 0, \alpha > \beta > 0, \lambda > 0, t_0 > 0$ *are constants to be tuned.*

Assumption 4.2 *(Learning rates)* $2(1-\epsilon_\beta) < \min\{2(\epsilon_\beta - \epsilon_\alpha), \epsilon_\alpha\}, 0 \leq \epsilon_\lambda < \frac{1-\epsilon_\beta}{2}$

We remark that Assumptions 4.1 and 4.2 are only one of many possible forms of learning rates and we choose this particular form to ease presentation. To ensure each update to θ_t does not change the dynamics of the induced Markov chain too fast, we impose the following assumption on the parameterization of μ_θ.

Assumption 4.3 *(Lipschitz continuity) There exists* $L_\mu > 0$ *such that* $\forall \theta, \theta', a, s$,

$$\|\mu_\theta(a|s) - \mu_{\theta'}(a|s)\| \leq L_\mu \big\|\theta - \theta'\big\|.$$

We remark that given the softmax parameterization of π_θ, it is well-known (see, e.g., Lemma 1 of Wang and Zou 2020) that π_θ is also Lipschitz continuous, i.e., there exists $L_\pi > 0$ such that $\forall \theta, \theta', a, s$

$$\|\pi_\theta(a|s) - \pi_{\theta'}(a|s)\| \leq L_\pi \big\|\theta - \theta'\big\|. \tag{12}$$

To ensure sufficient exploration, we impose the following assumption on the behavior policy.

Assumption 4.4 *(Uniform ergodicity) Let* $\bar{\Lambda}_\mu$ *be the closure of* $\left\{\mu_\theta \mid \theta \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}\right\}$. *For any* $\mu \in \bar{\Lambda}_\mu$, *the chain induced by* μ *is ergodic and* $\mu(a|s) > 0$.

Assumption 4.4 is easy to fulfill in practice. Assuming the chain induced by a uniformly random policy is ergodic, which we believe is a necessary condition for any assumption regarding ergodicity, one possible choice for μ_θ is to mix an arbitrary behavior policy μ'_θ satisfying the Lipschitz continuous requirement with the uniformly random policy, i.e.,

$$\mu_\theta(\cdot|s) \doteq (1-\epsilon)\mathcal{U}_\mathcal{A} + \epsilon\mu'_\theta(\cdot|s) \tag{13}$$

with any $\epsilon \in (0,1)$. From now on, we use $d_\mu \in \mathbb{R}^{|\mathcal{S}|}$ to denote the invariant state distribution of the chain induced by a policy μ and also overload $d_\mu \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$ to denote the invariant state action distribution under policy μ. With all assumptions stated, we are ready to present our convergence results.

4.1 Convergence of the Critic

In this section, we study the convergence of the critic by invoking Theorem 2 with the update to q_t in Algorithm 1 expressed as (8). Assumption 3.3 requires us to study the expected operator

$$\bar{F}_\theta(q) \doteq \sum_{s,a,s'} d_{\mu_\theta}(s)\mu_\theta(a|s)p(s'|s,a)F_\theta(q,s,a,s').$$

Simple algebraic manipulation yields

$$\begin{aligned}\bar{F}_\theta(q) &= D_{\mu_\theta}(r+\gamma P_{\pi_\theta}q - q) + q \\ &= (I - D_{\mu_\theta}(I-\gamma P_{\pi_\theta}))q + D_{\mu_\theta} r,\end{aligned} \tag{14}$$

where $D_{\mu_\theta} \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|\times|\mathcal{S}\times\mathcal{A}|}$ is a diagonal matrix with $D_{\mu_\theta}((s,a),(s,a)) \doteq d_{\mu_\theta}(s)\mu_\theta(a|s)$ and $P_{\pi_\theta} \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|\times|\mathcal{S}\times\mathcal{A}|}$ is the state-action pair transition matrix under policy π_θ, i.e.,

$$P_{\pi_\theta}((s,a),(s',a')) \doteq p(s'|s,a)\pi_\theta(a'|s').$$

We now verify Assumption 3.3 with Lemma 3.

Lemma 3 *(Uniform contraction) Let Assumption 4.4 hold. Then, there exists an ℓ_p norm and a constant $\kappa \in (0,1)$ such that for any $\theta, q, q' \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$,*

$$\left\|\bar{F}_\theta(q) - \bar{F}_\theta(q')\right\|_p \leq \kappa \left\|q - q'\right\|_p.$$

Further, q_{π_θ} is the unique fixed point of $\bar{F}_\theta$.

The proof of Lemma 3 is provided in Section B.1. Next, we are able to prove the convergence of the critic.

Proposition 4 *(Convergence of the critic) Let Assumptions 4.1, 4.3, and 4.4 hold. For any*

$$\epsilon_q \in (0, \min\{2(\epsilon_\beta - \epsilon_\alpha), \epsilon_\alpha\}),$$

if t_0 is sufficiently large, the iterates $\{q_t\}$ generated by Algorithm 1 satisfy

$$\mathbb{E}\left[\left\|q_t - q_{\pi_{\theta_t}}\right\|_p^2\right] = \mathcal{O}\left(\frac{1}{t^{\epsilon_q}}\right).$$

The proof of Proposition 4 is provided in Section B.2. Proposition 4 confirms that the critic is able to track the true value function in the limit, where the dependence between the convergence rate and the mixing parameter of the Markov chains are hidden in $\mathcal{O}(\cdot)$. Similar trackability has also been established in Konda (2002); Zhang et al. (2020c); Wu et al. (2020). Those, however, rely on the uniform negative-definiteness of the limiting update matrix. Konda (2002) proves that the uniform negative-definiteness holds in the on-policy actor critic with linear function approximation (Lemma 4.18 of Konda 2002) and establishes this trackability asymptotically. Wu et al. (2020) assume the uniform negative-definiteness holds (the second half of Assumption 4.1 of Wu et al. 2020) in the on-policy actor critic with linear function approximation and establish this trackability nonasymptotically. Zhang et al. (2020c) achieve this uniform negative-definiteness via introducing extra ridge regularization and using full gradients (c.f. Gradient TD, Sutton et al. 2009) instead of semi-gradients (c.f. TD, Sutton 1988) for the critic update in the off-policy actor critic with function approximation and achieve this trackability asymptotically. In our off-policy actor critic setting, the limiting update matrix of the critic can be computed as

$$D_{\mu_\theta}(\gamma P_{\pi_\theta} - I).$$

To achieve the desired uniform negative-definiteness, we would need to prove that there exists a constant $\zeta > 0$ such that for all x, θ,

$$x^\top D_{\mu_\theta}(\gamma P_{\pi_\theta} - I)x \leq -\xi \|x\|^2.$$

We, however, do not expect the above inequality to hold without making strong assumptions. Instead, we resort to uniform contraction. As demonstrated by Lemma 3 and Proposition 4, uniform contraction is indeed an effective alternative tool for establishing such trackability. Moreover, Khodadadian et al. (2022) establish this trackability for a natural actor critic (Kakade, 2001) with a Lyapunov method in a quasi-off-policy setting. The setting Khodadadian et al. (2022) consider is a quasi-off-policy setting in that they prescribe a special form of the behavior policy such that the difference between the behavior policy and the target policy diminishes as time progresses. By contrast, we work on a general off-policy setting in that at any time step the behavior policy can always be arbitrarily different from the target policy. A weaker trackability of the critic can be obtained with the results from Chen et al. (2021) directly without using our extension (i.e., Theorem 2), as done by Chen et al. (2022); Khodadadian et al. (2021) in their analysis of a natural actor critic. However, since Chen et al. (2021) require both the dynamics of the Markov chain and the update operator to be fixed, Chen et al. (2022); Khodadadian et al. (2021) have to keep both the behavior policy and the target policy (actor) fixed when updating the critic. That being said, Chen et al. (2022); Khodadadian et al. (2021) have an inner loop for updating the critic and an outer loop for updating the actor. For the critic to be sufficiently accurate, the inner loop has to take sufficiently many steps. Chen et al. (2022); Khodadadian et al. (2021), therefore, have a flavor of bi-level optimization. Further, as long as the steps of the inner loop is finite, the bias from using a learned critic instead of the true value function will not diminish in the limit. This bias eventually translates into a suboptimality of the policy that will not vanish in the limit. By contrast, Theorem 2 allows us to consider multi-timescales directly without incurring nested loops, which ensures that the bias from the critic diminishes in the limit.

4.2 Convergence of the Actor

With the critic able to track the true value function, we are now ready to present the optimality of the actor.

Theorem 5 *(Optimality of the actor) Let Assumptions 4.1 - 4.4 hold. Fix*

$$\epsilon_q \in \Big(2(1 - \epsilon_\beta), \min\{2(\epsilon_\beta - \epsilon_\alpha), \epsilon_\alpha\}\Big).$$

Let t_0 be sufficiently large. For the iterates $\{\theta_t\}$ generated by Algorithm 1 and any $t > 0$, if k is uniformly randomly selected from the set $\left\{\lceil \frac{t}{2} \rceil, \lceil \frac{t}{2} \rceil + 1, \dots, t\right\}$ where $\lceil \cdot \rceil$ is the ceiling function, then

$$J(\pi_{\theta_k}; p_0) \geq J(\pi_*; p_0) - \mathcal{O}\left(\lambda_k\right) \tag{15}$$

holds with probability at least

$$1 - \mathcal{O}\left(\frac{1}{t^{1-\epsilon_\beta-2\epsilon_\lambda}} + \frac{\log^2 t}{t^{\epsilon_\beta-2\epsilon_\lambda}} + \frac{1}{t^{\epsilon_q-2\epsilon_\lambda}}\right), \tag{16}$$

where π_* can be any optimal policy.

The proof of Theorem 5 is provided in Section B.3. We remark that the $\frac{1}{2}$ in $\lceil \frac{t}{2} \rceil$ is purely ad-hoc. We can use any positive constant smaller than 1 and the new rate will be different from the current one in only the constants hidden by $\mathcal{O}(\cdot)$. We now optimize the selection of ϵ_α and ϵ_β. Let ϵ_0 by any positive scalar sufficiently close to 0 and set

$$\epsilon_\beta = \frac{3}{4} + \epsilon_0,\, \epsilon_\alpha = \frac{1}{2} + \epsilon_0,\, \epsilon_q = \frac{1}{2} - \epsilon_0. \tag{17}$$

Then the high probability in (16) becomes

$$1 - \mathcal{O}\left(t^{-\left(\frac{1}{4}-\epsilon_0-2\epsilon_\lambda\right)}\right)$$

and the suboptimality in (15) remains

$$J(\pi_{\theta_k}; p_0) \geq J(\pi_*; p_0) - \mathcal{O}\left(k^{-\epsilon_\lambda}\right).$$

It now becomes clear that the selection of

$$\epsilon_\lambda \in (0, \frac{1}{8}) \tag{18}$$

trades off suboptimality and high probability. When ϵ_λ is large, the suboptimality diminishes quickly but the high probability approaches one slowly and vice versa. To our best knowledge, Theorem 5 is the first to establish the global optimality and convergence rate of a naive off-policy actor critic algorithm without density ratio correction even in the tabular setting. We leave the improvement of the convergence rate for future work.

Importantly, Theorem 5 does not make any assumption on the initial distribution p_0. By contrast, to obtain the asymptotic optimality in Agarwal et al. (2020) or to obtain the convergence rate in Mei et al. (2020), $p_0(s) > 0$ is assumed to hold for all states. Both Agarwal et al. (2020) and Mei et al. (2020) leave it an open problem whether $p_0(s) > 0$ is a necessary condition for optimality. Our results show that at least in the off-policy setting, this is not necessary. The intuition is simple. Let p_0' be another initial distribution such that $p_0'(s) > 0$ holds for all states. Then we could optimize $J(\pi_\theta; p_0')$ instead of $J(\pi_\theta; p_0)$ since the optimal policy w.r.t. $J(\pi_\theta; p_0')$ must also be optimal w.r.t. $J(\pi_\theta; p_0)$. To optimize $J(\pi_\theta; p_0')$, we would need samples starting from p_0', which is impractical in the on-policy setting since the initial distribution of the MDP is p_0. In the off-policy setting, we can, however, use samples starting from p_0' and make corrections with the density ratio. Since our results show that density ratio correction actually does not matter in the tabular setting we consider, we can then simply ignore the density ratio, yielding Algorithm 1. Agarwal et al. (2020); Mei et al. (2020) refer to the assumption $p_0(s) > 0$ as the sufficient exploration assumption. Unfortunately, the initial distribution p_0 is usually considered as part of the problem and thus is not controlled by the user. In our off-policy setting, we instead achieve sufficient exploration by making assumptions on the behavior policy (Assumption 4.4), which demonstrates the flexibility of off-policy learning in terms of exploration. Moreover, to obtain the nonasymptotic convergence rate of the off-policy actor critic with exact update, Laroche and Tachet (2021) require the optimal policy π_* to be unique. By contrast, Theorem 5 does not assume any such uniqueness.

5. Soft Actor Critic

In this section, we study the convergence of soft actor critic in the framework of maximum entropy RL, which penalizes deterministic policies via adding the entropy of the policy into the reward (Williams and Peng, 1991; Mnih et al., 2016; Nachum et al., 2017; Haarnoja et al., 2018). The soft state value function of a policy π is defined as

$$\begin{aligned}\tilde{v}_{\pi,\eta}(s) \doteq& \mathbb{E}\left[\sum_{i=0}^{\infty}\gamma^i\Big(r(S_{t+i},A_{t+i})+\eta\mathbb{H}\left(\pi(\cdot|S_{t+i})\right)\Big)\mid S_t=s,\pi\right]\\ =&v_\pi(s)+\eta\mathbb{E}\left[\sum_{i=0}^{\infty}\gamma^i\mathbb{H}\left(\pi(\cdot|S_{t+i})\right)\mid S_t=s,\pi\right],\end{aligned}$$

where

$$\mathbb{H}\left(\pi(\cdot|s)\right)\doteq-\sum_a\pi(a|s)\log\pi(a|s)$$

is the entropy and $\eta \geq 0$ is the parameter controlling the strength of entropy regularization. Correspondingly, the soft action value function of a policy π is defined as

$$\tilde{q}_{\pi,\eta}(s,a)\doteq r(s,a)+\gamma\sum_{s'}p(s'|s,a)\tilde{v}_{\pi,\eta}(s'), \tag{19}$$

which satisfies the recursive equation

$$\tilde{q}_{\pi,\eta}(s,a)=r(s,a)+\gamma\sum_{s',a'}p(s'|s,a)\pi(a'|s')\left(\tilde{q}_{\pi,\eta}(s,a)-\eta\log\pi(a'|s')\right).$$

The entropy regularized discounted total rewards is then

$$\tilde{J}_\eta(\pi;p_0)\doteq\sum_s p_0(s)\tilde{v}_{\pi,\eta}(s)=J(\pi;p_0)+\frac{\eta}{1-\gamma}\sum_s d_{\pi,\gamma,p_0}(s)\mathbb{H}\left(\pi(\cdot|s)\right). \tag{20}$$

We still consider the softmax parameterization for the policy π. Similar to the canonical policy gradient theorem, it can be computed (Levine, 2018) that

$$\nabla\tilde{J}_\eta(\pi_\theta;p_0)=\frac{1}{1-\gamma}\sum_s d_{\pi_\theta,\gamma,p_0}(s)\sum_a\left(\tilde{q}_{\pi_\theta,\eta}(s,a)-\eta\log\pi_\theta(a|s)\right)\nabla\pi_\theta(a|s).$$

To get unbiased estimates of $\nabla \tilde{J}_\eta(\pi_\theta; p_0)$, one would need to sample states from $d_{\pi_\theta,\gamma,p_0}$, which is, however, impractical in off-policy settings. Practitioners, instead, directly use states obtained by following the behavior policy (see, e.g., Algorithm 2), yielding a distribution mismatch.

Algorithm 2: Expected Soft Actor-Critic

$S_0 \sim p_0(\cdot)$
$t \leftarrow 0$
while *True* **do**
  Sample $A_t \sim \mu_{\theta_t}(\cdot|S_t)$
  Execute A_t, get R_{t+1}, S_{t+1}
  $\delta_t \leftarrow R_{t+1} + \gamma \sum_{a'} \pi_{\theta_t}(a'|S_{t+1}) \left(q_t(S_{t+1}, a') - \lambda_t \log \pi_{\theta_t}(a'|S_{t+1})\right) - q_t(S_t, A_t)$
  $q_{t+1}(s,a) \leftarrow \begin{cases} q_t(s,a) + \alpha_t \delta_t, & (s,a) = (S_t, A_t) \\ q_t(s,a), & \text{otherwise} \end{cases}$
  $\theta_{t+1} \leftarrow \theta_t + \beta_t \sum_a \pi_{\theta_t}(a|S_t) \nabla_\theta \log \pi_{\theta_t}(a|S_t) \Big(\Pi(q_t(S_t, a)) - \lambda_t \log \pi_{\theta_t}(a|S_t)\Big)$
  $t \leftarrow t + 1$
end

In Algorithm 2, we still consider the learning rates specified in Assumption 4.1 and use a projection Π in the actor update with

$$C_\Pi \doteq \frac{r_{max} + \lambda \log |\mathcal{A}|}{1 - \gamma},$$

which is the maximum possible soft action value given our selection of λ_t since the entropy is always bounded by $\log |\mathcal{A}|$. We still consider Assumptions 4.3 and 4.4 for the behavior policy. Importantly, in Algorithm 2, we consider expected actor updates (Ciosek and Whiteson, 2020) that update the policy for all actions instead of just the executed action A_t. This is mainly for technical consideration. If we use stochastic update akin to Algorithm 1, the update to θ_t in Algorithm 2 will have the term $\log \pi_{\theta_t}(A_t|S_t)$. As $\{\lambda_t\}$ decreases over time, we would expect that π_{θ_t} becomes more and more deterministic. Consequently, $|\log \pi_{\theta_t}(A_t|S_t)|$ tends to go to infinity, imposing additional challenges in verifying (11) unless we ensure $\{\lambda_t\}$ decays sufficiently fast (e.g., using $\epsilon_\lambda > 1 - \epsilon_\beta$) such that $|\lambda_t \log \pi_{\theta_t}(A_t|S_t)|$ remains bounded. By using expected updates instead, we are able to verify (11) without imposing any additional condition on $\{\lambda_t\}$. We remark that Algorithm 2 makes expected updates across only actions. At each time step, Algorithm 2 still update the policy only for the current state. Algorithm 2 shares the same spirit of the canonical soft actor critic algorithm (Algorithm 1 in Haarnoja et al. 2018). Haarnoja et al. (2018) derive the canonical soft actor critic algorithm from a soft policy iteration perspective, where the policy evaluation of the soft value function and the policy improvement of the actor are performed alternatively. Importantly, during the soft policy iteration, both the policy evaluation and the policy improvement steps are assumed to be fully executed. By contrast, the soft actor critic algorithm conduct only several gradient steps for both the policy evaluation and the policy improvement. As a consequence, the results concerning the optimality of the soft policy iteration in Haarnoja et al. (2018) do not apply to soft actor critic. The convergence of soft actor critic with a fixed regularization weight (η) remains an open problem, and convergence with a decaying regularization, to optimality even more so. In this work, we instead derive the soft actor critic algorithm from the policy gradient perspective directly, akin to the canonical actor critic, and establish its convergence.[1]

We first study the convergence of $\{q_t\}$ in Algorithm 2. Different from Algorithm 1, the iterates $\{q_t\}$ now depend on not only θ_t but also λ_t. In light of this, we consider their concatenation and define

$$\zeta_t \doteq \begin{bmatrix}\lambda_t \\ \theta_t\end{bmatrix}, \zeta \doteq \begin{bmatrix}\eta \\ \theta\end{bmatrix}.$$

Here ζ is the placeholder for ζ_t used for defining functions. The update of $\{q_t\}$ in Algorithm 2 can then be expressed in a compact way as

$$q_{t+1} = q_t + \alpha_t(F_{\zeta_t}(q_t, S_t, A_t, S_{t+1}) - q_t),$$

where

$$\begin{aligned} F_\zeta(q, s_0, a_0, s_1)[s, a] &\doteq \delta_\zeta(q, s_0, a_0, s_1)\mathbb{I}_{(s_0,a_0)=(s,a)} + q(s, a), \\ \delta_\zeta(q, s_0, a_0, s_1) &\doteq r(s_0, a_0) + \gamma \sum_{a_1} \pi_\theta(a_1|s_1) \left(q(s_1, a_1) - \eta \log \pi_\theta(a_1|s_1)\right) - q(s_0, a_0). \end{aligned}$$

We can then establish the convergence of $\{q_t\}$ similarly to Proposition 4.

Proposition 6 *(Convergence of the critic) Let Assumptions 4.1, 4.3, and 4.4 hold. Then there exists an ℓ_p norm such that for any*

$$\epsilon_q \in (0, \min\{2(\epsilon_\beta - \epsilon_\alpha), \epsilon_\alpha\}),$$

if t_0 is sufficiently large, the iterates $\{q_t\}$ generated by Algorithm 1 satisfy

$$\mathbb{E}\left[\left\|q_t - \tilde{q}_{\pi_{\theta_t},\lambda_t}\right\|_p^2\right] = \mathcal{O}\left(\frac{1}{t^{\epsilon_q}}\right).$$

The proof of Proposition 6 is provided in Section C.1 and is more convoluted than that of Proposition 4 since we now need to verify the assumptions of Theorem 2 for the concatenated vector ζ_t instead of just θ_t. With the help of Proposition 6, we now establish the convergence of $\{\theta_t\}$, akin to Theorem 5.

Theorem 7 *(Convergence of the actor) Let Assumptions 4.1, 4.3, and 4.4 hold. Fix any*

$$\epsilon_q \in \Big(0, \min\{2(\epsilon_\beta - \epsilon_\alpha), \epsilon_\alpha\}\Big).$$

[1] Following existing works, e.g., Konda (2002); Zhang et al. (2020c); Wu et al. (2020); Xu et al. (2021), by convergence of the actor, we mean that the gradients converge to 0.

Let t_0 be sufficiently large. Fix any $\epsilon_0 > 0$ and any state distribution p_0'. For the iterates $\{\theta_t\}$ generated by Algorithm 2 and any $t > 0$, if k is uniformly randomly selected from the set $\left\{\lceil \frac{t}{2} \rceil, \lceil \frac{t}{2} \rceil + 1, \dots, t\right\}$, then

$$\left\|\nabla \tilde{J}_{\lambda_k}(\pi_{\theta_k}; p_0')\right\|^2 \leq \frac{1}{k^{\epsilon_0}}$$

holds with at least probability

$$1 - \mathcal{O}\left(\frac{1}{t^{1-\epsilon_\beta-\epsilon_0}} + \frac{\log^2 t}{t^{\epsilon_\beta-\epsilon_0}} + \frac{1}{t^{\epsilon_q-\epsilon_0}}\right).$$

The proof of Theorem 7 is provided in Section C.2. Theorem 7 confirms the convergence of the actor to stationary points, where the additional ϵ_0 trades off the rate at which the gradient vanishes and the rate at which the probability goes to one. This ϵ_0 is just to present the results and is not a hyperparameter of Algorithm 2. To our best knowledge, Theorem 7 is the first to establish the convergence of soft actor critic with a decaying entropy regularization weight.

Based on Theorem 7, the following corollary gives a *partial* result concerning the optimality of Algorithm 2.

Corollary 8 *(Optimality of the actor) Let Assumptions 4.1, 4.3, and 4.4 hold. Fix any*

$$\epsilon_q \in \Big(0, \min\left\{2(\epsilon_\beta - \epsilon_\alpha), \epsilon_\alpha\right\}\Big).$$

Let t_0 be sufficiently large. Let $\{\delta_t\}$ be any positive decreasing sequence converging to 0. For the iterates $\{\theta_t\}$ generated by Algorithm 2 and any $t > 0$, if k is uniformly randomly selected from the set $\left\{\lceil \frac{t}{2} \rceil, \lceil \frac{t}{2} \rceil + 1, \dots, t\right\}$, then

$$J(\pi_{\theta_k}; p_0) \geq J(\pi_*; p_0) - \mathcal{O}\left(\lambda_k\right) - \mathcal{O}\left(\frac{\delta_k}{\lambda_k \left(\min_{s,a} \pi_{\theta_k}(a|s)\right)^2}\right)$$

holds with at least probability

$$1 - \frac{\mathcal{O}\left(t^{-(1-\epsilon_\beta)} + t^{-\epsilon_\beta}\log^2 t + t^{-\epsilon_q}\right)}{\delta_t},$$

where π_* can be any optimal policy in (2).

The proof of Corollary 8 is provided in Section C.3. The sequence $\{\delta_t\}$ in Corollary 8 trades off the suboptimality and the high probability. For Corollary 8 to be nontrivial (i.e., the suboptimality diminishes and the high probability approaches one), one sufficient condition is that

$$\lim_{t\to\infty} \frac{t^{-(1-\epsilon_\beta)} + t^{-\epsilon_\beta}\log^2 t + t^{-\epsilon_q}}{\lambda_t \left(\min_{s,a} \pi_{\theta_t}(a|s)\right)^2} = 0. \tag{21}$$

This requires us to study the decay rate of $\min_{s,a} \pi_{\theta_t}(a|s)$. We conjecture that when λ_t decays slower, $\min_{s,a} \pi_{\theta_t}(a|s)$ also decays slower. Consequently, we expect (21) to hold when

λ_t decays sufficiently slow and the form of the learning rates α_t and β_t are adjusted correspondingly according to the form of $\min_{s,a} \pi_{\theta_t}(a|s)$'s decay rate. We leave the investigation of this rate for future work.

We remark that though Corollary 8 is only a partial result, it still advances the state of the art regarding the optimality of soft policy gradient (policy gradient in the maximum entropy RL framework) methods in Mei et al. (2020). Theorem 8 of Mei et al. (2020) gives a convergence rate of soft policy gradient methods, also with a dependence on the rate at which $\min_{s,a} \pi_{\theta_t}(a|s)$ diminishes. They too leave the investigation of the rate as an open problem. Theorem 8 of Mei et al. (2020), however, only considers a bandit setting with the exact soft policy gradient and leaves the general MDP setting for future work. By contrast, Corollary 8 applies to general MDPs with approximate and stochastic update steps.

6. Related Work

Our Theorem 2 regarding the finite sample analysis of stochastic approximation algorithms follows the line of research of Chen et al. (2020, 2021). In particular, Chen et al. (2020) consider (9) with an expected operator (i.e., $F_{\theta_t}(w_t, Y_t)$ is replaced by $\bar{F}(w_t)$). Chen et al. (2021) extend Chen et al. (2020) in that the expected operator is replaced by the stochastic operator $F(w_t, Y_t)$, though $\{Y_t\}$ here is a Markov chain with fixed dynamics. We further extend Chen et al. (2021) from time-homogeneous stochastic operator and dynamics to time-inhomogeneous stochastic operator and dynamics. This line of research depends on properties of contraction mappings. There are also ODE-based analysis for stochastic approximation algorithms (Benveniste et al., 1990; Kushner and Yin, 2003; Borkar, 2009) and we refer the reader to Chen et al. (2020, 2021) for a more detailed review.

In this work, we focus on the optimality of naive actor critic algorithms that do not use second order information. With the help of the Fisher information, the optimality of natural actor critic (Kakade, 2001; Peters and Schaal, 2008; Bhatnagar et al., 2009) is also established in both on-policy settings (Agarwal et al., 2020; Wang et al., 2019; Liu et al., 2020; Khodadadian et al., 2022) and off-policy settings (Khodadadian et al., 2021; Chen et al., 2022). In particular, Agarwal et al. (2020); Khodadadian et al. (2022, 2021) establish the optimality of natural actor critic in the tabular setting. They, however, make synchronous updates to the actor. In other words, they update the policy for all states at each time step. Consequently, the state distribution is not important there. By contrast, the naive actor critic this work considers makes asynchronous updates to the actor. In other words, at each time step, we only update the policy for the current state. This asynchronous update is more practical in large scale experiments. Moreover, Xu et al. (2021) establish the convergence to stationary points of an off-policy actor critic with density ratio correction and a fixed sampling distribution. To study the optimality of the stationary points, Xu et al. (2021) also make some assumptions about the Fisher information. In this work, we do not use any second order information. How this work achieves optimality (i.e., vanilla actor critic with decaying KL regularization) is fundamentally different from natural actor critic.

Liu et al. (2020) improve the results of Agarwal et al. (2020) regarding the optimality of policy gradient methods from exact gradient to stochastic and approximate gradient. Liu et al. (2020), however, work on on-policy settings and require nested loops. By contrast, we work on off-policy settings and consider three-timescale updates.

Degris et al. (2012) also study the convergence of an off-policy actor critic without using density ratio to correct the state distribution mismatch. As noted in the Errata of Degris et al. (2012), their results also exclusively apply to tabular settings. Additionally, Degris et al. (2012) establish asymptotic convergence to only some locally asymptotically stable points of an ODE without any convergence rate. And the optimality of those locally asymptotically stable points remains unclear. Further, Degris et al. (2012) assume the transitions are identically and independently sampled. By contrast, our transitions are obtained by following a time-inhomogeneous behavior policy.

In this paper, we focus on the tabular setting as a starting point for this line of research. When linear function approximation is used for the critic, compatible features (Sutton et al., 1999; Konda, 2002; Zhang et al., 2020c) can be used to eliminate the bias resulting from the limit of the representation capacity. With the help of compatible features, Liu et al. (2020) show the optimality of their on-policy actor critic and Xu et al. (2021) show the optimality of their off-policy actor critic. We leave the study of linear function approximation in our settings with compatible features for future work.

7. Experiments

In this section, we provide some empirical results in complement to our theoretical analysis. The implementation is made publicly available to facilitate future research.[2] In particular, we are interested in the following three questions:

(i). Can the claimed convergence and optimality of Algorithm 1 in Theorem 5 be observed in computational experiments?

(ii). Can the claimed convergence of Algorithm 2 in Theorem 7 and its conjectured optimality from (21) and Corollary 8 be observed in computational experiments?

(iii). How is the KL-based regularization (c.f. Algorithm 1) qualitatively different from the entropy-based regularization (c.f. Algorithm 2)?

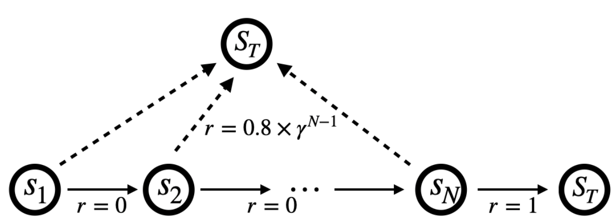

Figure 1: The chain domain from Laroche and Tachet (2021) with $\gamma = 0.99$.

We use the chain domain from Laroche and Tachet (2021) as our testbed. As described in Figure 1, there are N non-terminal states in the chain and the agent is always initialized

[2] `https://github.com/ShangtongZhang/DeepRL`

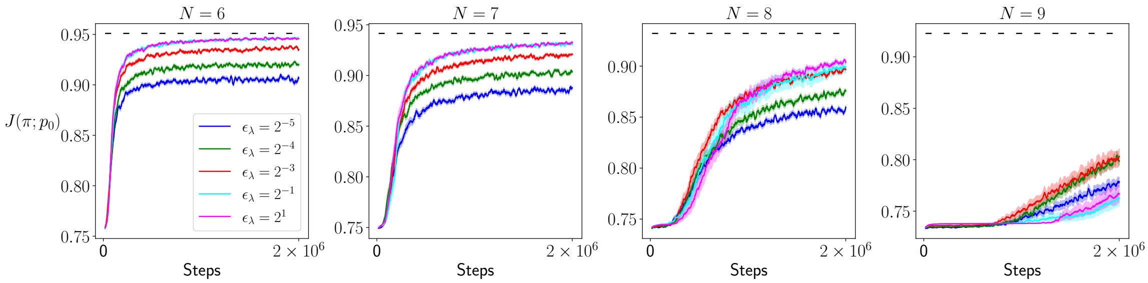

Figure 2: Evaluation performance against training steps of Algorithm 1

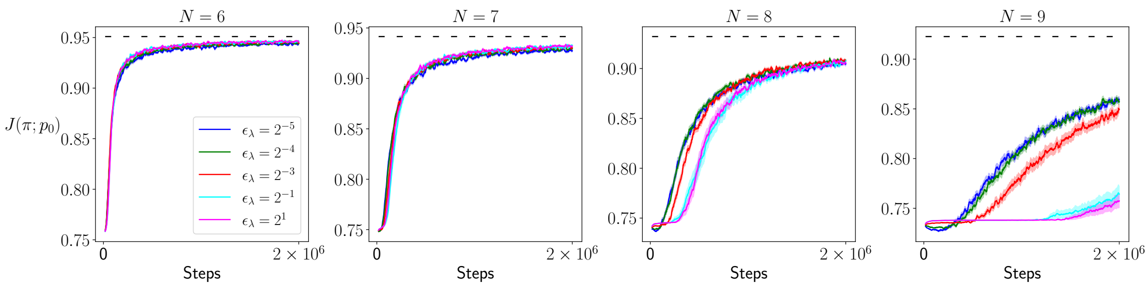

Figure 3: Evaluation performance against training steps of Algorithm 2

at state s_1. There are two actions available in each state. The `solid` action leads the agent from s_i to s_{i+1} and yields a reward of 0 for all $i < N$. At s_N, the `solid` action instead leads to the terminal state and yields a reward of 1. The `dotted` action always leads to the terminal state directly and yields a reward $0.8 \times \gamma^{N-1}$. Trivially, the optimal policy is to always choose the `solid` action, which will yield an episodic return of γ^{N-1}. As noted by Laroche and Tachet (2021), the challenge of this chain domain is to overcome the immediate rewards pushing the agent towards suboptimal policies. We remark that though this chain has a finite horizon, we can indeed reformalize it into an infinite-horizon chain with transition-dependent discounting. We refer the reader to White (2017) for more details about this technique and we believe our theoretical results can be easily extended to transition-dependent discounting.

We run Algorithms 1 and 2 in the chain domain. According to (13), we use the behavior policy

$$\mu_\theta(\cdot|s) \doteq 0.1 \times \frac{1}{2} + 0.9 \times \frac{\exp(0.1 \times \theta_{s,\cdot})}{\exp(0.1 \times \theta_{s,\texttt{solid}}) + \exp(0.1 \times \theta_{s,\texttt{dotted}})}.$$

According to (17) and Assumption 4.1, we set $\{\alpha_t, \beta_t, \lambda_t\}$ as

$$\alpha_t \doteq \frac{100+1}{(t+10^5)^{0.5+0.001}},$$
$$\beta_t \doteq \frac{100}{(t+10^5)^{0.75+0.001}},$$
$$\lambda_t \doteq \frac{0.025}{(t+10^5)^{\epsilon_\lambda}},$$

where we test a range of ϵ_λ from $\{2^{-5}, 2^{-4}, 2^{-3}, 2^{-1}, 2\}$[3]. We run both Algorithms 1 and 2 for 2×10^6 steps and evaluate the target policy every 2×10^3 steps, where we execute it for 10 episodes and take the mean episodic return. The evaluation performance is reported in Figures 2 and 3 respectively. Curves are averaged over 30 independent runs with shaded regions indicating standard errors. The black dotted lines are the performance of the optimal policy.

As suggested by Figure 2 with $N \in \{6, 7\}$, when $\epsilon_\lambda \in \{2^{-1}, 2\}$, the target policy found by Algorithm 1 is indeed very close to the optimal policy at the end of training, which gives an affirmative answer to the question (i). It is important to note that neither $\epsilon_\lambda = 2^{-1}$ nor $\epsilon_\lambda = 2$ is recommended by (18). This is expected as Assumption 4.1 is only sufficient and the convergence rate in Theorem 5 can possibly be significantly improved. Further, with the increase of N, the suboptimality of the target policy at the end of training also increases. This is expected as increasing N makes the problem more challenging. We, however, remark that though with $N \in \{8, 9\}$, the target policy is not close to the optimal policy at the end of training, all curves are monotonically improving as time progresses. Similarly, the results in Figure 3 give an affirmative answer to the question (ii). Comparing Figures 2 and 3, it is easy to see that Algorithm 1 is much more sensitive to ϵ_λ than Algorithm 2. As shown by Figure 2, the selection of ϵ_λ significantly affects the rate that the suboptimality diminishes in Algorithm 1. By contrast, Figure 3 suggests that the rate that the suboptimality diminishes is barely affected by ϵ_λ in Algorithm 2. This comparison gives an intuitive answer the question (iii). This difference is because the KL regularization is much more aggressive than the entropy regularization. To be more specific, the entropy of the policy is always bounded but the KL divergence used here can be unbounded when the policy becomes deterministic.

8. Conclusion

In this paper, we demonstrate the optimality of the off-policy actor critic algorithm even without using a density ratio to correct the state distribution mismatch. This result is significant in two aspects. First, it advances the understanding of the optimality of policy gradient methods in the tabular setting from Agarwal et al. (2020); Mei et al. (2020); Laroche and Tachet (2021). Second, it provides, to certain extent, a theoretical justification for the practice of ignoring state distribution mismatch in large scale RL experiments (Wang et al., 2017; Espeholt et al., 2018; Vinyals et al., 2019; Schmitt et al., 2020; Zahavy et al., 2020). One important ingredient of our results is the finite sample analysis of a generic stochastic approximation algorithm with time-inhomogeneous update operators on time-inhomogeneous Markov chains, which we believe can be used to analyze more RL algorithms and has interest beyond RL.

Acknowledgments

[3] We omit $\epsilon_\lambda = 2^{-2}$ and $\epsilon_\lambda = 1$ to improve the readability of the figures. The corresponding curves are similar to $\epsilon_\lambda = 2^{-1}$ and $\epsilon_\lambda = 2$.

Part of this work was done during SZ's internship at Microsoft Research Montreal and SZ's DPhil at the University of Oxford. SZ is also funded by the Engineering and Physical Sciences Research Council (EPSRC) during his DPhil.

Appendix A. Proofs of Section 3

A.1 Proof of Lemma 1

Lemma 9 *(Uniform ergodicity implies uniform mixing) Let Assumption 3.2 hold. Then, there exist constants $C_0 > 0$ and $\tau \in (0, 1)$, independent of θ, such that for any $n > 0$,*

$$\sup_{y,\theta} \sum_{y'} \left| P_\theta^n(y, y') - d_\theta(y') \right| \leq C_0 \tau^n. \tag{10}$$

Proof Theorem 4.9 of Levin and Peres (2017) confirms the geometric mixing for a single ergodic chain. Here we adapt its proof to show the uniform mixing.

For any $P \in \bar{\Lambda}_P$, define the indicator matrix $\mathbb{I}_P \in \{0, 1\}^{|\mathcal{Y}| \times |\mathcal{Y}|}$ such that

$$\mathbb{I}_P[y, y'] = \begin{cases} 1, & P[y, y'] > 0 \\ 0, & P[y, y'] = 0 \end{cases}.$$

Consider the stochastic matrix M_P defined as

$$M_P(i, j) = \frac{\mathbb{I}_P(i, j)}{\sum_{j'} \mathbb{I}_P(i, j')}.$$

Since the chain induced by P is ergodic, it is easy to see the chain induced by M_P is also ergodic. This is because (1) a finite chain is ergodic if and only if it is irreducible and aperiodic; (2) the connectivity of the chain induced by M_P is the same as that by P; and (3) irreducibility and aperiodicity depend only on connectivity, not on the specific probability of each transition.

The proof of Proposition 1.7 of Levin and Peres (2017) then asserts that there exists a constant $t_P > 0$ such that for all $t \geq t_P$,

$$M_P^t(i, j) > 0$$

holds for any i, j. Hence $P^t(i, j) > 0$ also holds. Since $\mathcal{Y}$ is finite, the set $\left\{\mathbb{I}_P | P \in \bar{\Lambda}_P\right\}$ is also finite (at most $2^{|\mathcal{Y}| \times |\mathcal{Y}|}$ elements), and so are the sets $\left\{M_P | P \in \bar{\Lambda}_P\right\}$ and $\left\{t_P | P \in \bar{\Lambda}_P\right\}$. Let

$$t_* \doteq \max_{P \in \bar{\Lambda}_P} \{t_P\},$$

we then have for any $P \in \bar{\Lambda}_P$, $P^{t_*}(i, j) > 0$ always holds. Importantly, t_* is independent of P. Then the extreme value theorem implies that

$$\delta \doteq \inf_{P \in \bar{\Lambda}_P, i, j} P^{t_*}(i, j) > 0.$$

Let d_P be the invariant distribution of the chain induced by P and take any δ' such that $0 < \delta' < \delta$, then

$$P^{t_*}(i,j) > \delta' \geq \delta' d_P(j)$$

holds for any $P \in \bar{\Lambda}_P, i, j$.

For any $P \in \bar{\Lambda}_P$, let Π be a matrix, each row of which is $d_P^\top$, and define

$$\zeta \doteq 1 - \delta'.$$

We now verify that the matrix

$$Q \doteq \frac{P^{t_*} + \zeta\Pi - \Pi}{\zeta}$$

is a stochastic matrix. First, its row sums are 1:

$$(Q\mathbf{1})(i) \doteq \frac{1 + \zeta - 1}{\zeta} = 1.$$

Second, its elements are nonnegative:

$$Q(i,j) = \frac{P^{t_*}(i,j) + \zeta d_P(j) - d_P(j)}{\zeta} \geq \frac{\delta' d_P(j) + \zeta d_P(j) - d_P(j)}{\zeta} = 0.$$

Rearranging terms yields

$$P^{t_*} = (1 - \zeta)\Pi + \zeta Q. \tag{22}$$

We now use induction to show that for any $k \geq 1$,

$$P^{t_* k} = (1 - \zeta^k)\Pi + \zeta^k Q^k. \tag{23}$$

For $k = 1$, we know (23) holds from (22). Suppose (23) holds for $k = n$, then

$$\begin{aligned}
P^{t_*(n+1)} &= P^{t_* n} P^{t_*} \\
&= \left((1 - \zeta^n)\Pi + \zeta^n Q^n\right) P^{t_*} \\
&= (1 - \zeta^n)\Pi P^{t_*} + \zeta^n Q^n \left((1 - \zeta)\Pi + \zeta Q\right) \\
&= (1 - \zeta^n)\Pi P^{t_*} + (1 - \zeta)\zeta^n Q^n \Pi + \zeta^{n+1} Q^{n+1} \\
&= (1 - \zeta^n)\Pi + (1 - \zeta)\zeta^n Q^n \Pi + \zeta^{n+1} Q^{n+1} \\
&\qquad \text{(Property of invariant distribution)} \\
&= (1 - \zeta^n)\Pi + (1 - \zeta)\zeta^n \Pi + \zeta^{n+1} Q^{n+1} \\
&\qquad (Q\Pi = Q1d_P^\top = 1d_P^\top = \Pi \text{ for any stochastic matrix } Q) \\
&= (1 - \zeta^{n+1})\Pi + \zeta^{n+1} Q^{n+1},
\end{aligned}$$

which completes the induction. Consequently, for any $l \in \{0, 1, \dots, t_* - 1\}$, multiplying by P^l both sides of (23) yields

$$\begin{aligned} P^{t_* k + l} &= (1 - \zeta^k)\Pi P^l + \zeta^k Q^k P^l \\ &= (1 - \zeta^k)\Pi + \zeta^k Q^k P^l. \end{aligned}$$

Rearranging terms yields

$$P^{t_* k + l} - \Pi = \zeta^k (Q^k P^l - \Pi),$$

implying for any i,

$$\begin{aligned} \sum_j \left| P^{t_* k + l}(i, j) - d_P(j) \right| &= \zeta^k \sum_j \left| (Q^k P^l)(i, j) - d_P(j) \right| \\ &\leq 2\zeta^k \quad \text{(Boundedness of total variation)} \\ &= 2\zeta^{-\frac{l}{t_*}} \left(\zeta^{\frac{1}{t_*}} \right)^{t_* k + l} \\ &\leq 2\zeta^{-\frac{t_* - 1}{t_*}} \left(\zeta^{\frac{1}{t_*}} \right)^{t_* k + l}. \end{aligned}$$

Let

$$C_0' \doteq 2\zeta^{-\frac{t_* - 1}{t_*}}, \tau \doteq \zeta^{\frac{1}{t_*}}.$$

It is easy to see $C_0' > 0, \tau \in (0, 1)$ and both C_0 and τ are independent of P. Consequently, for any $n \geq t_*$, we have

$$\sum_j |P^n(i, j) - d_P(j)| \leq C_0' \tau^n.$$

By the boundedness of total variation, for $n \in \{0, 1, \dots, t_* - 1\}$, we have

$$\sum_j |P^n(i, j) - d_P(j)| \leq 2 \leq \frac{2}{\tau^{t_*}} \tau^n.$$

Setting $C_0 \doteq \max \left\{ C_0', \frac{2}{\tau^{t_*}} \right\}$ completes the proof. ∎

A.2 Proof of Theorem 2

Theorem 2 *Let Assumptions 3.1 - 3.6 hold. For any*

$$\epsilon_w \in (0, \min\{2(\epsilon_\beta - \epsilon_\alpha), \epsilon_\alpha\}),$$

if t_0 is sufficiently large, then $\forall t$,

$$\mathbb{E}\left[\left\| w_t - w^*_{\theta_t} \right\|_c^2 \right] = \mathcal{O}\left(\frac{1}{(t + t_0)^{\epsilon_w}} \right).$$

Proof Since the theorem is a generalization of the results in Chen et al. (2021), we follow their framework to complete the proof. In our setting, the dynamics of the Markov chain changes every time step according to a secondary random sequence $\{\theta_t\}$. Consequently, we have many new error terms which are not controlled by Chen et al. (2021) and that we handle using techniques from Zou et al. (2019).

Following Chen et al. (2021), we use a Lyapunov method for the proof with the generalized Moreau envelope of $\frac{1}{2}\|\cdot\|_c^2$ as the Lyapunov function. In particular, we consider the Lyapunov function

$$M(w) \doteq \inf_{u \in \mathbb{R}^K} \left\{ \frac{1}{2}\|u\|_c^2 + \frac{1}{2\xi}\|w - u\|_s^2 \right\},$$

where $\xi > 0$ is a constant to be tuned, $\|\cdot\|_c$ is the norm w.r.t. which $\bar{F}_\theta$ is contractive (c.f. Assumption 3.3), and $\|\cdot\|_s$ is an arbitrary norm such that $\frac{1}{2}\|\cdot\|_s^2$ is L-smooth (Lemma 45). It can, e.g., be an ℓ_p norm with $p \geq 2$ (Example 5.11 of Beck (2017)). Due to the equivalence between norms, there exist positive constants l_{cs} and u_{cs} such that

$$l_{cs}\|w\|_s \leq \|w\|_c \leq u_{cs}\|w\|_s$$

holds for any w. The following lemma proved by Chen et al. (2021) describes some properties of M.

Lemma 10 *(Proposition A.1 of Chen et al. (2021))*

(i). *$M(w)$ is convex, and $\frac{L}{\xi}$-smooth w.r.t. $\|\cdot\|_s$.*

(ii). *There exists a norm $\|\cdot\|_m$ such that $M(w) = \frac{1}{2}\|w\|_m^2$.*

(iii). *Define*

$$l_{cm} = \sqrt{(1 + \xi l_{cs}^2)}$$
$$u_{cm} = \sqrt{(1 + \xi u_{cs}^2)},$$

then $\forall w$,

$$l_{cm}\|w\|_m \leq \|w\|_c \leq u_{cm}\|w\|_m.$$

Lemma 10 (i) and Lemma 45 imply that for any x, x',

$$M(x') \leq M(x) + \left\langle \nabla M(x), x' - x \right\rangle + \frac{L}{2\xi}\left\|x - x'\right\|_s^2.$$

Using $x' = w_{t+1} - w^*_{\theta_{t+1}}$ and $x = w_t - w^*_{\theta_t}$ in the above inequality and the update equation (9):

$$w_{t+1} \doteq w_t + \alpha_t(F_{\theta_t}(w_t, Y_t) - w_t + \epsilon_t),$$

Lemma 10 (ii) yields

$$
\begin{aligned}
&\frac{1}{2}\left\|w_{t+1}-w^*_{\theta_{t+1}}\right\|_m^2 \\
\leq&\frac{1}{2}\left\|w_t-w^*_{\theta_t}\right\|_m^2+\left\langle \nabla M(w_t-w^*_{\theta_t}), w_{t+1}-w_t+w^*_{\theta_t}-w^*_{\theta_{t+1}}\right\rangle \\
&+\frac{L}{2\xi}\left\|w_{t+1}-w_t+w^*_{\theta_t}-w^*_{\theta_{t+1}}\right\|_s^2 \\
=&\frac{1}{2}\left\|w_t-w^*_{\theta_t}\right\|_m^2 \\
&+\underbrace{\left\langle \nabla M(w_t-w^*_{\theta_t}), w^*_{\theta_t}-w^*_{\theta_{t+1}}\right\rangle}_{T_1} \\
&+\alpha_t\underbrace{\left\langle \nabla M(w_t-w^*_{\theta_t}), \bar{F}_{\theta_t}(w_t)-w_t\right\rangle}_{T_2} \\
&+\alpha_t\underbrace{\left\langle \nabla M(w_t-w^*_{\theta_t}), F_{\theta_t}(w_t,Y_t)-\bar{F}_{\theta_t}(w_t)\right\rangle}_{T_3} \\
&+\alpha_t\underbrace{\left\langle \nabla M(w_t-w^*_{\theta_t}), \epsilon_t\right\rangle}_{T_4} \\
&+\alpha_t^2\underbrace{\frac{L}{\xi}\left\|F_{\theta_t}(w_t,Y_t)-w_t+\epsilon_t\right\|_s^2}_{T_5} \\
&+\underbrace{\frac{L}{\xi}\left\|w^*_{\theta_t}-w^*_{\theta_{t+1}}\right\|_s^2}_{T_6}.
\end{aligned} \tag{24}
$$

We now bound T_1 - T_6 one by one. T_1 and T_6 are errors resulting from changing dynamics and are not controlled in Chen et al. (2021). T_2, T_4, and T_5 can be bounded similarly to Chen et al. (2021). To bound T_3, we further decompose it as

$$
\begin{aligned}
T_3 =&\left\langle \nabla M(w_t-w^*_{\theta_t}), F_{\theta_t}(w_t,Y_t)-\bar{F}_{\theta_t}(w_t)\right\rangle \\
=&\underbrace{\left\langle \nabla M(w_t-w^*_{\theta_t})-\nabla M(w_{t-\tau_{\alpha_t}}-w^*_{\theta_{t-\tau_{\alpha_t}}}), F_{\theta_t}(w_t,Y_t)-\bar{F}_{\theta_t}(w_t)\right\rangle}_{T_{31}} \\
&+\underbrace{\left\langle \nabla M(w_{t-\tau_{\alpha_t}}-w^*_{\theta_{t-\tau_{\alpha_t}}}), F_{\theta_t}(w_t,Y_t)-F_{\theta_t}(w_{t-\tau_{\alpha_t}},Y_t)+\bar{F}_{\theta_t}(w_{t-\tau_{\alpha_t}})-\bar{F}_{\theta_t}(w_t)\right\rangle}_{T_{32}} \\
&+\underbrace{\left\langle \nabla M(w_{t-\tau_{\alpha_t}}-w^*_{\theta_{t-\tau_{\alpha_t}}}), F_{\theta_t}(w_{t-\tau_{\alpha_t}},Y_t)-\bar{F}_{\theta_t}(w_{t-\tau_{\alpha_t}})\right\rangle}_{T_{33}},
\end{aligned}
$$

where

$$
\tau_{\alpha_t} \doteq \min\left\{n\geq 0 \mid C_0\tau^n \leq \alpha_t\right\}, \tag{25}
$$

and C_0 and τ are defined in Lemma 1. τ_{α_t} denotes the number of steps the chain needs to mix to an accuracy of α_t. T_{31} and T_{32} can be bounded similarly to Chen et al. (2021). The bound for T_{33} is however significantly different. We decompose T_{33} as

$$\begin{aligned}
T_{33} =& \left\langle \nabla M(w_{t-\tau_{\alpha_t}} - w^*_{\theta_{t-\tau_{\alpha_t}}}), F_{\theta_t}(w_{t-\tau_{\alpha_t}}, Y_t) - \bar{F}_{\theta_t}(w_{t-\tau_{\alpha_t}}) \right\rangle \\
=& \underbrace{\left\langle \nabla M(w_{t-\tau_{\alpha_t}} - w^*_{\theta_{t-\tau_{\alpha_t}}}), F_{\theta_{t-\tau_{\alpha_t}}}(w_{t-\tau_{\alpha_t}}, \tilde{Y}_t) - \bar{F}_{\theta_{t-\tau_{\alpha_t}}}(w_{t-\tau_{\alpha_t}}) \right\rangle}_{T_{331}} + \\
& \underbrace{\left\langle \nabla M(w_{t-\tau_{\alpha_t}} - w^*_{\theta_{t-\tau_{\alpha_t}}}), F_{\theta_{t-\tau_{\alpha_t}}}(w_{t-\tau_{\alpha_t}}, Y_t) - F_{\theta_{t-\tau_{\alpha_t}}}(w_{t-\tau_{\alpha_t}}, \tilde{Y}_t) \right\rangle}_{T_{332}} + \\
& \underbrace{\left\langle \nabla M(w_{t-\tau_{\alpha_t}} - w^*_{\theta_{t-\tau_{\alpha_t}}}), F_{\theta_t}(w_{t-\tau_{\alpha_t}}, Y_t) - F_{\theta_{t-\tau_{\alpha_t}}}(w_{t-\tau_{\alpha_t}}, Y_t) \right\rangle}_{T_{333}} + \\
& \underbrace{\left\langle \nabla M(w_{t-\tau_{\alpha_t}} - w^*_{\theta_{t-\tau_{\alpha_t}}}), \bar{F}_{\theta_{t-\tau_{\alpha_t}}}(w_{t-\tau_{\alpha_t}}) - \bar{F}_{\theta_t}(w_{t-\tau_{\alpha_t}}) \right\rangle}_{T_{334}}.
\end{aligned}$$

Here $\left\{\tilde{Y}_t\right\}$ is an auxiliary chain inspired from Zou et al. (2019). Before time $t - \tau_{\alpha_t} - 1$, $\left\{\tilde{Y}_t\right\}$ is exactly the same as $\{Y_t\}$. After time $t - \tau_{\alpha_t} - 1$, $\tilde{Y}_t$ evolves according to the *fixed* kernel $P_{\theta_{t-\tau_{\alpha_t}}}$ while Y_t evolves according the changing kernel $P_{\theta_{t-\tau_{\alpha_t}}}, P_{\theta_{k-\tau_{\alpha_t}+1}}, \ldots$.

$$\begin{aligned}
\left\{\tilde{Y}_t\right\} &: \cdots \to Y_{t-\tau_{\alpha_t}-1} \underbrace{\to}_{P_{\theta_{t-\tau_{\alpha_t}}}} Y_{t-\tau_{\alpha_t}} \underbrace{\to}_{P_{\theta_{t-\tau_{\alpha_t}}}} \tilde{Y}_{t-\tau_{\alpha_t}+1} \underbrace{\to}_{P_{\theta_{t-\tau_{\alpha_t}}}} \tilde{Y}_{t-\tau_{\alpha_t}+2} \to \ldots \\
\{Y_t\} &: \cdots \to Y_{t-\tau_{\alpha_t}-1} \underbrace{\to}_{P_{\theta_{t-\tau_{\alpha_t}}}} Y_{t-\tau_{\alpha_t}} \underbrace{\to}_{P_{\theta_{t-\tau_{\alpha_t}+1}}} Y_{t-\tau_{\alpha_t}+1} \underbrace{\to}_{P_{\theta_{t-\tau_{\alpha_t}+2}}} Y_{t-\tau_{\alpha_t}+2} \to \ldots.
\end{aligned}$$

We are now ready to present bounds for each of the above terms. To begin, we define some shorthand and study their properties:

$$\begin{aligned}
\alpha_{t_1,t_2} \doteq \sum_{t=t_1}^{t_2} \alpha_t, \quad \beta_{t_1,t_2} \doteq \sum_{t=t_1}^{t_2} \beta_t \\
A \doteq U_\epsilon + L_F + 1, \quad B \doteq U_F + U'_\epsilon, \quad C \doteq AU_w + B + A + A(1 + U'_F + U''_F).
\end{aligned} \tag{26}$$

Lemma 11 *For sufficiently large t_0,*

$$\begin{aligned}
\tau_{\alpha_t} = \mathcal{O}(\log(t+t_0)), \quad \alpha_{t-\tau_{\alpha_t},t-1} = \mathcal{O}\left(\frac{\log(t+t_0)}{(t+t_0)^{\epsilon_\alpha}}\right), \\
\beta_{t-\tau_{\alpha_t},t-1} = \mathcal{O}\left(\frac{\log(t+t_0)}{(t+t_0)^{\epsilon_\alpha}}\right), \quad \frac{\alpha_t \alpha_{t-\tau_{\alpha_t},t-1}}{\beta_t} = \mathcal{O}\left(\frac{\log(t+t_0)}{(t+t_0)^{2\epsilon_\alpha - \epsilon_\beta}}\right).
\end{aligned}$$

The proof of Lemma 11 is provided in Section E.11. Lemma 11 asserts that we can select a t_0 sufficiently large such that

$$\alpha_{t-\tau_{\alpha_t},t-1} \leq \frac{1}{4A}$$

holds for all t. This condition is crucial for Lemma 47, which plays an important role in the following bounds.

Lemma 12 *(Bound of T_1)*

$$T_1 \leq \frac{L_w L_\theta \beta_t}{l_{cm}} \left\| w_t - w^*_{\theta_t} \right\|_m.$$

The proof of Lemma 12 is provided in Section E.1.

Lemma 13 *(Bound of T_2)*

$$T_2 \leq -(1 - \kappa \frac{u_{cm}}{l_{cm}}) \left\| w_t - w^*_{\theta_t} \right\|_m^2.$$

The proof of Lemma 13 is provided in Section E.2.

Lemma 14 *(Bound of T_{31})*

$$T_{31} \leq \frac{8L(L_w L_\theta + 1)\alpha_{t-\tau_{\alpha_t},t-1}}{\xi l_{cs}^2} \left(u_{cm}^2 A^2 \left\| w_t - w^*_{\theta_t} \right\|_m^2 + C^2 \right).$$

The proof of Lemma 14 is provided in Section E.3.

Lemma 15 *(Bound of T_{32})*

$$T_{32} \leq \frac{32L\alpha_{t-\tau_{\alpha_t},t-1}(1 + L_w L_\theta \beta_{t-\tau_{\alpha_t},t-1})}{\xi l_{cs}^2} \left(u_{cm}^2 A^2 \left\| w_t - w^*_{\theta_t} \right\|_m^2 + C^2 \right).$$

The proof of Lemma 15 is provided in Section E.4.

Lemma 16 *(Bound of T_{331})*

$$\mathbb{E}\left[T_{331}\right] \leq \frac{8L\alpha_t(1 + L_w L_\theta \beta_{t-\tau_{\alpha_t},t-1})}{A\xi l_{cs}^2} \left(u_{cm}^2 A^2 \mathbb{E}\left[\left\| w_t - w^*_{\theta_t} \right\|_m^2 \right] + C^2 \right).$$

The proof of Lemma 16 is provided in Section E.5.

Lemma 17 *(Bound of T_{332})*

$$\mathbb{E}\left[T_{332}\right] \leq \frac{8|\mathcal{Y}|L_P L_\theta \sum_{j=t-\tau_{\alpha_t}}^{t-1} \beta_{t-\tau_{\alpha_t},j} L(1 + L_w L_\theta \beta_{t-\tau_{\alpha_t},t-1})}{A\xi l_{cs}^2} \left(u_{cm}^2 A^2 \mathbb{E}\left[\left\| w_t - w^*_{\theta_t} \right\|_m^2 \right] + C^2 \right).$$

The proof of Lemma 17 is provided in Section E.6.

Lemma 18 *(Bound of T_{333})*

$$T_{333} \leq \frac{8LL'_F L_\theta \beta_{t-\tau_{\alpha_t},t-1}(1 + L_w L_\theta \beta_{t-\tau_{\alpha_t},t-1})}{A^2 \xi l_{cs}^2} \left(u_{cm}^2 A^2 \left\| w_t - w_{\theta_t}^* \right\|_m^2 + C^2 \right).$$

The proof of Lemma 18 is provided in Section E.7.

Lemma 19 *(Bound of T_{334})*

$$T_{334} \leq \frac{8LL''_F L_\theta \beta_{t-\tau_{\alpha_t},t-1}(1 + L_w L_\theta \beta_{t-\tau_{\alpha_t},t-1})}{A^2 \xi l_{cs}^2} \left(u_{cm}^2 A^2 \left\| w_t - w_{\theta_t}^* \right\|_m^2 + C^2 \right).$$

The proof of Lemma 19 is provided in Section E.8.

Lemma 20 *(Bound of T_4)*

$$\mathbb{E}\left[T_4\right] = 0.$$

The proof of Lemma 20 is provided in Section E.9.

Lemma 21 *(Bound of T_5)*

$$T_5 \leq \frac{2L}{\xi l_{cs}^2} \left(A^2 u_{cm}^2 \left\| w_t - w_{\theta_t}^* \right\|_m^2 + C^2 \right).$$

The proof of Lemma 21 is provided in Section E.10.

Lemma 22 *(Bound of T_6)*

$$T_6 = \frac{L}{\xi} \left\| w_{\theta_t}^* - w_{\theta_{t+1}}^* \right\|_s^2 \leq \frac{LL_w^2 L_\theta^2 \beta_t^2}{\xi l_{cs}^2}.$$

Lemma 22 follows immediately from Assumptions 3.4 and 3.6.

We now assemble the bounds in Lemmas 12 - 22 back into (24). By the definition of u_{cm} and l_{cm} in Lemma 10, we have

$$\lim_{\xi \to 0} \frac{u_{cm}}{l_{cm}} = 1.$$

Since $\kappa < 1$, we can select a sufficiently small $\xi > 0$ such that

$$\psi_1 \doteq \frac{2}{9}(1 - \kappa \frac{u_{cm}}{l_{cm}})$$

satisfies $\psi_1 \in (0, 1)$, implying

$$\begin{aligned} T_2 &\leq -\frac{9}{2}\psi_1 \left\| w_t - w_{\theta_t}^* \right\|_m^2, \\ \alpha_t T_2 &\leq -\frac{9}{2}\alpha_t \psi_1 \left\| w_t - w_{\theta_t}^* \right\|_m^2. \end{aligned} \tag{27}$$

Let ψ_2 be a positive constant to be tuned. For T_1, suppose ψ_2 is large enough, then we have

$$T_1 \leq \frac{1}{2}\beta_t\psi_2\left\|w_t - w^*_{\theta_t}\right\|_m. \tag{28}$$

For T_{31}, Lemmas 11 and 14 assert that we can select sufficiently large t_0 and ψ_2 such that

$$\begin{aligned} T_{31} &\leq \frac{1}{2}\psi_1\left\|w_t - w^*_{\theta_t}\right\|_m^2 + \frac{1}{2}\alpha_{t-\tau_{\alpha_t},t-1}\psi_2, \\ \alpha_t T_{31} &\leq \frac{1}{2}\alpha_t\psi_1\left\|w_t - w^*_{\theta_t}\right\|_m^2 + \frac{1}{2}\alpha_t\alpha_{t-\tau_{\alpha_t},t-1}\psi_2. \end{aligned} \tag{29}$$

For T_{32}, Lemma 11 implies that for t_0 large enough

$$\beta_{t-\tau_{\alpha_t},t-1} \leq 1.$$

Hence, Lemma 15 guarantees that we can select sufficiently large t_0 and ψ_2 such that

$$\begin{aligned} T_{32} &\leq \frac{1}{2}\psi_1\left\|w_t - w^*_{\theta_t}\right\|_m^2 + \frac{1}{2}\alpha_{t-\tau_{\alpha_t},t-1}\psi_2, \\ \alpha_t T_{32} &\leq \frac{1}{2}\alpha_t\psi_1\left\|w_t - w^*_{\theta_t}\right\|_m^2 + \frac{1}{2}\alpha_t\alpha_{t-\tau_{\alpha_t},t-1}\psi_2. \end{aligned} \tag{30}$$

For T_{331}, similarly, we can select sufficiently large t_0 and ψ_2 such that

$$\begin{aligned} \mathbb{E}\left[T_{331}\right] &\leq \frac{1}{2}\psi_1\mathbb{E}\left[\left\|w_t - w^*_{\theta_t}\right\|_m^2\right] + \frac{1}{2}\alpha_t\psi_2 \\ &\leq \frac{1}{2}\psi_1\mathbb{E}\left[\left\|w_t - w^*_{\theta_t}\right\|_m^2\right] + \frac{1}{2}\alpha_{t-\tau_{\alpha_t},t-1}\psi_2, \\ \alpha_t\mathbb{E}\left[T_{331}\right] &\leq \frac{1}{2}\alpha_t\psi_1\mathbb{E}\left[\left\|w_t - w^*_{\theta_t}\right\|_m^2\right] + \frac{1}{2}\alpha_t\alpha_{t-\tau_{\alpha_t},t-1}\psi_2. \end{aligned} \tag{31}$$

For T_{332}, we have

$$\begin{aligned} \frac{\sum_{j=t-\tau_{\alpha_t}}^{t-1}\beta_{t-\tau_{\alpha_t},j}}{\alpha_{t-\tau_{\alpha_t},t-1}} &\leq \frac{\tau_{\alpha_t}\tau_{\alpha_t}\beta_{t-\tau_{\alpha_t}}}{\tau_{\alpha_t}\alpha_t} = \frac{\tau_{\alpha_t}\beta_{t-\tau_{\alpha_t}}}{\alpha_t} = \mathcal{O}\left(\frac{\log(t+t_0)\beta_{t-\tau_{\alpha_t}}}{\alpha_t}\right) \\ &= \mathcal{O}\left(\frac{\log(t+t_0)\beta_t}{\alpha_t}\right) \quad \text{(for } t_0 \text{ sufficiently large).} \end{aligned}$$

Since the RHS of the above inequality approaches 0 when t_0 is sufficiently large, we can select sufficiently large t_0 such that

$$\sum_{j=t-\tau_{\alpha_t}}^{t-1}\beta_{t-\tau_{\alpha_t},j} \leq \alpha_{t-\tau_{\alpha_t},t-1}.$$

Then it is easy to see for sufficiently large t_0 and ψ_2,

$$\begin{aligned} \mathbb{E}\left[T_{332}\right] &\leq \frac{1}{2}\psi_1\mathbb{E}\left[\left\|w_t - w^*_{\theta_t}\right\|_m^2\right] + \frac{1}{2}\alpha_{t-\tau_{\alpha_t},t-1}\psi_2, \\ \alpha_t\mathbb{E}\left[T_{332}\right] &\leq \frac{1}{2}\alpha_t\psi_1\mathbb{E}\left[\left\|w_t - w^*_{\theta_t}\right\|_m^2\right] + \frac{1}{2}\alpha_t\alpha_{t-\tau_{\alpha_t},t-1}\psi_2. \end{aligned} \tag{32}$$

Similarly, for sufficiently large t_0 and ψ_2,

$$\alpha_t T_{333} \leq \frac{1}{2}\alpha_t \psi_1 \left\| w_t - w^*_{\theta_t} \right\|_m^2 + \frac{1}{2}\alpha_t \alpha_{t-\tau_{\alpha_t}, t-1} \psi_2, \tag{33}$$

$$\alpha_t T_{334} \leq \frac{1}{2}\alpha_t \psi_1 \left\| w_t - w^*_{\theta_t} \right\|_m^2 + \frac{1}{2}\alpha_t \alpha_{t-\tau_{\alpha_t}, t-1} \psi_2. \tag{34}$$

For T_5, it is easy to see for sufficiently large t_0 and ψ_2,

$$\begin{aligned} \alpha_t^2 T_5 &\leq \frac{1}{2}\alpha_t \psi_1 \left\| w_t - w^*_{\theta_t} \right\|_m^2 + \frac{1}{2}\alpha_t^2 \psi_2 \\ &\leq \frac{1}{2}\alpha_t \psi_1 \left\| w_t - w^*_{\theta_t} \right\|_m^2 + \frac{1}{2}\alpha_t \alpha_{t-\tau_{\alpha_t}, t-1} \psi_2. \end{aligned} \tag{35}$$

For T_6, since $\beta_t < \alpha_t$, we can similarly select sufficiently large t_0 and ψ_2 such that

$$T_6 \leq \frac{1}{2}\alpha_t \alpha_{t-\tau_{\alpha_t}, t-1} \psi_2. \tag{36}$$

Putting (27), (28), (29), (30), (31), (32), (33), (34), (35), and (36) back to (24) yields

$$\begin{aligned} &\mathbb{E}\left[\left\| w_{t+1} - w^*_{\theta_{t+1}} \right\|_m^2\right] \\ \leq &(1 - \psi_1 \alpha_t)\mathbb{E}\left[\left\| w_t - w^*_{\theta_t} \right\|_m^2\right] + \beta_t \psi_2 \mathbb{E}\left[\left\| w_t - w^*_{\theta_t} \right\|_m\right] + 8\alpha_t \alpha_{t-\tau_{\alpha_t}, t-1} \psi_2 \\ \leq &(1 - \psi_1 \alpha_t)\mathbb{E}\left[\left\| w_t - w^*_{\theta_t} \right\|_m^2\right] + \beta_t \psi_2 \sqrt{\mathbb{E}\left[\left\| w_t - w^*_{\theta_t} \right\|_m^2\right]} + 8\alpha_t \alpha_{t-\tau_{\alpha_t}, t-1} \psi_2 \\ &\text{(Jensen's inequality).} \end{aligned} \tag{37}$$

(37) applies only for t such that $t - \tau_{\alpha_t} \geq 0$. According to Lemma 11, we can select a sufficiently large t_0 such that for all $t \geq t_0$, we have $t - \tau_{\alpha_t} \geq 0$. We now bound $\mathbb{E}\left[\left\| w_t - w^*_{\theta_t} \right\|_m^2\right]$ for both $t \leq t_0$ and $t \geq t_0$.

Lemma 23 *There exists a constant C_{t_0, w_0} such that for all $t \leq t_0$,*

$$\mathbb{E}\left[\left\| w_t - w^*_{\theta_t} \right\|_m^2\right] \leq C_{t_0, w_0}.$$

The proof of Lemma 23 is provided in Section E.12. We now proceed to the case of $t \geq t_0$. When t_0 is sufficiently large, Lemma 11 asserts that there exists a constant ψ_3 such that

$$8\alpha_t \alpha_{t-\tau_{\alpha_t}, t-1} \psi_2 \leq \psi_3 \frac{\log(t + t_0)}{(t + t_0)^{2\epsilon_\alpha}}.$$

Then using

$$z_t \doteq \sqrt{\mathbb{E}\left[\left\| w_t - w^*_{\theta_t} \right\|_m^2\right]}$$

as a shorthand, we get from (37) that

$$z_{t+1}^2 \leq (1 - \frac{\alpha\psi_1}{(t+t_0)^{\epsilon_\alpha}})z_t^2 + \frac{\beta\psi_2 z_t}{(t+t_0)^{\epsilon_\beta}} + \psi_3 \frac{\log(t+t_0)}{(t+t_0)^{2\epsilon_\alpha}}.$$

We now use an induction to show that $\forall t \geq t_0$,

$$z_t \leq \frac{C_0}{(t+t_0)^\epsilon}, \tag{38}$$

where $C_0 > 1$ and $\epsilon \in (0,1)$ are constants to be tuned. Since Lemma 23 asserts that $z_{t_0} \leq C_{t_0,w_0}$, we can select

$$C_0 \geq C_{t_0,w_0}(2t_0)^\epsilon$$

such that (38) holds for $t = t_0$. Now assume that (38) holds for $t = n$, then for $t = n+1$, we have

$$\begin{aligned}
&z_{n+1}^2 \\
\leq &(1 - \frac{\alpha\psi_1}{(n+t_0)^{\epsilon_\alpha}})z_n^2 + \frac{\beta\psi_2 z_n}{(n+t_0)^{\epsilon_\beta}} + \psi_3 \frac{\log(n+t_0)}{(n+t_0)^{2\epsilon_\alpha}} \\
\overset{(i)}{\leq} &(1 - \frac{\alpha\psi_1}{(n+t_0)^{\epsilon_\alpha}})\frac{C_0^2}{(n+t_0)^{2\epsilon}} + \frac{\beta\psi_2 C_0}{(n+t_0)^{\epsilon_\beta+\epsilon}} + \psi_3 \frac{\log(n+t_0)}{(n+t_0)^{2\epsilon_\alpha}} \\
= &\frac{C_0^2}{(n+t_0)^{2\epsilon}} - \frac{\alpha\psi_1 C_0^2}{(n+t_0)^{\epsilon_\alpha+2\epsilon}} + \frac{\beta\psi_2 C_0}{(n+t_0)^{\epsilon_\beta+\epsilon}} + \psi_3 \frac{\log(n+t_0)}{(n+t_0)^{2\epsilon_\alpha}} \\
\overset{(ii)}{\leq} &\frac{C_0^2}{(n+1+t_0)^{2\epsilon}} + \frac{2C_0^2}{(n+t_0)^{2\epsilon+1}} - \frac{\alpha\psi_1 C_0^2}{(n+t_0)^{\epsilon_\alpha+2\epsilon}} + \frac{\beta\psi_2 C_0}{(n+t_0)^{\epsilon_\beta+\epsilon}} + \psi_3 \frac{\log(n+t_0)}{(n+t_0)^{2\epsilon_\alpha}} \\
\leq &\frac{C_0^2}{(n+1+t_0)^{2\epsilon}} + \underbrace{\left(\frac{2}{(n+t_0)^{2\epsilon+1}} - \frac{\alpha\psi_1}{(n+t_0)^{\epsilon_\alpha+2\epsilon}} + \frac{\beta\psi_2}{(n+t_0)^{\epsilon_\beta+\epsilon}} + \psi_3 \frac{\log(n+t_0)}{(n+t_0)^{2\epsilon_\alpha}}\right)C_0^2}_{z_n'}.
\end{aligned}$$

Here (i) results from the inductive hypothesis and (ii) results from the fact that

$$x^{-2\epsilon} \leq (x+1)^{-2\epsilon} + \frac{2}{x^{2\epsilon+1}}.$$

To see the above inequality, consider

$$f(x) = x^{-2\epsilon},$$

which is convex on $(0, +\infty)$, implying

$$f(x) - f(x+1) \leq f'(x)\left(x - (x+1)\right).$$

To complete the induction, it is sufficient to ensure that $\forall n$,

$$z_n' \leq 0.$$

One way to achieve this is to select ϵ such that

$$\begin{cases} \epsilon_\alpha + 2\epsilon < 2\epsilon + 1 \\ \epsilon_\alpha + 2\epsilon < \epsilon_\beta + \epsilon \\ \epsilon_\alpha + 2\epsilon < 2\epsilon_\alpha \end{cases} \iff \begin{cases} \epsilon_\alpha < 1 \\ \epsilon < \epsilon_\beta - \epsilon_\alpha \\ \epsilon < \frac{\epsilon_\alpha}{2} \end{cases}$$

and pick t_0 sufficiently large (depending on the chosen ϵ).

With the induction completed, (38) implies that $\forall t \geq t_0$,

$$\mathbb{E}\left[\left\|w_t - w^*_{\theta_t}\right\|_m^2\right] \leq \frac{C_0^2}{(t+t_0)^{2\epsilon}}. \tag{39}$$

Combining (39) and Lemma 23, we conclude that for any

$$\epsilon_w \in (0, \min\{2(\epsilon_\beta - \epsilon_\alpha), \epsilon_\alpha\}),$$

if t_0 is sufficiently large, then $\forall t$,

$$\mathbb{E}\left[\left\|w_t - w^*_{\theta_t}\right\|_c^2\right] = \mathcal{O}\left(\frac{1}{(t+t_0)^{\epsilon_w}}\right),$$

which completes the proof. ■

Appendix B. Proofs of Section 4

B.1 Proof of Lemma 3

Lemma 24 *(Uniform contraction) Let Assumption 4.4 hold. Then, there exists an ℓ_p norm and a constant $\kappa \in (0,1)$ such that for any $\theta, q, q' \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$,*

$$\left\|\bar{F}_\theta(q) - \bar{F}_\theta(q')\right\|_p \leq \kappa \left\|q - q'\right\|_p.$$

Further, q_{π_θ} is the unique fixed point of $\bar{F}_\theta$.

Proof Assumption 4.4 implies that for any $\mu \in \bar{\Lambda}_\mu$, we have

$$d_\mu(s,a) > 0.$$

Then by the continuity of invariant distribution (Lemma 48) and the extreme value theorem, we have

$$d_{\mu,min} \doteq \inf_{\mu \in \bar{\Lambda}_\mu, s, a} d_\mu(s,a) > 0.$$

Let

$$A_\theta \doteq I - D_{\mu_\theta}(I - \gamma P_{\pi_\theta}), \tag{40}$$

then

$$\bar{F}_\theta(q) - \bar{F}_\theta(q') = A_\theta(q - q').$$

The matrix A_θ has the following properties (i). Each element of A_θ is always nonnegative

(ii). The column sum of A_θ is always smaller than 2

(iii). The row sum of A_θ is always smaller than $\kappa_0 \doteq 1-(1-\gamma)d_{\mu,min}$ and greater than 0.

To see (i), for any diagonal entry, we have

$$A_\theta(i,i) = 1 - d_{\mu_\theta}(i) + \gamma d_{\mu_\theta}(i) P_{\pi_\theta}(i,i) \geq 0;$$

for any off-diagonal entry, we have

$$A_\theta(i,j) = \gamma (D_{\mu_\theta} P_{\pi_\theta})(i,j) \geq 0.$$

To see (ii), we have

$$\mathbf{1}^\top A_\theta = \mathbf{1}^\top - d_{\mu_\theta}^\top + \gamma d_{\mu_\theta}^\top P_{\pi_\theta}.$$

Then (ii) follows immediately from the fact that $d_{\mu_\theta}^\top P_{\pi_\theta}$ is a valid probability distribution. To see (iii), we have

$$A_\theta \mathbf{1} = \mathbf{1} - d_{\mu_\theta} + \gamma d_{\mu_\theta} = \mathbf{1} - (1-\gamma) d_{\mu_\theta}.$$

Then for each i, $\left(A_\theta \mathbf{1}\right)(i) > 0$ and

$$\left(A_\theta \mathbf{1}\right)(i) = 1 - (1-\gamma) d_{\mu_\theta}(i) \leq 1 - (1-\gamma) d_{\mu,min} = \kappa_0 < 1.$$

With those three properties, for any ℓ_p norm with $p > 1$, we have

$$\begin{aligned}
&\|A_\theta x\|_p^p \\
=&\sum_i \left| \sum_j A_\theta(i,j) x_j \right|^p \\
=&\sum_i \left(\sum_k A_\theta(i,k) \right)^p \left| \sum_j \frac{A_\theta(i,j)}{\sum_k A_\theta(i,k)} x_j \right|^p \qquad \text{(Row sum of } A_\theta \text{ is strictly positive)} \\
\leq&\sum_i \left(\sum_k A_\theta(i,k) \right)^p \sum_j \frac{A_\theta(i,j)}{\sum_k A_\theta(i,k)} \left| x_j \right|^p \\
& \qquad \text{(Jensen's inequality and convexity of } |\cdot|^p) \\
=&\sum_i \left(\sum_k A_\theta(i,k) \right)^{p-1} \sum_j A_\theta(i,j) \left| x_j \right|^p \\
\leq&\sum_i \kappa_0^{p-1} \sum_j A_\theta(i,j) \left| x_j \right|^p \qquad \text{(Row sum of } A_\theta \text{ is smaller than } \kappa_0) \\
=&\kappa_0^{p-1} \sum_j \left| x_j \right|^p \sum_i A_\theta(i,j) \\
\leq& 2\kappa_0^{p-1} \sum_j \left| x_j \right|^p,
\end{aligned}$$

implying

$$\|A_\theta x\|_p \leq (2\kappa_0^{p-1})^{\frac{1}{p}} \|x\|_p.$$

Since $\kappa_0 < 1$, for sufficiently large p, we have

$$2\kappa_0^{p-1} < 1,$$

implying

$$\kappa \doteq (2\kappa_0^{p-1})^{\frac{1}{p}} < 1.$$

Consequently,

$$\left\|\bar{F}_\theta(q) - \bar{F}_\theta(q')\right\|_p = \left\|A_\theta(q - q')\right\|_p \leq \kappa\left\|q - q'\right\|_p,$$

i.e., $\bar{F}_\theta$ is a κ-contraction w.r.t. $\|\cdot\|_p$ for all θ. Further,

$$\begin{aligned} &\bar{F}_\theta(q) = q, \\ \iff &D_{\mu_\theta}(r + \gamma P_{\pi_\theta} q - q) = 0, \\ \iff &r + \gamma P_{\pi_\theta} q - q = 0, \\ \iff &q = q_{\pi_\theta}, \end{aligned}$$

which completes the proof. ■

B.2 Proof of Proposition 4

Proposition 25 *(Convergence of the critic) Let Assumptions 4.1, 4.3, and 4.4 hold. For any*

$$\epsilon_q \in (0, \min\{2(\epsilon_\beta - \epsilon_\alpha), \epsilon_\alpha\}),$$

if t_0 is sufficiently large, the iterates $\{q_t\}$ generated by Algorithm 1 satisfy

$$\mathbb{E}\left[\left\|q_t - q_{\pi_{\theta_t}}\right\|_p^2\right] = \mathcal{O}\left(\frac{1}{t^{\epsilon_q}}\right).$$

Proof As previously described, the iterates $\{q_t\}$ in Algorithm 1 evolve according to (8). We, therefore, proceed by verifying Assumptions 3.1 - 3.6 in order to invoke Theorem 2.

To start with, define

$$\mathcal{Y} \doteq \left\{(s, a, s') \mid s \in \mathcal{S}, a \in \mathcal{A}, s' \in \mathcal{S}, p(s'|s, a) > 0\right\}, \tag{41}$$

$$Y_t \doteq (S_t, A_t, S_{t+1}),$$

$$P_\theta((s_1, a_1, s_1'), (s_2, a_2, s_2')) \doteq \begin{cases} 0 & s_1' \neq s_2 \\ \mu_\theta(a_2|s_2)p(s_2'|s_2, a_2) & s_1' = s_2. \end{cases}$$

According to the action selection rule for A_t specified in Algorithm 1, we have

$$\Pr(Y_{t+1} = y) = P_{\theta_{t+1}}(Y_t, y),$$

Assumption 3.1 is then fulfilled.

Assumption 3.2 is immediately implied by Assumption 4.4. In particular, for any θ, the invariant distribution of the chain induced by P_θ is $d_{\mu_\theta}(s)\mu_\theta(a|s)p(s'|s,a)$.

Assumption 3.3 is verified by Lemma 3.

We now verify Assumption 3.4. In particular, the norm $\|\cdot\|_c$ in Section 3 is now realized as the ℓ_p norm specified by Lemma 3. We will repeatedly use the equivalence between $\|\cdot\|_\infty$, $\|\cdot\|$, and $\|\cdot\|_p$, i.e., there exist positive constants $l_{\infty,p}, u_{\infty,p}, l_{2,p}, u_{2,p}$ such that $\forall x$,

$$\begin{aligned} l_{\infty,p}\|x\|_\infty \leq \|x\|_p \leq u_{\infty,p}\|x\|_\infty \\ l_{2,p}\|x\| \leq \|x\|_p \leq u_{2,p}\|x\|. \end{aligned}$$

To verify Assumption 3.4 (i), for any $y = (s_0, a_0, s_1)$, we have,

$$\big(F_\theta(q,y) - F_\theta(q',y)\big)(s,a) = \begin{cases} q(s,a) - q'(s,a), & (s,a) \neq (s_0,a_0) \\ \gamma \sum_{a_1} \pi_\theta(a_1|s_1)\left(q(s_1,a_1) - q'(s_1,a_1)\right), & (s,a) = (s_0,a_0) \end{cases}.$$

Hence

$$\big\|F_\theta(q,y) - F_\theta(q',y)\big\|_\infty \leq \big\|q - q'\big\|_\infty,$$

implying

$$\big\|F_\theta(q,y) - F_\theta(q',y)\big\|_p \leq \frac{u_{\infty,p}}{l_{\infty,p}}\big\|q - q'\big\|_p.$$

Assumption 3.4 (i) is then fulfilled.

To verify Assumption 3.4 (ii), for any $y = (s_0, a_0, s_1)$, we have

$$\left(F_\theta(q,y) - F_{\theta'}(q,y)\right)(s,a) = \begin{cases} 0, & (s,a) \neq (s_0,a_0) \\ \gamma \sum_{a_1} \left(\pi_\theta(a_1|s_1) - \pi_{\theta'}(a_1|s_1)\right) q(s_1,a_1), & (s,a) = (s_0,a_0) \end{cases}.$$

Hence

$$\|F_\theta(q,y) - F_{\theta'}(q,y)\|_\infty \leq \gamma|\mathcal{A}|L_\pi\big\|\theta - \theta'\big\|\|q\|_\infty \quad \text{(using (12))},$$

implying

$$\|F_\theta(q,y) - F_{\theta'}(q,y)\|_p \leq \frac{u_{\infty,p}\gamma|\mathcal{A}|L_\pi}{l_{\infty,p}l_{2,p}}\big\|\theta - \theta'\big\|_p\|q\|_p.$$

Assumption 3.4 (ii) is then fulfilled.

To verify Assumption 3.4 (iii), for any $y = (s_0, a_0, s_1)$, we have

$$\left(F_\theta(0,y)\right)(s,a) = \begin{cases} 0, & (s,a) \neq (s_0,a_0) \\ r(s_0,a_0), & (s,a) = (s_0,a_0) \end{cases}.$$

Then

$$\|F_\theta(0, y)\|_p \leq u_{\infty,p} \|F_\theta(0, y)\|_\infty \leq u_{\infty,p} r_{max}.$$

Assumption 3.4 (iii) is the fulfilled.

To verify Assumption 3.4 (iv), we have

$$\bar{F}_\theta(q) - \bar{F}_{\theta'}(q) = \left(D_{\mu_\theta} - D_{\mu_{\theta'}}\right) r + \gamma \left(D_{\mu_\theta} P_{\pi_\theta} - D_{\mu_{\theta'}} P_{\pi_{\theta'}}\right) q - \left(D_{\mu_\theta} - D_{\mu_{\theta'}}\right) q.$$

Since D_{μ_θ} is Lipschitz continuous in θ (Lemma 48) and $\|D_{\mu_\theta}\|$ is bounded from above, and P_{π_θ} is Lipschitz continuous in θ (see (12)) and $\|P_{\pi_\theta}\|$ is bounded from the above, Lemma 44 confirms the Lipschitz continuity of $\bar{F}_\theta$, which completes the verification of Assumption 3.4 (iv).

To verify Assumption 3.4 (v), recall that Lemma 3 asserts that the fixed point of $\bar{F}_\theta$ is q_{π_θ}. We have

$$q_{\pi_\theta} - q_{\pi_{\theta'}} = \left((I - \gamma P_{\pi_\theta})^{-1} - (I - \gamma P_{\pi_{\theta'}})^{-1}\right) r. \tag{42}$$

Using Lemma 49 yields

$$\left\|q_{\pi_\theta} - q_{\pi_{\theta'}}\right\|_p \leq \left\|(I - \gamma P_{\pi_\theta})^{-1}\right\|_p \left\|\gamma P_{\pi_\theta} - \gamma P_{\pi_{\theta'}}\right\|_p \left\|(I - \gamma P_{\pi_{\theta'}})^{-1}\right\|_p \|r\|_p.$$

Notice that (1) for any policy π, $(I - \gamma P_\pi)^{-1}$ is always well-defined; (2) $(I - \gamma P_\pi)^{-1}$ is continuous in π (this can be seen by writing the inverse explicitly with the adjugate matrix); (3) the space of all policies is compact, by the extreme value theorem we conclude that

$$\sup_\theta \left\|(I - \gamma P_{\pi_\theta})^{-1}\right\|_p < \infty,$$

which together with (12) completes the verification of Assumption 3.4 (v).

Assumption 3.4 (vi) follows immediately from the fact that

$$|q_{\pi_\theta}(s, a)| \leq \frac{r_{max}}{1 - \gamma}.$$

Assumption 3.4 (vii) follows immediately from Assumption 4.3.

Assumption 3.5 is automatically fulfilled since in our setting we have $\epsilon_t \equiv 0$.

Assumption 3.6 is identical to Assumption 4.1 except for (11). According to the updates of $\{\theta_t\}$ in Algorithm 1, we have

$$\begin{aligned}
&\|\theta_{t+1} - \theta_t\| \\
=&\beta_t \|\rho_t \nabla_\theta \log \pi_{\theta_t}(A_t|S_t) \Pi(q_t(S_t, A_t)) - \lambda_t \nabla_\theta \mathrm{KL}\left(\mathcal{U}_\mathcal{A} || \pi_{\theta_t}(\cdot|S_t)\right)\| \\
\leq&\beta_t \left(\|\rho_t\| \|\nabla_\theta \log \pi_{\theta_t}(A_t|S_t)\| \frac{r_{max}}{1 - \gamma} + \lambda_t \|\nabla_\theta \mathrm{KL}\left(\mathcal{U}_\mathcal{A} || \pi_{\theta_t}(\cdot|S_t)\right)\|\right).
\end{aligned}$$

Assumption 4.4 and the extreme value theorem ensures that

$$\inf_{\theta, s, a} \mu_\theta(a|s) > 0.$$

Hence

$$\rho_{max} \doteq \sup_{\theta,s,a} \frac{\pi_\theta(a|s)}{\mu_\theta(a|s)} < \infty,$$

implying $\forall t$,

$$\|\rho_t\| < \infty.$$

Assumption 4.1 ensures

$$\lambda_t \leq \lambda.$$

Lemma 50 ensures the boundedness of $\|\nabla_\theta \log \pi_{\theta_t}(A_t|S_t)\|$ and $\|\nabla_\theta \mathrm{KL}\left(\mathcal{U}_\mathcal{A}||\pi_{\theta_t}(\cdot|S_t)\right)\|$, from which it is easy to see that there exists a constant L_θ such that

$$\|\theta_{t+1} - \theta_t\|_p \leq \beta_t L_\theta, \tag{43}$$

completing the verification of Assumption 3.6.

With Assumptions 3.1 - 3.6 satisfied, invoking Theorem 2 completes the proof. ■

B.3 Proof of Theorem 5

Theorem 5 *(Optimality of the actor) Let Assumptions 4.1 - 4.4 hold. Fix*

$$\epsilon_q \in \Big(2(1-\epsilon_\beta), \min\left\{2(\epsilon_\beta - \epsilon_\alpha), \epsilon_\alpha\right\}\Big).$$

Let t_0 be sufficiently large. For the iterates $\{\theta_t\}$ generated by Algorithm 1 and any $t > 0$, if k is uniformly randomly selected from the set $\left\{\lceil \frac{t}{2} \rceil, \lceil \frac{t}{2} \rceil + 1, \dots, t\right\}$ where $\lceil \cdot \rceil$ is the ceiling function, then

$$J(\pi_{\theta_k}; p_0) \geq J(\pi_*; p_0) - \mathcal{O}\left(\lambda_k\right) \tag{15}$$

holds with probability at least

$$1 - \mathcal{O}\left(\frac{1}{t^{1-\epsilon_\beta - 2\epsilon_\lambda}} + \frac{\log^2 t}{t^{\epsilon_\beta - 2\epsilon_\lambda}} + \frac{1}{t^{\epsilon_q - 2\epsilon_\lambda}}\right), \tag{16}$$

where π_* can be any optimal policy.

Proof Sketch We start with a proof sketch and then proceed to the full proof. We first define a KL regularized objective

$$J_\eta(\pi; p_0) \doteq J(\pi; p_0) - \eta \mathbb{E}_{s\sim\mathcal{U}_\mathcal{S}}\left[\mathrm{KL}\left(\mathcal{U}_\mathcal{A}||\pi(\cdot|s)\right)\right],$$

where $\mathcal{U}_\mathcal{X}$ denotes the uniform distribution on the set $\mathcal{X}$. Key to our proof is the following lemma:

Lemma 26 *(Theorem 5.2 of Agarwal et al. (2020)) For any state distribution d and d', if*

$$\left\|\nabla J_\eta(\pi_\theta; d')\right\| \leq \frac{\eta}{2|\mathcal{S} \times \mathcal{A}|},$$

then

$$J(\pi_\theta; d) \geq J(\pi_*; d) - \frac{2\eta}{1-\gamma} \max_s \frac{d_{\pi_*,\gamma,d}(s)}{d'(s)},$$

where π_* can be any optimal policy in (2).

The above lemma establishes the suboptimality of the stationary points of the KL regularized objective. If we can find those stationary points and decay the weight of the KL regularization (η) properly, optimality is then expected.

There are, however, two caveats. First, for the above lemma to be nontrivial, we have to ensure $\forall s, d'(s) > 0$. Consequently, we cannot simply set $d = d' = p_0$ because we do not make any assumption about p_0. Instead, we consider an artificial state distribution p_0' such that $\forall s, p_0'(s) > 0$ and set $d = p_0, d' = p_0'$. The second caveat is the following. To use the above lemma, we now have to optimize $J_\eta(\pi_\theta; p_0')$ to find its stationary points. This objective involves state distributions p_0' and $\mathcal{U}_\mathcal{S}$. We, however, only have access to samples from

$$d_t(s) \doteq \Pr\big(S_t = s | p_0, \mu_{\theta_0}, \dots, \mu_{\theta_{t-1}}\big).$$

We, therefore, would need to reweight them using

$$\frac{d_{\pi_\theta,\gamma,p_0'}(s)}{d_t(s)} \quad \text{and} \quad \frac{\mathcal{U}_\mathcal{S}(s)}{d_t(s)}.$$

Obviously we do not know those quantities but fortunately we can bound them. As a consequence, the reweightings can be properly accounted for even without knowing them exactly (see in particular Lemma 27). With those two caveats addressed, we are now ready to present the full proof.

Proof This proof borrows ideas from Wu et al. (2020) but is much more convoluted since we have the additional decaying KL regularization and our algorithm is off-policy without using density ratio for correcting the state distribution mismatch. Define the KL regularized objective

$$J_\eta(\pi; p_0) \doteq J(\pi; p_0) - \eta \mathbb{E}_{s\sim\mathcal{U}_\mathcal{S}}\left[\mathrm{KL}\left(\mathcal{U}_\mathcal{A} || \pi(\cdot|s)\right)\right], \tag{44}$$

where $\mathcal{U}_\mathcal{X}$ denotes the uniform distribution on the set $\mathcal{X}$. Let p_0' be an arbitrary distribution on $\mathcal{S}$ such that $p_0'(s) > 0$ holds for all $s \in \mathcal{S}$. In the rest of this proof, we use as shorthand

$$\begin{aligned} J(\theta) &\doteq J(\pi_\theta; p_0'), \\ J_\eta(\theta) &\doteq J_\eta(\pi_\theta; p_0'), \\ d_{\pi,\gamma}(s) &\doteq d_{\pi,\gamma,p_0'}(s), \end{aligned} \tag{45}$$

i.e., we work on the initial distribution p_0' (instead of p_0) by default. Note that the sampling is still done with respect to p_0, p_0' is simply an auxiliary distribution used for the proof.

Similarly, the KL regularized objective is built with a uniform distribution that does not correspond to what the algorithm implements. This too is a proof artefact. Both mismatches are accounted for, in particular in Lemma 27.

According to Lemma 7 of Mei et al. (2020), $J(\theta)$ is L_J-smoothness for some positive constant L_J w.r.t $\|\cdot\|$. Consequently, the Hessian of $J(\theta)$ is bounded from above by L_J. From Lemma 50, it is easy to see the Hessian of $\mathbb{E}_{s\sim\mathcal{U}_\mathcal{S}}\left[\mathrm{KL}\left(\mathcal{U}_\mathcal{A}||\pi_\theta(\cdot|s)\right)\right]$ is also bounded from above by some positive constant L_{KL}. Consequently, the Hessian of $J_\eta(\theta)$ is bounded from above by $L_J+\eta L_{\mathrm{KL}}$, i.e., $J_\eta(\theta)$ is $(L_J+\eta L_{\mathrm{KL}})$-smooth. With $\eta=\lambda_t$, Lemma 45 then implies

$$\begin{aligned} J_{\lambda_t}(\theta_{t+1}) \geq & J_{\lambda_t}(\theta_t) + \langle \nabla J_{\lambda_t}(\theta_t), \theta_{t+1}-\theta_t\rangle - (L_J+\lambda_t L_{\mathrm{KL}})\|\theta_{t+1}-\theta_t\|^2 \\ \geq & J_{\lambda_t}(\theta_t) + \underbrace{\langle \nabla J_{\lambda_t}(\theta_t), \theta_{t+1}-\theta_t\rangle}_{M_1} - L_J' \underbrace{\|\theta_{t+1}-\theta_t\|^2}_{M_2}, \end{aligned} \tag{46}$$

where $L_J' \doteq L_J+\lambda L_{\mathrm{KL}}$. Using (43) to bound M_2 yields

$$M_2 \leq \frac{1}{l_{2,p}}\beta_t^2 L_\theta^2.$$

To bound M_1, let

$$Y_t \doteq (S_t, A_t). \tag{47}$$

Here different from (41), we redefine Y_t to consider only state action pairs to ease presentation. For $y=(s,a)$, we define

$$\begin{aligned} \Lambda(\theta, y, \eta) \doteq & \frac{\pi_\theta(a|s)}{\mu_\theta(a|s)}\nabla \log \pi_\theta(a|s) q_{\pi_\theta}(s,a) + \frac{\eta}{|\mathcal{A}|}\sum_{\bar a}\nabla \log \pi_\theta(\bar a|s) \\ \bar\Lambda(\theta,\eta) \doteq & \sum_{s,a} d_{\mu_\theta}(s)\mu_\theta(a|s)\Lambda(\theta,y,\eta). \end{aligned} \tag{48}$$

Then we have

$$\begin{aligned} M_1 =& \langle \nabla J_{\lambda_t}(\theta_t), \theta_{t+1}-\theta_t\rangle \\ =& \beta_t\langle \nabla J_{\lambda_t}(\theta_t), \rho_t\nabla\log\pi_{\theta_t}(A_t|S_t)\Pi(q_t(S_t,A_t)) - \lambda_t\nabla\mathrm{KL}\left(\mathcal{U}_\mathcal{A}||\pi_{\theta_t}(\cdot|S_t)\right)\rangle \\ =& \beta_t\left\langle \nabla J_{\lambda_t}(\theta_t), \rho_t\nabla\log\pi_{\theta_t}(A_t|S_t)\Pi(q_t(S_t,A_t)) + \frac{\lambda_t}{|\mathcal{A}|}\sum_a \nabla\log\pi_{\theta_t}(a|S_t)\right\rangle \\ =& \beta_t\left\langle \nabla J_{\lambda_t}(\theta_t), \rho_t\nabla\log\pi_{\theta_t}(A_t|S_t)q_{\pi_{\theta_t}}(S_t,A_t) + \frac{\lambda_t}{|\mathcal{A}|}\sum_a \nabla\log\pi_{\theta_t}(a|S_t)\right\rangle \\ &+ \beta_t\left\langle \nabla J_{\lambda_t}(\theta_t), \rho_t\nabla\log\pi_{\theta_t}(A_t|S_t)\Big(\Pi\left(q_t(S_t,A_t)\right) - q_{\pi_{\theta_t}}(S_t,A_t)\Big)\right\rangle \\ =& \beta_t\underbrace{\langle \nabla J_{\lambda_t}(\theta_t), \bar\Lambda(\theta_t,\lambda_t)\rangle}_{M_{11}} + \beta_t\underbrace{\langle \nabla J_{\lambda_t}(\theta_t), \Lambda(\theta_t,Y_t,\lambda_t)-\bar\Lambda(\theta_t,\lambda_t)\rangle}_{M_{12}} \\ &+ \beta_t\underbrace{\left\langle \nabla J_{\lambda_t}(\theta_t), \rho_t\nabla\log\pi_{\theta_t}(A_t|S_t)\Big(\Pi\left(q_t(S_t,A_t)\right) - q_{\pi_{\theta_t}}(S_t,A_t)\Big)\right\rangle}_{M_{13}}. \end{aligned}$$

To bound M_{12}, define

$$\Lambda'(\theta, y, \eta) \doteq \left\langle \nabla J_\eta(\theta), \Lambda(\theta, y, \eta) - \bar{\Lambda}(\theta, \eta) \right\rangle. \tag{49}$$

Assumption 4.4 and Lemma 1 assert that there exist constants $C_0 > 0$ and $\tau \in (0, 1)$, independent of θ, such that for any $n > 0$,

$$\sup_{s,a,\theta} \sum_{s',a'} \left| P^n_{\mu_\theta}((s, a), (s', a')) - d_{\mu_\theta}(s')\mu_\theta(a'|s') \right| \le C_0 \tau^n,$$

which allows us to define

$$\tau_{\beta_t} \doteq \min \left\{ n \mid \sup_{s,a,\theta} \sum_{s',a'} \left| P^n_{\mu_\theta}((s, a), (s', a')) - d_{\mu_\theta}(s')\mu_\theta(a'|s') \right| \le \beta_t \right\}. \tag{50}$$

We then decompose M_{12} as

$$\begin{aligned}
M_{12} =& \Lambda'(\theta_t, Y_t, \lambda_t) \\
=& \underbrace{\Lambda'(\theta_t, Y_t, \lambda_t) - \Lambda'(\theta_{t-\tau_{\beta_t}}, Y_t, \lambda_t)}_{M_{121}} \\
&+ \underbrace{\Lambda'(\theta_{t-\tau_{\beta_t}}, Y_t, \lambda_t) - \Lambda'(\theta_{t-\tau_{\beta_t}}, \tilde{Y}_t, \lambda_t)}_{M_{122}} \\
&+ \underbrace{\Lambda'(\theta_{t-\tau_{\beta_t}}, \tilde{Y}_t, \lambda_t)}_{M_{123}}.
\end{aligned}$$

Here $\tilde{Y}_t$ is an auxiliary chain akin to Zou et al. (2019) and the one used in the proof of Theorem 2 in A.2 (for β_t instead of α_t). Before time $t - \tau_{\beta_t} - 1$, $\left\{\tilde{Y}_t\right\}$ is exactly the same as $\{Y_t\}$. After time $t - \tau_{\beta_t} - 1$, $\left\{\tilde{Y}_t\right\}$ evolves according to the fixed behavior policy $\mu_{\theta_{t-\tau_{\beta_t}}}$ while $\{Y_t\}$ evolves according to the changing behavior policy $\mu_{\theta_{t-\tau_{\beta_t}}}, \mu_{\theta_{t-\tau_{\beta_t}+1}}, \dots$.

$$\begin{aligned}
\left\{\tilde{Y}_t\right\} &: \cdots \to Y_{t-\tau_{\beta_t}-1} \underbrace{\to}_{\mu_{\theta_{t-\tau_{\beta_t}}}} Y_{t-\tau_{\beta_t}} \underbrace{\to}_{\mu_{\theta_{t-\tau_{\beta_t}}}} \tilde{Y}_{t-\tau_{\beta_t}+1} \underbrace{\to}_{\mu_{\theta_{t-\tau_{\beta_t}}}} \tilde{Y}_{t-\tau_{\beta_t}+2} \to \dots \\
\{Y_t\} &: \cdots \to Y_{t-\tau_{\beta_t}-1} \underbrace{\to}_{\mu_{\theta_{t-\tau_{\beta_t}}}} Y_{t-\tau_{\beta_t}} \underbrace{\to}_{\mu_{\theta_{t-\tau_{\beta_t}+1}}} Y_{t-\tau_{\beta_t}+1} \underbrace{\to}_{\mu_{\theta_{t-\tau_{\beta_t}+2}}} Y_{t-\tau_{\beta_t}+2} \to \dots.
\end{aligned} \tag{51}$$

Let us proceed to bounding each term defined above:

Lemma 27 *(Bound of M_{11}) There exists a constant $\chi_{11} > 0$ such that,*

$$M_{11} \ge \chi_{11} \|\nabla J_{\lambda_t}(\theta_t)\|^2.$$

The proof of Lemma 27 is provided in Section E.13.

Lemma 28 *(Bound of M_{121}) There exist constants $L^*_{\Lambda'} > 0$ such that*

$$\|M_{121}\| \leq \frac{L^*_{\Lambda'} L_\theta}{l_{2,p}} \beta_{t-\tau_{\beta_t},t-1}.$$

The proof of Lemma 28 is provided in Section E.14

Lemma 29 *(Bound of M_{122}) There exists a constant $U^*_{\Lambda'} > 0$ such that*

$$\|\mathbb{E}\left[M_{122}\right]\| \leq U^*_{\Lambda'} |\mathcal{S}||\mathcal{A}| L_\mu L_\theta \sum_{j=t-\tau_{\beta_t}}^{t-1} \beta_{t-\tau_{\beta_t},j}.$$

The proof of Lemma 29 is provided in Section E.15

Lemma 30 *(Bound of M_{123})*

$$\|\mathbb{E}\left[M_{123}\right]\| \leq U^*_{\Lambda'} \beta_t.$$

The proof of Lemma 30 is provided in Section E.16.

Lemma 31 *(Bound of M_{13}) There exists a constant $\rho_{max} > 0$ such that*

$$\|\mathbb{E}\left[M_{13}\right]\| \leq 2\rho_{max}\sqrt{|\mathcal{S} \times \mathcal{A}|}\sqrt{\mathbb{E}\left[\left\|q_t - q_{\pi_{\theta_t}}\right\|_\infty^2\right]}\sqrt{\mathbb{E}\left[\|\nabla J_{\lambda_t}(\theta_t)\|^2\right]}.$$

The proof of Lemma 31 is provided in Section E.17.

We now assemble the bounds of $M_{11}, M_{121}, M_{122}, M_{123}, M_{12}$ and M_2 back to (46). Similar to Lemma 11, it is easy to see for sufficiently large t_0,

$$\begin{aligned}
\tau_{\beta_t} &= \mathcal{O}\left(\log(t+t_0)\right), \\
\beta_{t-\tau_{\beta_t},t-1} &= \mathcal{O}\left(\frac{\log(t+t_0)}{(t+t_0)^{\epsilon_\beta}}\right), \\
\sum_{j=t-\tau_{\beta_t}}^{t-1} \beta_{t-\tau_{\beta_t},j} &= \mathcal{O}\left(\frac{\log^2(t+t_0)}{(t+t_0)^{\epsilon_\beta}}\right).
\end{aligned}$$

Hence if t_0 is sufficiently large, there exist positive constants $\chi_{12}, \chi_{13}, \chi_2$ such that

$$\begin{aligned}
\mathbb{E}\left[M_{121} + M_{122} + M_{123}\right] &\geq -\chi_{12}\frac{\log^2(t+t_0)}{(t+t_0)^{\epsilon_\beta}}, \\
\mathbb{E}\left[M_{13}\right] &\geq -\chi_{13}\sqrt{\mathbb{E}\left[\left\|q_t - q_{\pi_{\theta_t}}\right\|_p^2\right]}\sqrt{\mathbb{E}\left[\|\nabla J_{\lambda_t}(\theta_t)\|^2\right]}, \\
\mathbb{E}\left[M_2\right] &\leq \beta_t \chi_2 \frac{1}{(t+t_0)^{\epsilon_\beta}},
\end{aligned}$$

where the ℓ_p norm is defined by Proposition 4. Then, from (46), we get

$$\begin{aligned}\mathbb{E}\left[J_{\lambda_t}(\theta_{t+1})\right] \geq & \mathbb{E}\left[J_{\lambda_t}(\theta_t)\right] + \beta_t \chi_{11} \mathbb{E}\left[\|\nabla J_{\lambda_t}(\theta_t)\|^2\right] \\ & - \beta_t \chi_{12} \frac{\log^2(t+t_0)}{(t+t_0)^{\epsilon_\beta}} \\ & - \beta_t \chi_{13} \sqrt{\mathbb{E}\left[\left\|q_t - q_{\pi_{\theta_t}}\right\|_p^2\right]} \sqrt{\mathbb{E}\left[\|\nabla J_{\lambda_t}(\theta_t)\|^2\right]} \\ & - \beta_t \chi_2 \frac{1}{(t+t_0)^{\epsilon_\beta}}.\end{aligned} \tag{52}$$

Rearranging terms yields

$$\begin{aligned}\mathbb{E}\left[\|\nabla J_{\lambda_t}(\theta_t)\|^2\right] \leq & \frac{1}{\chi_{11}} \frac{1}{\beta_t} \left(\mathbb{E}\left[J_{\lambda_t}(\theta_{t+1})\right] - \mathbb{E}\left[J_{\lambda_t}(\theta_t)\right]\right) \\ & + \frac{\chi_{13}}{\chi_{11}} \sqrt{\mathbb{E}\left[\left\|q_t - q_{\pi_{\theta_t}}\right\|_p^2\right]} \sqrt{\mathbb{E}\left[\|\nabla J_{\lambda_t}(\theta_t)\|^2\right]} \\ & + \frac{\chi_{12} + \chi_2}{\chi_{11}} \frac{\log^2(t+t_0)}{(t+t_0)^{\epsilon_\beta}}.\end{aligned}$$

Defining

$$\chi_3 \doteq \frac{1}{\chi_{11}},\, \chi_4 \doteq \frac{\chi_{13}}{\chi_{11}},\, \chi_5 \doteq \frac{\chi_{12} + \chi_2}{\chi_{11}}$$

and telescoping the above inequality from $\lceil \frac{t}{2} \rceil$ to t yields

$$\begin{aligned}\sum_{k=\lceil \frac{t}{2} \rceil}^{t} \mathbb{E}\left[\|\nabla J_{\lambda_k}(\theta_k)\|^2\right] \leq & \chi_3 \sum_{k=\lceil \frac{t}{2} \rceil}^{t} \frac{1}{\beta_k} \left(\mathbb{E}\left[J_{\lambda_k}(\theta_{k+1})\right] - \mathbb{E}\left[J_{\lambda_k}(\theta_k)\right]\right) \\ & + \chi_4 \sum_{k=\lceil \frac{t}{2} \rceil}^{t} \sqrt{\mathbb{E}\left[\left\|q_k - q_{\pi_{\theta_k}}\right\|_p^2\right]} \sqrt{\mathbb{E}\left[\|\nabla J_{\lambda_k}(\theta_k)\|^2\right]} \\ & + \chi_5 \sum_{k=\lceil \frac{t}{2} \rceil}^{t} \frac{\log^2(k+t_0)}{(k+t_0)^{\epsilon_\beta}}.\end{aligned} \tag{53}$$

We now bound the right terms of the above inequality.

Lemma 32 *There exists a constant $U_{J,\lambda}$ such that for all t,*

$$\left|\mathbb{E}\left[J_{\lambda_t}(\theta_t)\right]\right| \leq U_{J,\lambda},\, \left|\mathbb{E}\left[J_{\lambda_t}(\theta_{t+1})\right]\right| \leq U_{J,\lambda}.$$

The proof of Lemma 32 is provided in Section E.18.

Lemma 33

$$\mathbb{E}\left[\sum_{k=\lceil \frac{t}{2} \rceil}^{t} \frac{1}{\beta_k} \left(J_{\lambda_k}(\theta_{k+1}) - J_{\lambda_k}(\theta_k)\right)\right] \leq \frac{2U_{J,\lambda}}{\beta}(t+t_0)^{\epsilon_\beta}$$

The proof of Lemma 33 is provided in Section E.19. Using Lemma 33, the Cauchy-Schwarz inequality, and

$$\sum_{k=\lceil\frac{t}{2}\rceil}^{t}\frac{\log^2(k+t_0)}{(k+t_0)^{\epsilon_\beta}}\leq\log^2(t+t_0)\int_{x=\lceil\frac{t}{2}\rceil-1}^{t}\frac{1}{(x+t_0)^{\epsilon_\beta}}dx\leq\frac{\log^2(t+t_0)}{1-\epsilon_\beta}(t+t_0)^{1-\epsilon_\beta}$$

to bound the RHS of (53) yields

$$\begin{aligned}\sum_{k=\lceil\frac{t}{2}\rceil}^{t}\mathbb{E}\left[\|\nabla J_{\lambda_k}(\theta_k)\|^2\right]\leq&\frac{2\chi_3U_{J,\lambda}}{\beta}(t+t_0)^{\epsilon_\beta}+\chi_5\frac{\log^2(t+t_0)}{1-\epsilon_\beta}(t+t_0)^{1-\epsilon_\beta}\\&+\chi_4\sqrt{\sum_{k=\lceil\frac{t}{2}\rceil}^{t}\mathbb{E}\left[\left\|q_k-q_{\pi_{\theta_k}}\right\|_p^2\right]}\sqrt{\sum_{k=\lceil\frac{t}{2}\rceil}^{t}\mathbb{E}\left[\|\nabla J_{\lambda_k}(\theta_k)\|^2\right]}.\end{aligned}$$

Multiplying $\frac{1}{t-\lceil\frac{t}{2}\rceil+1}$ in both sides yields

$$\begin{aligned}\underbrace{\frac{\sum_{k=\lceil\frac{t}{2}\rceil}^{t}\mathbb{E}\left[\|\nabla J_{\lambda_k}(\theta_k)\|^2\right]}{t-\lceil\frac{t}{2}\rceil+1}}_{z_t}\leq&\frac{2\chi_3U_{J,\lambda}}{\beta}\frac{(t+t_0)^{\epsilon_\beta}}{t-\lceil\frac{t}{2}\rceil+1}+\chi_5\frac{\log^2(t+t_0)}{1-\epsilon_\beta}\frac{(t+t_0)^{1-\epsilon_\beta}}{t-\lceil\frac{t}{2}\rceil+1}\\&+\chi_4\sqrt{\underbrace{\frac{\sum_{k=\lceil\frac{t}{2}\rceil}^{t}\mathbb{E}\left[\left\|q_k-q_{\pi_{\theta_k}}\right\|_p^2\right]}{t-\lceil\frac{t}{2}\rceil+1}}_{e_t}}\sqrt{\frac{\sum_{k=\lceil\frac{t}{2}\rceil}^{t}\mathbb{E}\left[\|\nabla J_{\lambda_k}(\theta_k)\|^2\right]}{t-\lceil\frac{t}{2}\rceil+1}}.\end{aligned}$$

It is then easy to see that there exist positive constants E_1, E_2, E_3, E_4 such that

$$\begin{aligned}z_t&\leq\frac{E_1}{t^{1-\epsilon_\beta}}+\frac{E_2\log^2 t}{t^{\epsilon_\beta}}+2E_3\sqrt{e_t}\sqrt{z_t}\\\implies(\sqrt{z_t}-E_3\sqrt{e_t})^2&\leq\frac{E_1}{t^{1-\epsilon_\beta}}+\frac{E_2\log^2 t}{t^{\epsilon_\beta}}+E_3^2e_t\\\implies\sqrt{z_t}-E_3\sqrt{e_t}&\leq\sqrt{\frac{E_1}{t^{1-\epsilon_\beta}}+\frac{E_2\log^2 t}{t^{\epsilon_\beta}}+E_3^2e_t}\\&\leq\sqrt{\frac{E_1}{t^{1-\epsilon_\beta}}+\frac{E_2\log^2 t}{t^{\epsilon_\beta}}}+E_3\sqrt{e_t}\\\implies z_t&\leq\frac{2E_1}{t^{1-\epsilon_\beta}}+\frac{2E_2\log^2 t}{t^{\epsilon_\beta}}+8E_3^2e_t.\end{aligned}$$

Proposition 4 implies that there exists a constant $E_5>0$ such that

$$e_t=\frac{\sum_{k=\lceil\frac{t}{2}\rceil}^{t}\frac{E_5}{k^{\epsilon_q}}}{t-\lceil\frac{t}{2}\rceil+1}\leq\frac{E_5t^{1-\epsilon_q}}{(1-\epsilon_q)(t-\lceil\frac{t}{2}\rceil+1)}.$$

It is then easy to see

$$e_t = \mathcal{O}\left(\frac{1}{t^{\epsilon_q}}\right),$$

implying

$$\frac{\sum_{k=\lceil \frac{t}{2} \rceil}^{t} \mathbb{E}\left[\|\nabla J_{\lambda_k}(\theta_k)\|^2\right]}{t - \lceil \frac{t}{2} \rceil + 1} = \mathcal{O}\left(\frac{1}{t^{1-\epsilon_\beta}} + \frac{\log^2 t}{t^{\epsilon_\beta}} + \frac{1}{t^{\epsilon_q}}\right). \tag{54}$$

The above inequality establishes the convergence to stationary points, with which we now study the optimality of the sequence $\{\theta_t\}$. We rely on the following lemma.

Lemma 34 *(Theorem 5.2 of Agarwal et al. (2020)) For any state distribution d and d', if*

$$\left\|\nabla J_\eta(\theta; d')\right\| \leq \frac{\eta}{2|\mathcal{S} \times \mathcal{A}|},$$

then

$$J(\theta; d) \geq J(\pi_*; d) - \frac{2\eta}{1-\gamma} \max_s \frac{d_{\pi_*, \gamma, d}(s)}{d'(s)},$$

where π_* can be any optimal policy in (2).

Obviously, for Lemma 34 to be nontrivial, we have to ensure $d'(s) > 0$.

Fix any $t > 0$. Then select a k uniformly randomly from $\left\{\lceil \frac{t}{2} \rceil, \lceil \frac{t}{2} \rceil + 1, \dots, t-1, t\right\}$. Now the random variable $\|\nabla J_{\lambda_k}(\theta_k)\|$ has randomness from both the random selection of k and the learning of θ_k. Using Markov's inequality yields

$$\begin{aligned}
\Pr\left(\|\nabla J_{\lambda_k}(\theta_k)\| \leq \frac{\lambda_t}{2|\mathcal{S} \times \mathcal{A}|}\right) =& \Pr\left(\|\nabla J_{\lambda_k}(\theta_k)\|^2 \leq \frac{\lambda_t^2}{4|\mathcal{S} \times \mathcal{A}|^2}\right) \\
\geq& 1 - \frac{4|\mathcal{S} \times \mathcal{A}|^2}{\lambda_t^2} \mathbb{E}\left[\|\nabla J_{\lambda_k}(\theta_k)\|^2\right] \\
=& 1 - \frac{4|\mathcal{S} \times \mathcal{A}|^2}{\lambda_t^2} \mathbb{E}\left[\mathbb{E}\left[\|\nabla J_{\lambda_k}(\theta_k)\|^2 \mid k\right]\right] \\
=& 1 - \frac{4|\mathcal{S} \times \mathcal{A}|^2}{\lambda_t^2} \frac{\sum_{i=\lceil \frac{t}{2} \rceil}^{t} \mathbb{E}\left[\|\nabla J_{\lambda_k}(\theta_k)\|^2 \mid k = i\right]}{t - \lceil \frac{t}{2} \rceil + 1} \\
\geq& 1 - \frac{1}{\lambda_t^2} \mathcal{O}\left(\frac{1}{t^{1-\epsilon_\beta}} + \frac{\log^2 t}{t^{\epsilon_\beta}} + \frac{1}{t^{\epsilon_q}}\right) \quad \text{(Using (54))} \\
\geq& 1 - C_t,
\end{aligned}$$

where

$$C_t \doteq \mathcal{O}\left(\frac{1}{t^{1-\epsilon_\beta - 2\epsilon_\lambda}} + \frac{\log^2 t}{t^{\epsilon_\beta - 2\epsilon_\lambda}} + \frac{1}{t^{\epsilon_q - 2\epsilon_\lambda}}\right).$$

Since $\lambda_k \geq \lambda_t$, we have

$$\|\nabla J_{\lambda_k}(\theta_k)\| \leq \frac{\lambda_t}{2|\mathcal{S} \times \mathcal{A}|} \implies \|\nabla J_{\lambda_k}(\theta_k)\| \leq \frac{\lambda_k}{2|\mathcal{S} \times \mathcal{A}|}.$$

Consequently,

$$\Pr\left(\|\nabla J_{\lambda_k}(\theta_k)\| \leq \frac{\lambda_k}{2|\mathcal{S} \times \mathcal{A}|}\right) \geq \Pr\left(\|\nabla J_{\lambda_k}(\theta_k)\| \leq \frac{\lambda_t}{2|\mathcal{S} \times \mathcal{A}|}\right) \geq 1 - C_t.$$

Let $d = p_0, d' = p_0', \eta = \lambda_k$ in Lemma 34 and recall (45), we get

$$J(\theta_k; p_0) \geq J(\pi_*; p_0) - 2\frac{\lambda_k}{1-\gamma} \max_s \frac{d_{\pi_*,\gamma,p_0}(s)}{p_0'(s)}.$$

holds with at least probability

$$1 - C_t,$$

which completes the proof.

■

Appendix C. Proofs of Section 5

C.1 Proof of Proposition 6

Proposition 35 *(Convergence of the critic) Let Assumptions 4.1, 4.3, and 4.4 hold. Then there exists an ℓ_p norm such that for any*

$$\epsilon_q \in (0, \min\{2(\epsilon_\beta - \epsilon_\alpha), \epsilon_\alpha\}),$$

if t_0 is sufficiently large, the iterates $\{q_t\}$ generated by Algorithm 1 satisfy

$$\mathbb{E}\left[\left\|q_t - \tilde{q}_{\pi_{\theta_t},\lambda_t}\right\|_p^2\right] = \mathcal{O}\left(\frac{1}{t^{\epsilon_q}}\right).$$

Proof The proof is similar to the proof of Proposition 4. To start with, define

$$\begin{aligned}
\mathcal{Y} &\doteq \left\{(s, a, s') \mid s \in \mathcal{S}, a \in \mathcal{A}, s' \in \mathcal{S}, p(s'|s,a) > 0\right\}, \\
Y_t &\doteq (S_t, A_t, S_{t+1}), \\
P_\zeta((s_1, a_1, s_1'), (s_2, a_2, s_2')) &\doteq \begin{cases} 0 & s_1' \neq s_2 \\ \mu_\theta(a_2|s_2) p(s_2'|s_2, a_2) & s_1' = s_2. \end{cases}
\end{aligned}$$

According to the action selection rule for A_t specified in Algorithm 2, we have

$$\Pr(Y_{t+1} = y) = P_{\zeta_{t+1}}(Y_t, y),$$

Assumption 3.1 is then fulfilled.

Assumption 3.2 is immediately implied by Assumption 4.4. In particular, for any ζ, the invariant distribution of the chain induced by P_ζ is $d_{\mu_\theta}(s)\mu_\theta(a|s)p(s'|s,a)$.

To verify Assumption 3.3, first notice that

$$\begin{aligned}
\bar{F}_\zeta(q) &= \sum_{s,a,s'} d_{\mu_\theta}(s)\mu_\theta(a|s)p(s'|s,a)F_\zeta(q,s,a,s') \\
&= D_{\mu_\theta}\left(r + \gamma P_{\pi_\theta}\left(q - \eta \log \pi_\theta\right) - q\right) + q \\
&= (I - D_{\mu_\theta}(I - \gamma P_{\pi_\theta}))q + D_{\mu_\theta}(r - \eta\gamma P_{\pi_\theta} \log \pi_\theta),
\end{aligned}$$

where π_θ denotes a vector in $\mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$ whose (s,a)-indexed element is $\pi_\theta(a|s)$ and $\log \pi_\theta$ is the elementwise logarithm of π_θ. Then, we have

$$\bar{F}_\zeta(q) - \bar{F}_\zeta(q') = A_\theta(q - q'),$$

where A_θ is defined as in (40):

$$A_\theta \doteq I - D_{\mu_\theta}(I - \gamma P_{\pi_\theta}).$$

According to the proof of Lemma 3, there exist a $\kappa \in (0,1)$ and an ℓ_p norm such that $\forall x$,

$$\left\|A_\theta x\right\|_p \leq \kappa \left\|x\right\|_p,$$

implying

$$\left\|\bar{F}_\zeta(q) - \bar{F}_\zeta(q')\right\|_p \leq \kappa \left\|q - q'\right\|_p.$$

Further,

$$\begin{aligned}
&\bar{F}_\zeta(q) = q \\
\iff &r + \gamma P_{\pi_\theta}(q - \eta \log \pi_\theta) - q = 0 \\
\iff &q = \tilde{q}_{\pi_\theta,\eta} \quad \text{(Lemma 1 of Haarnoja et al. (2018))},
\end{aligned}$$

which completes the verification of Assumption 3.3.

We now verify Assumption 3.4. In particular, the norm $\left\|\cdot\right\|_c$ in Section 3 is now realized as the ℓ_p norm above.

To verify Assumption 3.4 (i), for any $y = (s_0, a_0, s_1)$, we have

$$\left(F_\zeta(q,y) - F_\zeta(q',y)\right)(s,a) = \begin{cases} q(s,a) - q'(s,a), & (s,a) \neq (s_0,a_0) \\ \gamma \sum_{a_1} \pi_\theta(a_1|s_1)\left(q(s_1,a_1) - q'(s_1,a_1)\right), & (s,a) = (s_0,a_0) \end{cases}.$$

Hence

$$\left\|F_\zeta(q,y) - F_\zeta(q',y)\right\|_\infty \leq \left\|q - q'\right\|_\infty,$$

implying

$$\left\|F_\zeta(q,y) - F_\zeta(q',y)\right\|_p \leq \frac{u_{\infty,p}}{l_{\infty,p}}\left\|q - q'\right\|_p.$$

Assumption 3.4 (i) is then fulfilled.

To verify Assumption 3.4 (ii), for any $y = (s_0, a_0, s_1)$, we have

$$\begin{aligned}
&\left(F_{\zeta_t}(q,y) - F_{\zeta_k}(q,y)\right)(s,a) \\
=&\begin{cases} 0, & (s,a) \neq (s_0,a_0) \\ \gamma \sum_{a_1} \left(\pi_{\theta_t}(a_1|s_1) - \pi_{\theta_k}(a_1|s_1)\right) q(s_1,a_1) + \lambda_t \mathbb{H}\left(\pi_{\theta_t}(\cdot|s_1)\right) - \lambda_k \mathbb{H}\left(\pi_{\theta_k}(\cdot|s_1)\right), & (s,a) = (s_0,a_0) \end{cases}.
\end{aligned}$$

Since

$$\begin{aligned}
&\left|\lambda_t \mathbb{H}\left(\pi_{\theta_t}(\cdot|s_1)\right) - \lambda_k \mathbb{H}\left(\pi_{\theta_k}(\cdot|s_1)\right)\right| \\
\leq& |\lambda_t - \lambda_k| \mathbb{H}\left(\pi_{\theta_t}(\cdot|s_1)\right) + \lambda_k \left|\mathbb{H}\left(\pi_{\theta_t}(\cdot|s_1)\right) - \mathbb{H}\left(\pi_{\theta_k}(\cdot|s_1)\right)\right| \\
\leq& |\lambda_t - \lambda_k| \log|\mathcal{A}| + \lambda_k \left(\log|\mathcal{A}| + e^{-1}\right) \|\theta_t - \theta_k\| \quad \text{(Lemma 50)} \\
\leq& \left(\log|\mathcal{A}| + \lambda \log|\mathcal{A}| + \lambda e^{-1}\right) \|\zeta_t - \zeta_k\|,
\end{aligned}$$

we have

$$\begin{aligned}
&\left\|F_{\zeta_k}(q,y) - F_{\zeta_t}(q,y)\right\|_\infty \\
\leq& \gamma|\mathcal{A}| L_\pi \|\theta_t - \theta_k\| \|q\|_\infty + \left(\log|\mathcal{A}| + \lambda \log|\mathcal{A}| + \lambda e^{-1}\right) \|\zeta_t - \zeta_k\| \\
\leq& \gamma|\mathcal{A}| L_\pi \|\zeta_t - \zeta_k\| \|q\|_\infty + \left(\log|\mathcal{A}| + \lambda \log|\mathcal{A}| + \lambda e^{-1}\right) \|\zeta_t - \zeta_k\| \\
\leq& \frac{\gamma|\mathcal{A}| L_\pi}{l_{2,p} l_{\infty,p}} \|\zeta_t - \zeta_k\|_p \|q\|_p + \frac{\left(\log|\mathcal{A}| + \lambda \log|\mathcal{A}| + \lambda e^{-1}\right)}{l_{2,p}} \|\zeta_t - \zeta_k\|_p
\end{aligned}$$

Assumption 3.4 (ii) is then fulfilled.

To verify Assumption 3.4 (iii), for any $y = (s_0, a_0, s_1)$, we have

$$\left(F_{\zeta_t}(0,y)\right)(s,a) = \begin{cases} 0, & (s,a) \neq (s_0,a_0) \\ r(s_0,a_0) + \gamma \lambda_t \mathbb{H}\left(\pi_{\theta_t}(\cdot|s_1)\right), & (s,a) = (s_0,a_0) \end{cases}.$$

Then

$$\left\|F_{\zeta_t}(0,y)\right\|_p \leq u_{\infty,p} \|F_\theta(0,y)\|_\infty \leq u_{\infty,p}\left(r_{max} + \gamma\lambda \log|\mathcal{A}|\right).$$

Assumption 3.4 (iii) is then fulfilled.

To verify Assumption 3.4 (iv), we have

$$\bar{F}_{\zeta_t}(q) - \bar{F}_{\zeta_k}(q) = \bar{F}_{\theta_t}(q) - \bar{F}_{\theta_k}(q) - \gamma\lambda_t D_{\mu_{\theta_t}} P_{\pi_{\theta_t}} \log \pi_{\theta_t} + \gamma\lambda_k D_{\mu_{\theta_k}} P_{\pi_{\theta_k}} \log \pi_{\theta_k},$$

where $\bar{F}_\theta$ is defined in (14). In the proof of Proposition 4, we already show that there exist constants C_1 and C_2 such that

$$\left\|\bar{F}_{\theta_t}(q) - \bar{F}_{\theta_k}(q)\right\|_p \leq C_1 \|\theta_t - \theta_k\|_p \left(\|q\|_p + C_2\right) \leq C_1 \|\zeta_t - \zeta_k\|_p \left(\|q\|_p + C_2\right).$$

We now bound the remaining parts $-\gamma\lambda_t D_{\mu_{\theta_t}} P_{\pi_{\theta_t}} \log \pi_{\theta_t} + \gamma\lambda_k D_{\mu_{\theta_k}} P_{\pi_{\theta_k}} \log \pi_{\theta_k}$. First, notice that

$$\begin{aligned}
\left(P_{\pi_\theta} \log \pi_\theta\right)(s,a) &= \sum_{s'} p(s'|s,a) \sum_{a'} \pi_\theta(a'|s') \log \pi_\theta(a'|s') \\
&= -\sum_{s'} p(s'|s,a) \mathbb{H}\left(\pi_\theta(\cdot|s')\right).
\end{aligned}$$

It is easy to see $\mathbb{H}\left(\pi_\theta(\cdot|s')\right)$ is Lipschitz continuous in θ (Lemma 50) and is bounded by $\log|\mathcal{A}|$. We, therefore, conclude that $P_{\pi_\theta} \log \pi_\theta$ is Lipschitz continuous in θ and is bounded from above. Since D_{μ_θ} is also Lipschitz continuous in θ (Lemma 48) and is bounded from above, Lemma 44 asserts that there exists constants C_3 and C_4 such that

$$\begin{aligned}
\|D_{\mu_\theta} P_{\pi_\theta} \log \pi_\theta\| &\leq C_3, \\
\left\|D_{\mu_\theta} P_{\pi_\theta} \log \pi_\theta - D_{\mu_{\theta'}} P_{\pi_{\theta'}} \log \pi_{\theta'}\right\| &\leq C_4 \left\|\theta - \theta'\right\|,
\end{aligned}$$

implying

$$\begin{aligned}
&\left\|\lambda_k D_{\mu_{\theta_k}} P_{\pi_{\theta_k}} \log \pi_{\theta_k} - \lambda_t D_{\mu_{\theta_t}} P_{\pi_{\theta_t}} \log \pi_{\theta_t}\right\| \\
\leq &\|\lambda_k - \lambda_t\| \left\|D_{\mu_{\theta_k}} P_{\pi_{\theta_k}} \log \pi_{\theta_k}\right\| + \lambda_t \left\|D_{\mu_{\theta_k}} P_{\pi_{\theta_k}} \log \pi_{\theta_k} - D_{\mu_{\theta_t}} P_{\pi_{\theta_t}} \log \pi_{\theta_t}\right\| \\
\leq &C_3 \|\lambda_k - \lambda_t\| + \lambda C_4 \|\theta_t - \theta_k\| \\
\leq &(C_3 + \lambda C_4) \|\zeta_t - \zeta_k\|,
\end{aligned}$$

which completes the verification of Assumption 3.4 (iv).

To verify Assumptions 3.4 (v), it suffices to show that

$$\left\|\tilde{q}_{\pi_{\theta_t},\lambda_t} - \tilde{q}_{\pi_{\theta_k},\lambda_k}\right\|_p \leq C_5 \|\zeta_t - \zeta_k\|_p$$

holds for some positive constant C_5. According to (19), it suffices to show that for some positive constant C_6,

$$\left\|\tilde{v}_{\pi_{\theta_t},\lambda_t} - \tilde{v}_{\pi_{\theta_k},\lambda_k}\right\|_p \leq C_6 \|\zeta_t - \zeta_k\|_p.$$

Recall by definition

$$\tilde{v}_{\pi_{\theta_t},\lambda_t}(s) = v_{\pi_{\theta_t}}(s) + \lambda_t \underbrace{\mathbb{E}_{\pi_{\theta_t}}\left[\sum_{i=0}^{\infty} \gamma^i \mathbb{H}\left(\pi_{\theta_t}(\cdot|S_{t+i})\right) \mid S_t = s\right]}_{\mathbb{H}_{\theta_t}(s)}. \tag{55}$$

Clearly,

$$|\mathbb{H}_\theta(s)| \leq \frac{\log|\mathcal{A}|}{1-\gamma}.$$

We now show that $\mathbb{H}_\theta(s)$ is Lipschitz continuous in θ. Let $p_{0,s}$ denote the distribution on $\mathcal{S}$ such that all its mass concentrates on the state s, i.e., $p_{0,s}(s) = 1$. We can then express $\mathbb{H}_\theta(s)$ as

$$\mathbb{H}_\theta(s) = \frac{1}{1-\gamma} \sum_s d_{\pi_\theta,\gamma,p_{0,s}}(s) \mathbb{H}\left(\pi_\theta(\cdot|s)\right).$$

It is easy to see that

$$\begin{aligned}
&d_{\pi_\theta,\gamma,p_{0,s}}(s'') \\
=&(1-\gamma)\sum_{t=0}^\infty \gamma^t \Pr\big(S_t = s''|S_0 \sim p_{0,s}\big) \\
=&(1-\gamma)p_{0,s}(s'') + (1-\gamma)\sum_{t=1}^\infty \gamma^t \Pr\big(S_t = s''|S_0 \sim p_{0,s}\big) \\
=&(1-\gamma)p_{0,s}(s'') + (1-\gamma)\sum_{t=0}^\infty \gamma^{t+1} \Pr\big(S_{t+1} = s''|S_0 \sim p_{0,s}\big) \\
=&(1-\gamma)p_{0,s}(s'') + \gamma(1-\gamma)\sum_{t=0}^\infty \gamma^t \sum_{s'} \Pr\big(S_t = s'|S_0 \sim p_{0,s}\big) \Pr\big(S_{t+1} = s''|S_t = s'\big) \\
=&(1-\gamma)p_{0,s}(s'') + \gamma \sum_{s'} \Pr\big(S_{t+1} = s''|S_t = s'\big) d_{\pi_\theta,\gamma,p_{0,s}}(s').
\end{aligned}$$

In a matrix form, we have

$$\begin{aligned}
&d_{\pi_\theta,\gamma,p_{0,s}} =(1-\gamma)p_{0,s} + \gamma P_{\pi_\theta}^\top d_{\pi_\theta,\gamma,p_{0,s}} \\
\implies &d_{\pi_\theta,\gamma,p_{0,s}} =(1-\gamma)(I - \gamma P_{\pi_\theta}^\top)^{-1} p_{0,s},
\end{aligned}$$

where we have abused the notation a bit to use P_{π_θ} to also denote the state transition matrix under the policy π_θ. Similar to (42), we conclude that $d_{\pi_\theta,\gamma,p_{0,s}}$ is Lipschitz continuous in θ. Lemma 50 confirms that $\mathbb{H}\left(\pi_\theta(\cdot|s)\right)$ is Lipschitz continuous in θ. Hence Lemma 44 asserts that $\mathbb{H}_\theta(s)$ is Lipschitz continuous in θ, i.e., there exists a positive constant such that

$$|\mathbb{H}_\theta(s) - \mathbb{H}_{\theta'}(s)| \leq C_7 \big\|\theta - \theta'\big\|. \tag{56}$$

Similar to (42), we can also show that there exists a constant C_8 such that

$$\left|v_{\pi_\theta}(s) - v_{\pi_{\theta'}}(s)\right| \leq C_8 \big\|\theta - \theta'\big\|. \tag{57}$$

We, therefore, have

$$\begin{aligned}
&\left|\tilde{v}_{\pi_{\theta_t},\lambda_t}(s) - \tilde{v}_{\pi_{\theta_k},\lambda_k}(s)\right| \\
\leq&\left|v_{\pi_{\theta_t}}(s) - v_{\pi_{\theta_k}}(s)\right| + |\lambda_t - \lambda_k||\mathbb{H}_{\theta_t}(s)| + \lambda_k|\mathbb{H}_{\theta_t}(s) - \mathbb{H}_{\theta_k}(s)| \\
\leq&C_8\|\theta_t - \theta_k\| + |\lambda_t - \lambda_k|\frac{\log|\mathcal{A}|}{1-\gamma} + \lambda C_7\|\theta_t - \theta_k\| \\
\leq&\left(C_8 + \frac{\log|\mathcal{A}|}{1-\gamma} + \lambda C_7\right)\|\zeta_t - \zeta_k\|,
\end{aligned}$$

which completes the verification of Assumption 3.4 (v).

For Assumption 3.4 (vi), first notice that the soft action value function $\tilde{q}_{\pi,\eta}$ can be regarded as the normal action value function q_π w.r.t. to the reward

$$r(s,a) + \eta \sum_{s'} p(s'|s,a) \mathbb{H}\left(\pi(\cdot|s')\right)$$

Then it is easy to see

$$\left|\tilde{q}_{\pi_{\theta_t},\lambda_t}(s,a)\right| \leq U_{\tilde{J}} \doteq \frac{r_{max} + \lambda \log |\mathcal{A}|}{1-\gamma}, \tag{58}$$

which completes the verification of Assumption 3.4 (vi).

For Assumption 3.4 (vii), we have

$$\left|P_{\zeta_t}(y,y') - P_{\zeta_k}(y,y')\right| \leq C_9 \|\theta_t - \theta_k\| \leq C_9 \|\zeta_t - \zeta_k\|,$$

where the existence of the positive constant C_9 is ensured by Assumption 4.3.

Assumption 3.5 is automatically fulfilled since in our setting we have $\epsilon_t \equiv 0$.

Assumption 3.6 is automatically implied by Assumption 4.1 except for (11). According to the updates of $\{\theta_t\}$ in Algorithm 2, we have

$$\begin{aligned}
&\|\theta_{t+1} - \theta_t\| \\
=&\beta_t \left\| \sum_a \pi_{\theta_t}(a|S_t) \nabla_\theta \log \pi_{\theta_t}(a|S_t) \Big(\Pi\big(q_t(S_t,a)\big) - \lambda_t \log \pi_{\theta_t}(a|S_t) \Big) \right\| \\
=&\beta_t \left\| \sum_a \nabla \pi_{\theta_t}(a|S_t) \Big(\Pi\big(q_t(S_t,a)\big) - \lambda_t \log \pi_{\theta_t}(a|S_t) \Big) \right\| \\
\leq&\beta_t \left\| \sum_a \nabla \pi_{\theta_t}(a|S_t) \Pi\big(q_t(S_t,a)\big) \right\| + \beta_t \lambda_t \left\| \sum_a \nabla \pi_{\theta_t}(a|S_t) \log \pi_{\theta_t}(a|S_t) \right\| \\
\leq&\beta_t 2U_{\tilde{J}} + \beta_t \lambda_t \left\| \sum_a \nabla \pi_{\theta_t}(a|S_t) \log \pi_{\theta_t}(a|S_t) \right\| \quad \text{((58) and Lemma 50)} \\
=&\beta_t 2U_{\tilde{J}} + \beta_t \lambda_t \left\| \sum_a \nabla \pi_{\theta_t}(a|S_t) \log \pi_{\theta_t}(a|S_t) + \pi_{\theta_t}(a|S_t) \nabla \log \pi_{\theta_t}(a|S_t) \right\| \\
=&\beta_t 2U_{\tilde{J}} + \beta_t \lambda_t \|\nabla \mathbb{H}\left(\pi_{\theta_t}(\cdot|S_t)\right)\| \\
\leq&\beta_t 2U_{\tilde{J}} + \beta_t \lambda_t \left(\log |\mathcal{A}| + e^{-1}\right) \quad \text{(Lemma 50)} \\
\leq&\beta_t \underbrace{\left(2U_{\tilde{J}} + \lambda \log |\mathcal{A}| + \lambda e^{-1}\right)}_{L_\theta}.
\end{aligned} \tag{59}$$

Further,

$$
\begin{aligned}
&|\lambda_{t+1} - \lambda_t| \\
=&\lambda \left(\frac{1}{(t+t_0)^{\epsilon_\lambda}} - \frac{1}{(t+t_0+1)^{\epsilon_\lambda}} \right) \\
=&\lambda \frac{(t+t_0+1)^{\epsilon_\lambda} - (t+t_0)^{\epsilon_\lambda}}{(t+t_0)^{\epsilon_\lambda}(t+t_0+1)^{\epsilon_\lambda}} \\
=&\lambda \frac{(t+t_0+1)^{\epsilon_\lambda}(t+t_0)^{1-\epsilon_\lambda} - (t+t_0)}{(t+t_0)(t+t_0+1)^{\epsilon_\lambda}} \\
\leq&\lambda \frac{(t+t_0+1)^{\epsilon_\lambda}(t+t_0+1)^{1-\epsilon_\lambda} - (t+t_0)}{(t+t_0)(t+t_0)^{\epsilon_\lambda}} \\
=&\frac{\lambda}{(t+t_0)^{1+\epsilon_\lambda}} \\
=&\beta_t \frac{\lambda (t+t_0)^{\epsilon_\beta}}{\beta (t+t_0)^{1+\epsilon_\lambda}} \\
\leq&\beta_t \frac{\lambda}{\beta}.
\end{aligned}
$$

We, therefore, conclude that that there exists a constant L_ζ such that

$$
\left\|\zeta_{t+1} - \zeta_t\right\|_p \leq \beta_t L_\zeta,
$$

which completes the verification of Assumption 3.6.

With Assumptions 3.1 - 3.6 satisfied, invoking Theorem 2 completes the proof. ∎

C.2 Proof of Theorem 7

Theorem 7 *(Convergence of the actor) Let Assumptions 4.1, 4.3, and 4.4 hold. Fix any*

$$
\epsilon_q \in \Big(0, \min\left\{2(\epsilon_\beta - \epsilon_\alpha), \epsilon_\alpha\right\}\Big).
$$

Let t_0 be sufficiently large. Fix any $\epsilon_0 > 0$ and any state distribution p_0'. For the iterates $\{\theta_t\}$ generated by Algorithm 2 and any $t > 0$, if k is uniformly randomly selected from the set $\left\{\lceil \frac{t}{2} \rceil, \lceil \frac{t}{2} \rceil + 1, \dots, t\right\}$, then

$$
\left\|\nabla \tilde{J}_{\lambda_k}(\pi_{\theta_k}; p_0')\right\|^2 \leq \frac{1}{k^{\epsilon_0}}
$$

holds with at least probability

$$
1 - \mathcal{O}\left(\frac{1}{t^{1-\epsilon_\beta-\epsilon_0}} + \frac{\log^2 t}{t^{\epsilon_\beta-\epsilon_0}} + \frac{1}{t^{\epsilon_q-\epsilon_0}} \right).
$$

Proof In this proof, we use as shorthand

$$
\begin{aligned}
J(\theta) &\doteq J(\pi_\theta; p_0'), \\
\tilde{J}_\eta(\theta) &\doteq \tilde{J}_\eta(\pi_\theta; p_0'), \\
d_{\pi,\gamma}(s) &\doteq d_{\pi,\gamma,p_0'}(s),
\end{aligned}
$$

i.e., we work on the initial distribution p_0' (instead of p_0). Recall the entropy regularized discounted total rewards is defined as

$$
\tilde{J}_\eta(\pi_\theta; p_0') = J(\pi_\theta; p_0') + \eta \underbrace{\frac{1}{1-\gamma} \sum_s d_{\pi_\theta,\gamma,p_0'}(s) \mathbb{H}\left(\pi_\theta(\cdot|s)\right)}_{\mathbb{H}(\pi_\theta)}.
$$

According to Lemma 7 of Mei et al. (2020), $J(\theta)$ is L_J-smooth for some positive constant L_J w.r.t $\|\cdot\|$. According to Lemma 14 of Mei et al. (2020), $\mathbb{H}(\pi_\theta)$ is L_H-smooth for some positive constant L_H w.r.t. $\|\cdot\|$. Hence $\tilde{J}_\eta(\theta)$ is $(L_J+\eta L_H)$-smooth. With $\eta = \lambda_t$, Lemma 45 then implies

$$
\begin{aligned}
\tilde{J}_{\lambda_t}(\theta_{t+1}) \geq & \tilde{J}_{\lambda_t}(\theta_t) + \left\langle \nabla \tilde{J}_{\lambda_t}(\theta_t), \theta_{t+1} - \theta_t \right\rangle - (L_J + \lambda_t L_H) \|\theta_{t+1} - \theta_t\|^2 \\
\geq & \tilde{J}_{\lambda_t}(\theta_t) + \underbrace{\left\langle \nabla \tilde{J}_{\lambda_t}(\theta_t), \theta_{t+1} - \theta_t \right\rangle}_{\tilde{M}_1} - \tilde{L}_J' \underbrace{\|\theta_{t+1} - \theta_t\|^2}_{\tilde{M}_2},
\end{aligned}
$$

where $\tilde{L}_J' \doteq L_J + \lambda L_H$. Using (59) to bound $\tilde{M}_2$ yields

$$
\tilde{M}_2 \leq \frac{1}{l_{2,p}} \beta_t^2 L_\theta^2.
$$

To bound $\tilde{M}_1$, we reuse the Y_t and $\tilde{Y}_t$ defined in (47) and (51). For any s, we define

$$
\begin{aligned}
\Lambda_1(\theta, s, \eta) &\doteq \sum_a \pi_\theta(s,a) \nabla \log \pi_\theta(a|s) \left(\tilde{q}_{\pi_\theta,\eta}(s,a) - \eta \log \pi_\theta(a|s)\right), \qquad (60) \\
\bar{\Lambda}_1(\theta, \eta) &\doteq \sum_s d_{\mu_\theta}(s) \Lambda(\theta, s, \eta).
\end{aligned}
$$

Then we have

$$
\begin{aligned}
&\tilde{M}_1 \\
=&\left\langle \nabla \tilde{J}_{\lambda_t}(\theta_t), \theta_{t+1} - \theta_t \right\rangle \\
=&\beta_t \left\langle \nabla \tilde{J}_{\lambda_t}(\theta_t), \sum_a \pi_{\theta_t}(a|S_t) \nabla \log \pi_{\theta_t}(a|S_t) \Big(\Pi(q_t(S_t, a)) - \lambda_t \log \pi_{\theta_t}(a|S_t) \Big) \right\rangle \\
=&\beta_t \underbrace{\left\langle \nabla \tilde{J}_{\lambda_t}(\theta_t), \bar{\Lambda}_1(\theta_t, \lambda_t) \right\rangle}_{\tilde{M}_{11}} \\
&+ \beta_t \underbrace{\left\langle \nabla \tilde{J}_{\lambda_t}(\theta_t), \sum_a \nabla \pi_{\theta_t}(a|S_t) \Big(\tilde{q}_{\pi_{\theta_t}, \lambda_t}(S_t, a) - \lambda_t \log \pi_{\theta_t}(a|S_t) \Big) - \bar{\Lambda}_1(\theta_t, \lambda_t) \right\rangle}_{\tilde{M}_{12}} \\
&+ \beta_t \underbrace{\left\langle \nabla \tilde{J}_{\lambda_t}(\theta_t), \sum_a \nabla \pi_{\theta_t}(a|S_t) \Big(\Pi(q_t(S_t, a)) - \tilde{q}_{\pi_{\theta_t}, \lambda_t}(S_t, a) \Big) \right\rangle}_{\tilde{M}_{13}}.
\end{aligned}
$$

To bound $\tilde{M}_{12}$, define

$$
\Lambda_1'(\theta, s, \eta) \doteq \left\langle \nabla \tilde{J}_\eta(\theta), \Lambda_1(\theta, s, \eta) - \bar{\Lambda}_1(\theta, \eta) \right\rangle. \tag{61}
$$

We then decompose $\tilde{M}_{12}$ as

$$
\begin{aligned}
&\tilde{M}_{12} = \Lambda_1'(\theta_t, S_t, \lambda_t) \\
=&\underbrace{\Lambda_1'(\theta_t, S_t, \lambda_t) - \Lambda_1'(\theta_{t-\tau_{\beta_t}}, S_t, \lambda_t)}_{\tilde{M}_{121}} + \underbrace{\Lambda_1'(\theta_{t-\tau_{\beta_t}}, S_t, \lambda_t) - \Lambda_1'(\theta_{t-\tau_{\beta_t}}, \tilde{S}_t, \lambda_t)}_{\tilde{M}_{122}} \\
&+ \underbrace{\Lambda_1'(\theta_{t-\tau_{\beta_t}}, \tilde{S}_t, \lambda_t)}_{\tilde{M}_{123}},
\end{aligned}
$$

where we recall that $\tilde{S}_t$ is defined as part of $\tilde{Y}_t$ in (51). Let us proceed to bounding each term defined above.

Lemma 36 *(Bound of $\tilde{M}_{11}$) There exists a constant $\chi_{11} > 0$ such that,*

$$
\tilde{M}_{11} \geq \chi_{11} \left\| \nabla \tilde{J}_{\lambda_t}(\theta_t) \right\|^2.
$$

The proof of Lemma 36 is provided in Section E.20.

Lemma 37 *(Bound of $\tilde{M}_{121}$) There exist constants $L^*_{\Lambda_1'} > 0$ such that*

$$
\left\| \tilde{M}_{121} \right\| \leq L^*_{\Lambda_1'} L_\theta \beta_{t-\tau_{\beta_t}, t-1},
$$

where L_θ is defined in (59).

The proof of Lemma 37 is provided in Section E.21

Lemma 38 *(Bound of $\tilde{M}_{122}$) There exists a constant $U^*_{\Lambda'_1} > 0$ such that*

$$\left\|\mathbb{E}\left[\tilde{M}_{122}\right]\right\| \leq U^*_{\Lambda'}|\mathcal{S}||\mathcal{A}|L_\mu L_\theta \sum_{j=t-\tau_{\beta_t}}^{t-1} \beta_{t-\tau_{\beta_t},j}.$$

The proof of Lemma 38 is identical to the proof of Lemma 29 in Section E.15 up to change of notations and is thus omitted.

Lemma 39 *(Bound of $\tilde{M}_{123}$)*

$$\left\|\mathbb{E}\left[\tilde{M}_{123}\right]\right\| \leq U^*_{\Lambda'_1}\beta_t.$$

The proof of Lemma 39 is identical to the proof of Lemma 30 in Section E.16 up to change of notations and is thus omitted.

Lemma 40 *(Bound of $\tilde{M}_{13}$) The exists a constant $\chi_{13} > 0$ such that*

$$\left\|\mathbb{E}\left[\tilde{M}_{13}\right]\right\| \leq \chi_{13}\sqrt{\mathbb{E}\left[\left\|q_t - \tilde{q}_{\pi_{\theta_t},\lambda_t}\right\|_\infty^2\right]}\sqrt{\mathbb{E}\left[\left\|\nabla \tilde{J}_{\lambda_t}(\theta_t)\right\|^2\right]}.$$

The proof of Lemma 40 is identical to the proof Lemma 31 in Section E.17 up to change of notations and is thus omitted.

Now using exactly the same routine as the proof of Theorem 5 in Section B.3, we obtain that there exists some positive constants χ_3, χ_4 and χ_5 such that

$$\begin{aligned}
\sum_{k=\lceil \frac{t}{2} \rceil}^{t} \mathbb{E}\left[\left\|\nabla \tilde{J}_{\lambda_k}(\theta_k)\right\|^2\right] \leq& \chi_3 \sum_{k=\lceil \frac{t}{2} \rceil}^{t} \frac{1}{\beta_k}\left(\mathbb{E}\left[\tilde{J}_{\lambda_k}(\theta_{k+1})\right] - \mathbb{E}\left[\tilde{J}_{\lambda_k}(\theta_k)\right]\right) \\
&+ \chi_4 \sum_{k=\lceil \frac{t}{2} \rceil}^{t} \sqrt{\mathbb{E}\left[\left\|q_k - \tilde{q}_{\pi_{\theta_k},\lambda_k}\right\|_p^2\right]}\sqrt{\mathbb{E}\left[\left\|\nabla \tilde{J}_{\lambda_k}(\theta_k)\right\|^2\right]} \\
&+ \chi_5 \sum_{k=\lceil \frac{t}{2} \rceil}^{t} \frac{\log^2(k+t_0)}{(k+t_0)^{\epsilon_\beta}},
\end{aligned}$$

where the ℓ_p norm is defined in Proposition 6. To continue mimicing the proof of Theorem 5, we need to establish counterparts of Lemmas 32 and 33 to bound the first summation in the RHS of the above inequality. The counterpart of Lemma 32 is trivial since by the definition of $\tilde{J}_\eta(\theta)$ we have $\forall t, \theta$

$$\left|\tilde{J}_{\lambda_t}(\theta)\right| \leq U_{\tilde{J}},$$

where $U_{\tilde{J}}$ is defined in (58). This simplification is because that $\mathbb{H}\left(\pi(\cdot|s)\right)$ is always bounded by $\log|\mathcal{A}|$ but $\mathrm{KL}\left(\mathcal{U}_\mathcal{A}||\pi(\cdot|s)\right)$ can be unbounded. Then we have

Lemma 41

$$\sum_{k=\lceil \frac{t}{2} \rceil}^{t} \frac{1}{\beta_k} \left(\tilde{J}_{\lambda_k}(\theta_{k+1}) - \tilde{J}_{\lambda_k}(\theta_k) \right) \leq \frac{3\lambda\beta \log |\mathcal{A}|}{1-\gamma} + \frac{2U_{\tilde{J}}}{\beta}(t+t_0)^{\epsilon_\beta}$$

The proof of Lemma 41 is provided in Section E.22. Using the same routine as the proof of Theorem 5 yields

$$\frac{\sum_{k=\lceil \frac{t}{2} \rceil}^{t} \mathbb{E}\left[\left\|\nabla \tilde{J}_{\lambda_k}(\theta_k)\right\|^2\right]}{t - \lceil \frac{t}{2} \rceil + 1} = \underbrace{\mathcal{O}\left(\frac{1}{t^{1-\epsilon_\beta}} + \frac{\log^2 t}{t^{\epsilon_\beta}} + \frac{1}{t^{\epsilon_q}}\right)}_{\tilde{C}_t}. \tag{62}$$

We now analyze the above equality from a probabilistic perspective. Consider a positive non-increasing sequence $\{\delta_t\}$ to be tuned. Fix any $t > 0$. Then select a k uniformly randomly from $\left\{\lceil \frac{t}{2} \rceil, \lceil \frac{t}{2} \rceil + 1, \dots, t-1, t\right\}$. Now the random variable $\left\|\nabla \tilde{J}_{\lambda_k}(\theta_k)\right\|$ has randomness from both the random selection of k and the learning of θ_k. Using Markov's inequality yields

$$\begin{aligned}
\Pr\left(\left\|\nabla \tilde{J}_{\lambda_k}(\theta_k)\right\|^2 \leq \delta_t\right) \geq& 1 - \frac{1}{\delta_t}\mathbb{E}\left[\left\|\nabla \tilde{J}_{\lambda_k}(\theta_k)\right\|^2\right] \\
=& 1 - \frac{1}{\delta_t}\mathbb{E}\left[\mathbb{E}\left[\left\|\nabla \tilde{J}_{\lambda_k}(\theta_k)\right\|^2 \mid k\right]\right] \\
=& 1 - \frac{1}{\delta_t}\frac{\sum_{i=\lceil \frac{t}{2} \rceil}^{t} \mathbb{E}\left[\left\|\nabla \tilde{J}_{\lambda_k}(\theta_k)\right\|^2 \mid k = i\right]}{t - \lceil \frac{t}{2} \rceil + 1} \\
\geq& 1 - \frac{1}{\delta_t}\tilde{C}_t. \quad \text{(Using (62))}
\end{aligned}$$

Since $\delta_k \geq \delta_t$, we have

$$\left\|\nabla \tilde{J}_{\lambda_k}(\theta_k)\right\|^2 \leq \delta_t \implies \left\|\nabla \tilde{J}_{\lambda_k}(\theta_k)\right\|^2 \leq \delta_k.$$

Consequently,

$$\Pr\left(\left\|\nabla \tilde{J}_{\lambda_k}(\theta_k)\right\|^2 \leq \delta_k\right) \geq \Pr\left(\left\|\nabla \tilde{J}_{\lambda_k}(\theta_k)\right\|^2 \leq \delta_t\right) \geq 1 - \frac{1}{\delta_t}\tilde{C}_t.$$

Letting

$$\delta_t \doteq \frac{1}{t^{\epsilon_0}}$$

then completes the proof. ■

C.3 Proof of Corollary 8

Corollary 42 *(Optimality of the actor) Let Assumptions 4.1, 4.3, and 4.4 hold. Fix any*

$$\epsilon_q \in \Big(0, \min\left\{2(\epsilon_\beta - \epsilon_\alpha), \epsilon_\alpha\right\}\Big).$$

Let t_0 be sufficiently large. Let $\{\delta_t\}$ be any positive decreasing sequence converging to 0. For the iterates $\{\theta_t\}$ generated by Algorithm 2 and any $t > 0$, if k is uniformly randomly selected from the set $\left\{\lceil \frac{t}{2} \rceil, \lceil \frac{t}{2} \rceil + 1, \dots, t\right\}$, then

$$J(\pi_{\theta_k}; p_0) \geq J(\pi_*; p_0) - \mathcal{O}\left(\lambda_k\right) - \mathcal{O}\left(\frac{\delta_k}{\lambda_k \left(\min_{s,a} \pi_{\theta_k}(a|s)\right)^2}\right)$$

holds with at least probability

$$1 - \frac{\mathcal{O}\left(t^{-(1-\epsilon_\beta)} + t^{-\epsilon_\beta}\log^2 t + t^{-\epsilon_q}\right)}{\delta_t},$$

where π_* can be any optimal policy in (2).

Proof Fix any state distribution p_0' satisfying $\forall s, p_0'(s) > 0$. Then, from the proof of Theorem 7 in Section C.2, we conclude that

$$\left\|\nabla \tilde{J}_{\lambda_k}(\pi_{\theta_k}; p_0')\right\|^2 \leq \delta_k \tag{63}$$

holds with probability at least

$$1 - \frac{\tilde{C}_t}{\delta_t}.$$

With the convergence to stationary points established in (63), we now use the following lemma from Mei et al. (2020) to study the optimality. Let $\pi_{*,\eta}$ be the optimal policy w.r.t. the soft value function, i.e., $\forall \pi, s$,

$$\tilde{v}_{\pi,\eta}(s) \leq \tilde{v}_{\pi_{*,\eta},\eta}(s),$$

then we have

Lemma 43 *(Lemma 15 of Mei et al. (2020)) For any state distribution d and d',*

$$\tilde{J}_\eta(\pi_\theta; d) \geq \tilde{J}_\eta(\pi_{*,\eta}; d) - \frac{|\mathcal{S}|\left\|\nabla \tilde{J}_\eta(\pi_\theta; d')\right\|^2}{2\eta \min_s d'(s) \left(\min_{s,a} \pi_\theta(a|s)\right)^2} \max_s \frac{d_{\pi_{*,\eta},\gamma,d}(s)}{d_{\pi_\theta,\gamma,d'}(s)}.$$

Obviously, for Lemma 43 to be nontrivial, we have to ensure $\forall s, d'(s) > 0$.

Letting $d = p_0, d' = p_0'$ in Lemma 43 and using (63) yield that

$$\begin{aligned}
\tilde{J}_{\lambda_k}(\pi_{\theta_k}; p_0) \geq& \tilde{J}_{\lambda_k}(\pi_{*,\lambda_k}; p_0) - \frac{|\mathcal{S}|\left\|\nabla \tilde{J}_{\lambda_k}(\pi_{\theta_k}; p_0')\right\|^2}{2\lambda_k \min_s p_0'(s) \left(\min_{s,a} \pi_{\theta_k}(a|s)\right)^2} \max_s \frac{d_{\pi_{*,\lambda_k},\gamma,p_0}(s)}{d_{\pi_{\theta_k},\gamma,p_0'}(s)} \\
\geq& \tilde{J}_{\lambda_k}(\pi_{*,\lambda_k}; p_0) - \frac{|\mathcal{S}|\delta_k}{2\lambda_k \min_s p_0'(s) \left(\min_{s,a} \pi_{\theta_k}(a|s)\right)^2} \frac{1}{\frac{\min_s p_0'(s)}{1-\gamma}}
\end{aligned} \tag{64}$$

holds with probability at least

$$1 - \frac{\tilde{C}_t}{\delta_t}.$$

According to Proposition 2 of Dai et al. (2018), we have

$$\max_s \left| \tilde{v}_{\pi_{*,\eta},\eta}(s) - v_{\pi_*}(s) \right| \leq \frac{\eta \log |\mathcal{A}|}{1-\gamma},$$

implying

$$\left| \tilde{J}_\eta(\pi_{*,\eta}; p_0) - J(\pi_*; p_0) \right| \leq \frac{\eta \log |\mathcal{A}|}{1-\gamma},$$

i.e.,

$$\tilde{J}_\eta(\pi_{*,\eta}; p_0) \geq J(\pi_*; p_0) - \frac{\eta \log |\mathcal{A}|}{1-\gamma}. \tag{65}$$

From (20), it is easy to see

$$\tilde{J}_\eta(\pi; p_0) \leq J(\pi; p_0) + \frac{\eta \log |\mathcal{A}|}{1-\gamma}. \tag{66}$$

Putting (65) and (66) back to (64) yields

$$J(\pi_{\theta_k}; p_0) \geq J(\pi_*; p_0) - \frac{2\lambda_k \log |\mathcal{A}|}{1-\gamma} - \frac{(1-\gamma)|\mathcal{S}|\delta_k}{2\lambda_k \left(\min_s p_0'(s) \min_{s,a} \pi_{\theta_k}(a|s)\right)^2},$$

which completes the proof. ■

Appendix D. Technical Lemmas

Lemma 44 *Let $f_1(x), f_2(x)$ be two Lipschitz continuous functions with Lipschitz constants L_1, L_2. Assume $\|f_1(x)\| \leq U_1, \|f_2(x)\| \leq U_2$, then $L_1U_2 + L_2U_1$ is a Lipschitz constant of $f(x) \doteq f_1(x)f_2(x)$.*

Proof

$$\begin{aligned}
&\|f_1(x)f_2(x) - f_1(y)f_2(y)\| \\
\leq &\|f_1(x)\|\|f_2(x) - f_2(y)\| + \|f_2(y)\|\|f_1(x) - f_1(y)\| \\
\leq &(U_1L_2 + U_2L_1)\|x - y\|.
\end{aligned}$$

■

Lemma 45 *The following statements about a differentiable function $f(x)$ are equivalent:*

(i). $f(x)$ *is* L*-smooth w.r.t. a norm* $\|\cdot\|_s$*.*

(ii). $\|\nabla f(x) - \nabla f(y)\|_s^* \le L\|x - y\|_s$*.*

(iii). $|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \le \frac{L}{2}\|x - y\|_s^2$*.*

Proof See e.g. Definition 5.1 and Lemma 5.7 of Beck (2017). ■

Lemma 46 *For any* x, x'*,*

$$
\begin{aligned}
\langle \nabla M(x), x' \rangle &\le \|x\|_m \|x'\|_m, \\
\langle \nabla M(x), x \rangle &\ge \|x\|_m^2.
\end{aligned}
$$

Proof The proof is taken from Section A.2 of Chen et al. (2020) and we include it here for completeness. Since $M(x) = \frac{1}{2}\|x\|_m^2$, by Theorem 3.47 of Beck (2017),

$$
\nabla M(x) = \|x\|_m v_x,
$$

where v_x is a subgradient of $\|x\|_m$ at x. Consequently,

$$
\begin{aligned}
\langle \nabla M(x), x' \rangle &= \|x\|_m \langle v_x, x' \rangle \\
&\le \|x\|_m \|v_x\|_m^* \|x'\|_m \\
&\le \|x\|_m \|x'\|_m,
\end{aligned}
$$

where the first inequality results from Holder's inequality and the last inequality results from the fact that $\|v_x\|_m^* \le 1$ (Lemma A.1 of Chen et al. (2020)).

Further, notice that $\|x\|_m$ is convex, we thus have

$$
\|x\|_m \le \|0\|_m + \langle v_x, x - 0 \rangle,
$$

implying

$$
\langle \nabla M(x), x \rangle = \|x\|_m \langle v_x, x \rangle \ge \|x\|_m^2.
$$

■

Lemma 47 *Given positive integers* $t_1 < t_2$ *satisfying*

$$
\alpha_{t_1, t_2 - 1} \le \frac{1}{4A},
$$

we have, for any $t \in [t_1, t_2]$*,*

$$
\begin{aligned}
\|w_t - w_{t_1}\|_c &\le 2\alpha_{t_1, t_2-1}(A\|w_{t_1}\|_c + B), && (67) \\
\|w_t - w_{t_1}\|_c &\le 4\alpha_{t_1, t_2-1}(A\|w_{t_2}\|_c + B), && (68) \\
\|w_t - w_{t_1}\|_c &\le \min\{\|w_{t_1}\|_c, \|w_{t_2}\|_c\} + \frac{B}{A}. && (69)
\end{aligned}
$$

Proof Notice that

$$
\begin{aligned}
&\|w_{t+1}\|_c - \|w_t\|_c \\
\leq&\|w_{t+1} - w_t\|_c \\
\leq&\alpha_t \|F_{\theta_t}(w_t, Y_t) - w_t + \epsilon_t\|_c \\
\leq&\alpha_t \left(\|F_{\theta_t}(w_t, Y_t)\|_c + \|w_t\|_c + \|\epsilon_t\|_c\right) \\
\leq&\alpha_t (U_F + (L_F + 1)\|w_t\|_c + \|\epsilon_t\|_c) \quad \text{(Lemma 61)} \\
\leq&\alpha_t (U_F + U'_\epsilon + (U_\epsilon + L_F + 1)\|w_t\|_c) \quad \text{(Assumption 3.5)} \\
\leq&\alpha_t (A\|w_t\|_c + B) \quad \text{(Using (26))}
\end{aligned} \tag{70}
$$

The rest of the proof is exactly the same as the proof of Lemma A.2 of Chen et al. (2021) up to changes of notations. We include it for completeness. Rearranging terms of the above inequality yields

$$
\|w_{t+1}\|_c + \frac{B}{A} \leq (1 + \alpha_t A)\left(\|w_t\|_c + \frac{B}{A}\right),
$$

implying that for any $t \in (t_1, t_2]$,

$$
\|w_t\|_c + \frac{B}{A} \leq \prod_{j=t_1}^{t-1}(1 + A\alpha_j)\left(\|w_{t_1}\|_c + \frac{B}{A}\right).
$$

Notice that for any $x \in [0, \frac{1}{2}]$, $1 + x \leq \exp(x) \leq 1 + 2x$ always hold. Hence

$$
\alpha_{t_1, t_2 - 1} \leq \frac{1}{4A}
$$

implies

$$
\prod_{j=t_1}^{t-1}(1 + A\alpha_j) \leq \exp(A\alpha_{t_1, t-1}) \leq 1 + 2A\alpha_{t_1, t-1}.
$$

Consequently, for any $t \in (t_1, t_2]$, we have

$$
\begin{aligned}
&\|w_t\|_c + \frac{B}{A} \leq (1 + 2A\alpha_{t_1, t-1})\left(\|w_{t_1}\|_c + \frac{B}{A}\right) \\
\Longrightarrow\ &\|w_t\|_c \leq (1 + 2A\alpha_{t_1, t-1})\,\|w_{t_1}\|_c + 2B\alpha_{t_1, t-1},
\end{aligned}
$$

which together with (70) yields that for any $t \in (t_1, t_2 - 1]$

$$
\begin{aligned}
\|w_{t+1} - w_t\|_c &\leq \alpha_t\left(A\|w_t\|_c + B\right) \\
&\leq \alpha_t\left(A\left(1 + 2A\alpha_{t_1, t-1}\right)\|w_{t_1}\|_c + 2AB\alpha_{t_1, t-1} + B\right) \\
&\leq 2\alpha_t (A\|w_{t_1}\|_c + B) \quad \left(\text{Using } \alpha_{t_1, t-1} \leq \frac{1}{4A}\right) \quad .
\end{aligned}
$$

Consequently, for any $t \in (t_1, t_2]$, we have

$$\begin{aligned}\|w_t - w_{t_1}\|_c \leq \sum_{j=t_1}^{t-1} \|w_{j+1} - w_j\|_c \leq \sum_{j=t_1}^{t-1} 2\alpha_j (A\|w_{t_1}\|_c + B) \\ = 2\alpha_{t_1,t-1}(A\|w_{t_1}\|_c + B) \leq 2\alpha_{t_1,t_2-1}(A\|w_{t_1}\|_c + B),\end{aligned}$$

which completes the proof of (67). For (68), we have

$$\begin{aligned}\|w_{t_2} - w_{t_1}\|_c \leq & 2\alpha_{t_1,t_2-1}(A\|w_{t_1}\|_c + B) \\ \leq & 2\alpha_{t_1,t_2-1}(A\|w_{t_1} - w_{t_2}\|_c + A\|w_{t_2}\|_c + B) \\ \leq & \frac{1}{2}\|w_{t_1} - w_{t_2}\|_c + 2\alpha_{t_1,t_2-1}(A\|w_{t_2}\|_c + B),\end{aligned}$$

implying

$$\|w_{t_2} - w_{t_1}\|_c \leq 4\alpha_{t_1,t_2-1}(A\|w_{t_2}\|_c + B).$$

Consequently, for any $t \in [t_1, t_2]$,

$$\begin{aligned}\|w_t - w_{t_1}\|_c \leq & 2\alpha_{t_1,t_2-1}(A\|w_{t_1}\|_c + B) \\ \leq & 2\alpha_{t_1,t_2-1}(A\|w_{t_1} - w_{t_2}\|_c + A\|w_{t_2}\|_c + B) \\ \leq & 2\alpha_{t_1,t_2-1}\left(A4\alpha_{t_1,t_2-1}(A\|w_{t_2}\|_c + B) + A\|w_{t_2}\|_c + B\right) \\ \leq & 4\alpha_{t_1,t_2-1}(A\|w_{t_2}\|_c + B) \quad (\text{Using } \alpha_{t_1,t_2-1} \leq \frac{1}{4A}) \quad ,\end{aligned}$$

which completes the proof of (68). (67) implies

$$\|w_t - w_{t_1}\| \leq \|w_{t_1}\|_c + \frac{B}{A},$$

(68) implies

$$\|w_t - w_{t_1}\| \leq \|w_{t_2}\|_c + \frac{B}{A},$$

then (69) follows immediately, which completes the proof. ■

Lemma 48 *Let Assumptions 4.3 and 4.4 hold. Then there exists a constant L'_μ such that $\forall \theta, \theta', a, s$,*

$$\left|d_{\mu_\theta}(s,a) - d_{\mu_{\theta'}}(s,a)\right| \leq L'_\mu \|\theta - \theta'\|.$$

Proof See, e.g., Lemma 9 of Zhang et al. (2021). ■

Lemma 49 *For any $\|\cdot\|$, we have*

$$\left\|X^{-1} - Y^{-1}\right\| \leq \left\|X^{-1}\right\| \|X - Y\| \left\|Y^{-1}\right\|.$$

Proof

$$\left\|X^{-1} - Y^{-1}\right\| = \left\|X^{-1}YY^{-1} - X^{-1}XY^{-1}\right\| \leq \left\|X^{-1}\right\|\|X - Y\|\left\|Y^{-1}\right\|.$$

■

Lemma 50 *With softmax parameterization,*

$$\frac{\mathrm{d}\pi_\theta(a|s)}{\mathrm{d}\theta_{s',a'}} = \mathbb{I}_{s=s'}\pi_\theta(a|s)\left(\mathbb{I}_{a=a'} - \pi_\theta(a'|s)\right), \tag{71}$$

$$\frac{\mathrm{d}\log\pi_\theta(a|s)}{\mathrm{d}\theta_{s',a'}} = \mathbb{I}_{s=s'}\left(\mathbb{I}_{a=a'} - \pi_\theta(a'|s)\right), \tag{72}$$

$$\frac{\mathrm{d}KL\left(\mathcal{U}_\mathcal{A}||\pi_\theta(\cdot|s)\right)}{\mathrm{d}\theta_{s',a'}} = \mathbb{I}_{s=s'}(\pi_\theta(a'|s) - \frac{1}{|\mathcal{A}|}), \tag{73}$$

$$\sum_a \frac{\mathrm{d}\pi_\theta(a|s)}{\mathrm{d}\theta_{s',a'}} q_{\pi_\theta}(s,a) = \mathbb{I}_{s=s'}\pi_\theta(a'|s)Adv_{\pi_\theta}(s,a'), \tag{74}$$

$$\frac{\mathrm{d}J(\pi_\theta;p_0)}{\mathrm{d}\theta_{s,a}} = \frac{1}{1-\gamma}d_{\pi_\theta,\gamma,p_0}(s)\pi_\theta(a|s)Adv_{\pi_\theta}(s,a), \tag{75}$$

$$\|\nabla\mathbb{H}\left(\pi_\theta(\cdot|s)\right)\| \leq \log|\mathcal{A}| + e^{-1}, \tag{76}$$

$$\sum_a \frac{\mathrm{d}\pi_\theta(a|s)}{\mathrm{d}\theta_{s',a'}}\left(\tilde{q}_{\pi_\theta,\eta}(s,a) - \eta\log\pi_\theta(a|s)\right) = \mathbb{I}_{s=s'}\pi_\theta(a'|s)\tilde{Adv}_{\pi_\theta,\eta}(s,a) \tag{77}$$

$$\frac{\mathrm{d}\tilde{J}_\eta(\pi_\theta;p_0)}{\mathrm{d}\theta_{s,a}} = \frac{1}{1-\gamma}d_{\pi_\theta,\gamma,p_0}(s)\pi_\theta(a|s)\tilde{Adv}_{\pi_\theta,\eta}(s,a), \tag{78}$$

where

$$\tilde{Adv}_{\pi_\theta,\eta}(s,a) \doteq \tilde{q}_{\pi_\theta,\eta}(s,a') - \eta\log\pi_\theta(a'|s) - \tilde{v}_{\pi_\theta,\eta}(s),$$
$$Adv_{\pi_\theta,\eta}(s,a) \doteq q_{\pi_\theta}(s,a) - v_{\pi_\theta}(s).$$

Further, for any s*,* $\mathbb{H}\left(\pi_\theta(\cdot|s)\right)$ *is* $(4 + 8\log|\mathcal{A}|)$*-smooth.*

Proof (71) is well-known. For (72), we have

$$\frac{\mathrm{d}\log\pi_\theta(a|s)}{\mathrm{d}\theta_{s',a'}} = \frac{1}{\pi_\theta(a|s)}\frac{\mathrm{d}\pi_\theta(a|s)}{\mathrm{d}\theta_{s',a'}} = \mathbb{I}_{s=s'}\left(\mathbb{I}_{a=a'} - \pi_\theta(a'|s)\right).$$

For (73), we have

$$\begin{aligned}
\frac{\mathrm{dKL}\left(\mathcal{U}_\mathcal{A}||\pi_\theta(\cdot|s)\right)}{\mathrm{d}\theta_{s',a'}} &= -\frac{\mathbb{I}_{s=s'}}{|\mathcal{A}|}\sum_a \frac{\mathrm{d}\log\pi_\theta(a|s)}{\mathrm{d}\theta_{s',a'}} \\
&= -\frac{\mathbb{I}_{s=s'}}{|\mathcal{A}|}\sum_a \left(\mathbb{I}_{a=a'} - \pi_\theta(a'|s)\right).
\end{aligned}$$

Since

$$\sum_a \left(\mathbb{I}_{a=a'} - \pi_\theta(a'|s)\right) = \left(\sum_a \left(0 - \pi_\theta(a'|s)\right)\right) + 1 = 1 - |\mathcal{A}|\pi_\theta(a'|s),$$

we have

$$\frac{\mathrm{dKL}\left(\mathcal{U}||\pi_\theta(\cdot|s)\right)}{\mathrm{d}\theta_{s',a'}} = \mathbb{I}_{s=s'}(\pi(a'|s) - \frac{1}{|\mathcal{A}|}).$$

For (74),

$$\begin{aligned}
&\sum_a \frac{\mathrm{d}\pi_\theta(a|s)}{\mathrm{d}\theta_{s',a'}} q_{\pi_\theta}(s,a) \\
=&\sum_a \mathbb{I}_{s=s'}\pi_\theta(a|s)\left(\mathbb{I}_{a=a'} - \pi_\theta(a'|s)\right) q_{\pi_\theta}(s,a) \\
=&\mathbb{I}_{s=s'}\left(\pi_\theta(a'|s)q_{\pi_\theta}(s,a') + \sum_a \pi_\theta(a|s)\left(0 - \pi_\theta(a'|s)\right) q_{\pi_\theta}(s,a)\right) \\
=&\mathbb{I}_{s=s'}\left(\pi_\theta(a'|s)q_{\pi_\theta}(s,a') - \pi_\theta(a'|s)v_{\pi_\theta}(s)\right).
\end{aligned}$$

For (75), see, e.g., Lemma C.1 of Agarwal et al. (2020). For (76), we have

$$\begin{aligned}
&\frac{\mathrm{d}\mathbb{H}\left(\pi_\theta(\cdot|s)\right)}{\mathrm{d}\theta_{s',a'}} \\
=&-\mathbb{I}_{s=s'}\sum_a \frac{\mathrm{d}\pi_\theta(a|s)}{\mathrm{d}\theta_{s',a'}}\log\pi_\theta(a|s) + 0 \\
=&-\mathbb{I}_{s=s'}\sum_a \pi_\theta(a|s)\left(\mathbb{I}_{a=a'} - \pi_\theta(a'|s)\right)\log\pi_\theta(a|s) \\
=&-\mathbb{I}_{s=s'}\left(\pi_\theta(a'|s)\mathbb{H}\left(\pi_\theta(\cdot|s)\right) + \pi_\theta(a'|s)\log\pi_\theta(a'|s)\right),
\end{aligned}$$

implying

$$\|\nabla\mathbb{H}\left(\pi_\theta(\cdot|s)\right)\| \le \log|\mathcal{A}| + e^{-1}.$$

By setting $\gamma = 0$ and putting all the mass of ρ (initial distribution) in s in Lemma 14 of Mei et al. (2020), we obtain that $\mathbb{H}\left(\pi_\theta(\cdot|s)\right)$ is $(4 + 8\log|\mathcal{A}|)$-smooth. For (77), we have

$$\begin{aligned}
&\sum_a \frac{\mathrm{d}\pi_\theta(a|s)}{\mathrm{d}\theta_{s',a'}}\left(\tilde{q}_{\pi_\theta,\eta}(s,a) - \eta\log\pi_\theta(a|s)\right) \\
=&\sum_a \mathbb{I}_{s=s'}\pi_\theta(a|s)\left(\mathbb{I}_{a=a'} - \pi_\theta(a'|s)\right)\left(\tilde{q}_{\pi_\theta,\eta}(s,a) - \eta\log\pi_\theta(a|s)\right) \\
=&\mathbb{I}_{s=s'}\left(\pi_\theta(a'|s)\left(\tilde{q}_{\pi_\theta,\eta}(s,a') - \eta\log\pi_\theta(a'|s)\right) - \sum_a \pi_\theta(a|s)\pi_\theta(a'|s)\left(\tilde{q}_{\pi_\theta,\eta}(s,a) - \eta\log\pi_\theta(a|s)\right)\right) \\
=&\mathbb{I}_{s=s'}\pi_\theta(a'|s)\left(\tilde{q}_{\pi_\theta,\eta}(s,a') - \eta\log\pi_\theta(a'|s) - \tilde{v}_{\pi_\theta,\eta}(s)\right).
\end{aligned}$$

Since (78) is identical to Lemma 10 of Mei et al. (2020), we have completed the proof. ∎

Appendix E. Proof of Auxiliary Lemmas

E.1 Proof of Lemma 12

Lemma 51 *(Bound of T_1)*

$$T_1 \leq \frac{L_w L_\theta \beta_t}{l_{cm}} \left\| w_t - w^*_{\theta_t} \right\|_m.$$

Proof

$$\begin{aligned} T_1 =& \left\langle \nabla M(w_t - w^*_{\theta_t}), w^*_{\theta_t} - w^*_{\theta_{t+1}} \right\rangle \\ \leq& \left\| w_t - w^*_{\theta_t} \right\|_m \left\| w^*_{\theta_t} - w^*_{\theta_{t+1}} \right\|_m \quad \text{(Lemma 46)} \\ \leq& \left\| w_t - w^*_{\theta_t} \right\|_m \frac{L_w L_\theta \beta_t}{l_{cm}} \quad \text{(Assumptions 3.4, 3.6 and Lemma 10)} \quad . \end{aligned}$$

■

E.2 Proof of Lemma 13

Lemma 52 *(Bound of T_2)*

$$T_2 \leq -(1 - \kappa \frac{u_{cm}}{l_{cm}}) \left\| w_t - w^*_{\theta_t} \right\|_m^2.$$

Proof

$$\begin{aligned} T_2 =& \left\langle \nabla M(w_t - w^*_{\theta_t}), \bar{F}_{\theta_t}(w_t) - w_t \right\rangle \\ =& \left\langle \nabla M(w_t - w^*_{\theta_t}), \bar{F}_{\theta_t}(w_t) - \bar{F}_{\theta_t}(w^*_{\theta_t}) \right\rangle - \left\langle \nabla M(w_t - w^*_{\theta_t}), w_t - w^*_{\theta_t} \right\rangle \\ & \qquad (w^*_{\theta_t} \text{ is the fixed point}). \end{aligned}$$

To bound the first inner product, we have

$$\begin{aligned} & \left\langle \nabla M(w_t - w^*_{\theta_t}), \bar{F}_{\theta_t}(w_t) - \bar{F}_{\theta_t}(w^*_{\theta_t}) \right\rangle \\ \leq& \left\| w_t - w^*_{\theta_t} \right\|_m \left\| \bar{F}_{\theta_t}(w_t) - \bar{F}_{\theta_t}(w^*_{\theta_t}) \right\|_m \quad \text{(Lemma 46)} \\ \leq& \left\| w_t - w^*_{\theta_t} \right\|_m \frac{1}{l_{cm}} \kappa \left\| w_t - w^*_{\theta_t} \right\|_c \\ \leq& \frac{u_{cm}\kappa}{l_{cm}} \left\| w_t - w^*_{\theta_t} \right\|_m^2 \end{aligned}$$

For the second inner product, Lemma 46 implies that

$$\left\langle \nabla M(w_t - w^*_{\theta_t}), w_t - w^*_{\theta_t} \right\rangle \geq \left\| w_t - w^*_{\theta_t} \right\|_m^2.$$

Putting the bounds for the two inner products together completes the proof. ■

E.3 Proof of Lemma 14

Lemma 53 *(Bound of T_{31})*

$$T_{31} \leq \frac{8L(L_w L_\theta + 1)\alpha_{t-\tau_{\alpha_t},t-1}}{\xi l_{cs}^2} \left(u_{cm}^2 A^2 \left\| w_t - w_{\theta_t}^* \right\|_m^2 + C^2 \right).$$

Proof

$$\begin{aligned}
T_{31} =& \left\langle \nabla M(w_t - w_{\theta_t}^*) - \nabla M(w_{t-\tau_{\alpha_t}} - w_{\theta_{t-\tau_{\alpha_t}}}^*), F_{\theta_t}(w_t, Y_t) - \bar{F}_{\theta_t}(w_t) \right\rangle \\
\leq& \left\| \nabla M(w_t - w_{\theta_t}^*) - \nabla M(w_{t-\tau_{\alpha_t}} - w_{\theta_{t-\tau_{\alpha_t}}}^*) \right\|_s^* \left\| F_{\theta_t}(w_t, Y_t) - \bar{F}_{\theta_t}(w_t) \right\|_s.
\end{aligned}$$

To bound the first term,

$$\begin{aligned}
& \left\| \nabla M(w_t - w_{\theta_t}^*) - \nabla M(w_{t-\tau_{\alpha_t}} - w_{\theta_{t-\tau_{\alpha_t}}}^*) \right\|_s^* \\
\leq& \frac{L}{\xi} \left\| w_t - w_{t-\tau_{\alpha_t}} + w_{\theta_{t-\tau_{\alpha_t}}}^* - w_{\theta_t}^* \right\|_s \quad \text{(Lemmas 10 and 45)} \\
\leq& \frac{L}{\xi} \left\| w_t - w_{t-\tau_{\alpha_t}} \right\|_s + \frac{L}{\xi} \left\| w_{\theta_t}^* - w_{\theta_{t-\tau_{\alpha_t}}}^* \right\|_s \\
\leq& \frac{L}{\xi l_{cs}} \left\| w_t - w_{t-\tau_{\alpha_t}} \right\|_c + \frac{L}{\xi l_{cs}} L_w L_\theta \beta_{t-\tau_{\alpha_t},t-1} \\
\leq& \frac{4L\alpha_{t-\tau_{\alpha_t},t-1}}{\xi l_{cs}} (A\|w_t\|_c + B) + \frac{L}{\xi l_{cs}} L_w L_\theta \beta_{t-\tau_{\alpha_t},t-1} \quad \text{(Lemma 47)} \\
\leq& \frac{4L\alpha_{t-\tau_{\alpha_t},t-1}}{\xi l_{cs}} (A\left\|w_t - w_{\theta_t}^*\right\|_c + A\left\|w_{\theta_t}^*\right\|_c + B) + \frac{L}{\xi l_{cs}} L_w L_\theta \beta_{t-\tau_{\alpha_t},t-1} \\
\leq& \frac{4L\alpha_{t-\tau_{\alpha_t},t-1}}{\xi l_{cs}} (A\left\|w_t - w_{\theta_t}^*\right\|_c + A\left\|w_{\theta_t}^*\right\|_c + B) + \frac{4L}{\xi l_{cs}} (L_w L_\theta + 1)\alpha_{t-\tau_{\alpha_t},t-1} \\
\leq& \frac{4L(L_w L_\theta + 1)\alpha_{t-\tau_{\alpha_t},t-1}}{\xi l_{cs}} (A\left\|w_t - w_{\theta_t}^*\right\|_c + AU_w + B + 1).
\end{aligned}$$

To bound the second term,

$$\begin{aligned}
& \left\| F_{\theta_t}(w_t, Y_t) - \bar{F}_{\theta_t}(w_t) \right\|_s \\
\leq& \frac{1}{l_{cs}} \left\| F_{\theta_t}(w_t, Y_t) - \bar{F}_{\theta_t}(w_t) \right\|_c \\
\leq& \frac{1}{l_{cs}} \left(\|F_{\theta_t}(w_t, Y_t)\|_c + \left\| \bar{F}_{\theta_t}(w_t) - \bar{F}_{\theta_t}(w_{\theta_t}^*) \right\|_c + \left\| w_{\theta_t}^* \right\|_c \right) \\
\leq& \frac{1}{l_{cs}} \left(U_F + L_F \|w_t\|_c + \left\| w_t - w_{\theta_t}^* \right\|_c + \left\| w_{\theta_t}^* \right\|_c \right) \quad \text{(Lemma 61)} \\
\leq& \frac{1}{l_{cs}} \left(U_F + L_F \left\| w_t - w_{\theta_t}^* \right\|_c + L_F \left\| w_{\theta_t}^* \right\|_c + \left\| w_t - w_{\theta_t}^* \right\|_c + \left\| w_{\theta_t}^* \right\|_c \right) \\
\leq& \frac{1}{l_{cs}} \left(A\left\| w_t - w_{\theta_t}^* \right\|_c + A\left\| w_{\theta_t}^* \right\|_c + B \right).
\end{aligned}$$

Combining the two inequalities together yields

$$\begin{aligned}
&\left\langle \nabla M(w_t - w^*_{\theta_t}) - \nabla M(w_{t-\tau_{\alpha_t}} - w^*_{\theta_t}), F_{\theta_t}(w_t, Y_t) - \bar{F}_{\theta_t}(w_t) \right\rangle \\
\leq &\frac{4L(L_w L_\theta + 1)\alpha_{t-\tau_{\alpha_t},t-1}}{\xi l_{cs}^2} (A\left\|w_t - w^*_{\theta_t}\right\|_c + C)^2 \\
\leq &\frac{8L(L_w L_\theta + 1)\alpha_{t-\tau_{\alpha_t},t-1}}{\xi l_{cs}^2} (A^2 u_{cm}^2 \left\|w_t - w^*_{\theta_t}\right\|_m^2 + C^2),
\end{aligned}$$

which completes the proof. ∎

E.4 Proof of Lemma 15

Lemma 54 *(Bound of T_{32})*

$$T_{32} \leq \frac{32L\alpha_{t-\tau_{\alpha_t},t-1}(1 + L_w L_\theta \beta_{t-\tau_{\alpha_t},t-1})}{\xi l_{cs}^2} \left(u_{cm}^2 A^2 \left\|w_t - w^*_{\theta_t}\right\|_m^2 + C^2 \right).$$

Proof

$$\begin{aligned}
T_{32} = &\left\langle \nabla M(w_{t-\tau_{\alpha_t}} - w^*_{\theta_{t-\tau_{\alpha_t}}}), F_{\theta_t}(w_t, Y_t) - F_{\theta_t}(w_{t-\tau_{\alpha_t}}, Y_t) + \bar{F}_{\theta_t}(w_{t-\tau_{\alpha_t}}) - \bar{F}_{\theta_t}(w_t) \right\rangle \\
\leq &\left\| \nabla M(w_{t-\tau_{\alpha_t}} - w^*_{\theta_{t-\tau_{\alpha_t}}}) \right\|_s^* \left\| F_{\theta_t}(w_t, Y_t) - F_{\theta_t}(w_{t-\tau_{\alpha_t}}, Y_t) + \bar{F}_{\theta_t}(w_{t-\tau_{\alpha_t}}) - \bar{F}_{\theta_t}(w_t) \right\|_s \\
\leq &\frac{1}{l_{cs}} \left\| \nabla M(w_{t-\tau_{\alpha_t}} - w^*_{\theta_{t-\tau_{\alpha_t}}}) \right\|_s^* \left\| F_{\theta_t}(w_t, Y_t) - F_{\theta_t}(w_{t-\tau_{\alpha_t}}, Y_t) + \bar{F}_{\theta_t}(w_{t-\tau_{\alpha_t}}) - \bar{F}_{\theta_t}(w_t) \right\|_c
\end{aligned}$$

For the first term,

$$\begin{aligned}
&\left\| \nabla M(w_{t-\tau_{\alpha_t}} - w^*_{\theta_{t-\tau_{\alpha_t}}}) \right\|_s^* \qquad (79)\\
= &\left\| \nabla M(w_{t-\tau_{\alpha_t}} - w^*_{\theta_{t-\tau_{\alpha_t}}}) - \nabla M(w^*_{\theta_t} - w^*_{\theta_t}) \right\|_s^* \\
&\qquad \text{(Using } \nabla M(0) = 0\text{, see the proof of Lemma 46)} \\
\leq &\frac{L}{\xi} \left\| (w_{t-\tau_{\alpha_t}} - w^*_{\theta_{t-\tau_{\alpha_t}}}) - (w^*_{\theta_t} - w^*_{\theta_t}) \right\|_s \quad \text{(Lemmas 10 and 45)} \\
\leq &\frac{L}{\xi} \left\| w_{t-\tau_{\alpha_t}} - w^*_{\theta_t} \right\|_s + \frac{L}{\xi} \left\| w^*_{\theta_t} - w^*_{\theta_{t-\tau_{\alpha_t}}} \right\|_s \\
\leq &\frac{L}{\xi l_{cs}} \left\| w_{t-\tau_{\alpha_t}} - w^*_{\theta_t} \right\|_c + \frac{L}{\xi l_{cs}} L_w L_\theta \beta_{t-\tau_{\alpha_t},t-1} \\
\leq &\frac{L}{\xi l_{cs}} \left(\left\| w_{t-\tau_{\alpha_t}} - w_t \right\|_c + \left\| w_t - w^*_{\theta_t} \right\|_c \right) + \frac{L}{\xi l_{cs}} L_w L_\theta \beta_{t-\tau_{\alpha_t},t-1} \\
\leq &\frac{L}{\xi l_{cs}} \left(\left\| w_t \right\|_c + \frac{B}{A} + \left\| w_t - w^*_{\theta_t} \right\|_c \right) + \frac{L}{\xi l_{cs}} L_w L_\theta \beta_{t-\tau_{\alpha_t},t-1} \quad \text{(Lemma 47)} \\
\leq &\frac{L(1 + L_w L_\theta \beta_{t-\tau_{\alpha_t},t-1})}{\xi l_{cs}} \left(\left\| w^*_{\theta_t} \right\|_c + \left\| w_t - w^*_{\theta_t} \right\|_c + \frac{B}{A} + \left\| w_t - w^*_{\theta_t} \right\|_c + 1 \right) \\
\leq &\frac{2L(1 + L_w L_\theta \beta_{t-\tau_{\alpha_t},t-1})}{\xi l_{cs}} \left(U_w + \frac{B}{A} + \left\| w_t - w^*_{\theta_t} \right\|_c + 1 \right).
\end{aligned}$$

For the second term,

$$
\begin{aligned}
&\left\|F_{\theta_t}(w_t, Y_t) - F_{\theta_t}(w_{t-\tau_{\alpha_t}}, Y_t) + \bar{F}_{\theta_t}(w_{t-\tau_{\alpha_t}}) - \bar{F}_{\theta_t}(w_t)\right\|_c \\
\leq &\left\|F_{\theta_t}(w_t, Y_t) - F_{\theta_t}(w_{t-\tau_{\alpha_t}}, Y_t)\right\|_c + \left\|\bar{F}_{\theta_t}(w_{t-\tau_{\alpha_t}}) - \bar{F}_{\theta_t}(w_t)\right\|_c \\
\leq &L_F\left\|w_{t-\tau_{\alpha_t}} - w_t\right\|_c + \left\|\sum_y d_{\theta_t}(y)\left(F_{\theta_t}(w_{t-\tau_{\alpha_t}}, y) - F_{\theta_t}(w_t, y)\right)\right\|_c \\
\leq &2L_F\left\|w_{t-\tau_{\alpha_t}} - w_t\right\|_c \\
\leq &2A\left\|w_{t-\tau_{\alpha_t}} - w_t\right\|_c \\
\leq &8A\alpha_{t-\tau_{\alpha_t},t-1}\left(A\|w_t\|_c + B\right) \quad \text{(Lemma 47)} \\
\leq &8A\alpha_{t-\tau_{\alpha_t},t-1}(A\left\|w_t - w^*_{\theta_t}\right\|_c + A\left\|w^*_{\theta_t}\right\|_c + B).
\end{aligned}
$$

Combining the two inequalities together yields

$$
\begin{aligned}
&\left\langle \nabla M(w_{t-\tau_{\alpha_t}} - w^*_{\theta_t}), F_{\theta_t}(w_t, Y_t) - F_{\theta_t}(w_{t-\tau_{\alpha_t}}, Y_t) + \bar{F}_{\theta_t}(w_{t-\tau_{\alpha_t}}) - \bar{F}_{\theta_t}(w_t)\right\rangle \\
\leq &\frac{16L\alpha_{t-\tau_{\alpha_t},t-1}(1 + L_w L_\theta \beta_{t-\tau_{\alpha_t},t-1})}{\xi l_{cs}^2}(A\left\|w_t - w^*_{\theta_t}\right\|_c + AU_w + B + A)^2 \\
\leq &\frac{32L\alpha_{t-\tau_{\alpha_t},t-1}(1 + L_w L_\theta \beta_{t-\tau_{\alpha_t},t-1})}{\xi l_{cs}^2}(u_{cm}^2 A^2\left\|w_t - w^*_{\theta_t}\right\|_m^2 + C^2)
\end{aligned}
$$

which completes the proof. ■

E.5 Proof of Lemma 16

Lemma 55 *(Bound of T_{331})*

$$
\mathbb{E}\left[T_{331}\right] \leq \frac{8L\alpha_t(1 + L_w L_\theta \beta_{t-\tau_{\alpha_t},t-1})}{A\xi l_{cs}^2}\left(u_{cm}^2 A^2 \mathbb{E}\left[\left\|w_t - w^*_{\theta_t}\right\|_m^2\right] + C^2\right).
$$

Proof

$$
\begin{aligned}
&\mathbb{E}\left[T_{331}\right] \qquad (80)\\
=&\mathbb{E}\left[\left\langle \nabla M(w_{t-\tau_{\alpha_t}} - w^*_{\theta_{t-\tau_{\alpha_t}}}), F_{\theta_{t-\tau_{\alpha_t}}}(w_{t-\tau_{\alpha_t}}, \tilde{Y}_t) - \bar{F}_{\theta_{t-\tau_{\alpha_t}}}(w_{t-\tau_{\alpha_t}})\right\rangle\right] \\
=&\mathbb{E}\left[\mathbb{E}\left[\left\langle \nabla M(w_{t-\tau_{\alpha_t}} - w^*_{\theta_{t-\tau_{\alpha_t}}}), F_{\theta_{t-\tau_{\alpha_t}}}(w_{t-\tau_{\alpha_t}}, \tilde{Y}_t) - \bar{F}_{\theta_{t-\tau_{\alpha_t}}}(w_{t-\tau_{\alpha_t}})\right\rangle \mid \substack{\theta_{t-\tau_{\alpha_t}} \\ w_{t-\tau_{\alpha_t}} \\ Y_{t-\tau_{\alpha_t}}}\right]\right] \\
=&\mathbb{E}\left[\left\langle \nabla M(w_{t-\tau_{\alpha_t}} - w^*_{\theta_{t-\tau_{\alpha_t}}}), \mathbb{E}\left[F_{\theta_{t-\tau_{\alpha_t}}}(w_{t-\tau_{\alpha_t}}, \tilde{Y}_t) - \bar{F}_{\theta_{t-\tau_{\alpha_t}}}(w_{t-\tau_{\alpha_t}}) \mid \substack{\theta_{t-\tau_{\alpha_t}} \\ w_{t-\tau_{\alpha_t}} \\ Y_{t-\tau_{\alpha_t}}}\right]\right\rangle\right] \\
\leq&\mathbb{E}\left[\left\|\nabla M(w_{t-\tau_{\alpha_t}} - w^*_{\theta_{t-\tau_{\alpha_t}}})\right\|_s^*\left\|\mathbb{E}\left[F_{\theta_{t-\tau_{\alpha_t}}}(w_{t-\tau_{\alpha_t}}, \tilde{Y}_t) - \bar{F}_{\theta_{t-\tau_{\alpha_t}}}(w_{t-\tau_{\alpha_t}}) \mid \substack{\theta_{t-\tau_{\alpha_t}} \\ w_{t-\tau_{\alpha_t}} \\ Y_{t-\tau_{\alpha_t}}}\right]\right\|_s\right] \\
\leq&\frac{1}{l_{cs}}\mathbb{E}\left[\left\|\nabla M(w_{t-\tau_{\alpha_t}} - w^*_{\theta_{t-\tau_{\alpha_t}}})\right\|_s^*\left\|\mathbb{E}\left[F_{\theta_{t-\tau_{\alpha_t}}}(w_{t-\tau_{\alpha_t}}, \tilde{Y}_t) - \bar{F}_{\theta_{t-\tau_{\alpha_t}}}(w_{t-\tau_{\alpha_t}}) \mid \substack{\theta_{t-\tau_{\alpha_t}} \\ w_{t-\tau_{\alpha_t}} \\ Y_{t-\tau_{\alpha_t}}}\right]\right\|_c\right]
\end{aligned}
$$

We now bound the inner expectation.

$$
\begin{aligned}
&\left\|\mathbb{E}\left[F_{\theta_{t-\tau_{\alpha_t}}}(w_{t-\tau_{\alpha_t}},\tilde{Y}_t)-\bar{F}_{\theta_{t-\tau_{\alpha_t}}}(w_{t-\tau_{\alpha_t}}) \mid {\scriptstyle \begin{matrix}\theta_{t-\tau_{\alpha_t}}\\ w_{t-\tau_{\alpha_t}}\\ Y_{t-\tau_{\alpha_t}}\end{matrix}}\right]\right\|_c \\
=&\left\|\sum_y\left(\Pr\left(\tilde{Y}_t=y \mid {\scriptstyle \begin{matrix}\theta_{t-\tau_{\alpha_t}}\\ w_{t-\tau_{\alpha_t}}\\ Y_{t-\tau_{\alpha_t}}\end{matrix}}\right)-d_{\theta_{t-\tau_{\alpha_t}}}(y)\right)F_{\theta_{t-\tau_{\alpha_t}}}(w_{t-\tau_{\alpha_t}},y)\right\|_c \\
\leq&\max_y\left\|F_{\theta_{t-\tau_{\alpha_t}}}(w_{t-\tau_{\alpha_t}},y)\right\|_c\sum_y\left|\Pr\left(\tilde{Y}_t=y \mid {\scriptstyle \begin{matrix}\theta_{t-\tau_{\alpha_t}}\\ w_{t-\tau_{\alpha_t}}\\ Y_{t-\tau_{\alpha_t}}\end{matrix}}\right)-d_{\theta_{t-\tau_{\alpha_t}}}(y)\right| \\
\leq&\max_y\left\|F_{\theta_{t-\tau_{\alpha_t}}}(w_{t-\tau_{\alpha_t}},y)\right\|_c\alpha_t \quad \text{(Definition of } \tau_{\alpha_t}) \qquad (81)\\
\leq&\alpha_t(U_F+L_F\left\|w_{t-\tau_{\alpha_t}}\right\|_c) \quad \text{(Lemma 61)} \\
\leq&\alpha_t(U_F+L_F\left\|w_{t-\tau_{\alpha_t}}-w_t\right\|_c+L_F\|w_t\|_c) \\
\leq&\alpha_t(B+A(\|w_t\|_c+\frac{B}{A})+A\|w_t\|_c) \quad \text{(Lemma 47)} \\
\leq&\alpha_t(2B+(A+1)\|w_t\|_c) \\
\leq&2\alpha_t(B+A\|w_t\|_c) \\
\leq&2\alpha_t(B+A\left\|w_t-w^*_{\theta_t}\right\|_c+A\left\|w^*_{\theta_t}\right\|_c).
\end{aligned}
$$

Using the above inequality and (79) yields

$$
\begin{aligned}
&\mathbb{E}\left[T_{331}\right] \\
\leq&\mathbb{E}\left[\frac{4L\alpha_t(1+L_wL_\theta\beta_{t-\tau_{\alpha_t},t-1})}{A\xi l_{cs}^2}\left(AU_w+B+A\left\|w_t-w^*_{\theta_t}\right\|_c+A\right)^2\right] \\
\leq&\mathbb{E}\left[\frac{8L\alpha_t(1+L_wL_\theta\beta_{t-\tau_{\alpha_t},t-1})}{A\xi l_{cs}^2}\left(A^2u_{cm}^2\left\|w_t-w^*_{\theta_t}\right\|_m^2+C^2\right)\right],
\end{aligned}
$$

which completes the proof. ■

E.6 Proof of Lemma 17

Lemma 56 *(Bound of T_{332})*

$$
\mathbb{E}\left[T_{332}\right]\leq\frac{8|\mathcal{Y}|L_PL_\theta\sum_{j=t-\tau_{\alpha_t}}^{t-1}\beta_{t-\tau_{\alpha_t},j}L(1+L_wL_\theta\beta_{t-\tau_{\alpha_t},t-1})}{A\xi l_{cs}^2}\left(u_{cm}^2A^2\mathbb{E}\left[\left\|w_t-w^*_{\theta_t}\right\|_m^2\right]+C^2\right).
$$

Proof

$$
\begin{aligned}
&\mathbb{E}\left[T_{332}\right] \\
=&\mathbb{E}\left[\left\langle\nabla M(w_{t-\tau_{\alpha_t}}-w^*_{\theta_{t-\tau_{\alpha_t}}}),F_{\theta_{t-\tau_{\alpha_t}}}(w_{t-\tau_{\alpha_t}},Y_t)-F_{\theta_{t-\tau_{\alpha_t}}}(w_{t-\tau_{\alpha_t}},\tilde{Y}_t)\right\rangle\right] \\
\leq&\frac{1}{l_{cs}}\mathbb{E}\left[\left\|\nabla M(w_{t-\tau_{\alpha_t}}-w^*_{\theta_{t-\tau_{\alpha_t}}})\right\|_s^*\left\|\mathbb{E}\left[F_{\theta_{t-\tau_{\alpha_t}}}(w_{t-\tau_{\alpha_t}},Y_t)-F_{\theta_{t-\tau_{\alpha_t}}}(w_{t-\tau_{\alpha_t}},\tilde{Y}_t) \mid {\scriptstyle \begin{matrix}w_{t-\tau_{\alpha_t}}\\ \theta_{t-\tau_{\alpha_t}}\\ Y_{t-\tau_{\alpha_t}}\end{matrix}}\right]\right\|_s\right]
\end{aligned}
$$

(Similar to (80))

$$\leq \mathbb{E}\left[\frac{2L(1+L_wL_\theta\beta_{t-\tau_{\alpha_t},t-1})}{\xi l_{cs}^2}\left(\left\|w^*_{\theta_t}\right\|_c+\frac{B}{A}+\left\|w_t-w^*_{\theta_t}\right\|_c+1\right)\right.$$

$$\left.\times\, 2|\mathcal{Y}|L_PL_\theta\sum_{j=t-\tau_{\alpha_t}}^{t-1}\beta_{t-\tau_{\alpha_t},j}(B+A\left\|w_t-w^*_{\theta_t}\right\|_c+A\left\|w^*_{\theta_t}\right\|_c)\right]$$

(Using (79) and Lemma 62)

$$\leq \frac{8|\mathcal{Y}|L_PL_\theta\sum_{j=t-\tau_{\alpha_t}}^{t-1}\beta_{t-\tau_{\alpha_t},j}L(1+L_wL_\theta\beta_{t-\tau_{\alpha_t},t-1})}{A\xi l_{cs}^2}\left(u_{cm}^2A^2\mathbb{E}\left[\left\|w_t-w^*_{\theta_t}\right\|_m^2\right]+C^2\right),$$

which completes the proof. ■

E.7 Proof of Lemma 18

Lemma 57 *(Bound of T_{333})*

$$T_{333}\leq\frac{8LL'_FL_\theta\beta_{t-\tau_{\alpha_t},t-1}(1+L_wL_\theta\beta_{t-\tau_{\alpha_t},t-1})}{A^2\xi l_{cs}^2}\left(u_{cm}^2A^2\left\|w_t-w^*_{\theta_t}\right\|_m^2+C^2\right).$$

Proof

$$\begin{aligned}T_{333}=&\left\langle\nabla M(w_{t-\tau_{\alpha_t}}-w^*_{\theta_{t-\tau_{\alpha_t}}}),F_{\theta_t}(w_{t-\tau_{\alpha_t}},Y_t)-F_{\theta_{t-\tau_{\alpha_t}}}(w_{t-\tau_{\alpha_t}},Y_t)\right\rangle\\ \leq&\left\|\nabla M(w_{t-\tau_{\alpha_t}}-w^*_{\theta_{t-\tau_{\alpha_t}}})\right\|_s^*\left\|F_{\theta_t}(w_{t-\tau_{\alpha_t}},Y_t)-F_{\theta_{t-\tau_{\alpha_t}}}(w_{t-\tau_{\alpha_t}},Y_t)\right\|_s\\ \leq&\frac{2L(1+L_wL_\theta\beta_{t-\tau_{\alpha_t},t-1})}{\xi l_{cs}^2}\left(\left\|w^*_{\theta_t}\right\|_c+\frac{B}{A}+\left\|w_t-w^*_{\theta_t}\right\|_c+1\right)\\ &\times L'_FL_\theta\beta_{t-\tau_{\alpha_t},t-1}\left(\left\|w_{t-\tau_{\alpha_t}}\right\|_c+U'_F\right)\quad\text{(Using (79) and Assumption 3.4)}\quad.\end{aligned}$$

Since

$$\begin{aligned}&\left\|w_{t-\tau_{\alpha_t}}\right\|_c\\ \leq&\left\|w_{t-\tau_{\alpha_t}}-w_t\right\|_c+\left\|w_t\right\|_c\\ \leq&2\left\|w_t\right\|_c+\frac{B}{A}\quad\text{(Lemma 47)}\\ \leq&2\left\|w_t-w^*_{\theta_t}\right\|_c+2\left\|w^*_{\theta_t}\right\|_c+\frac{B}{A},\end{aligned}\tag{82}$$

we have

$$T_{333}\leq\frac{8LL'_FL_\theta\beta_{t-\tau_{\alpha_t},t-1}(1+L_wL_\theta\beta_{t-\tau_{\alpha_t},t-1})}{A^2\xi l_{cs}^2}\left(u_{cm}^2A^2\left\|w_t-w^*_{\theta_t}\right\|_m^2+(AU_x+A+B+AU'_F)^2\right),$$

which completes the proof. ■

E.8 Proof of Lemma 19

Lemma 58 *(Bound of T_{334})*

$$T_{334} \leq \frac{8LL_F'' L_\theta \beta_{t-\tau_{\alpha_t},t-1}(1+L_w L_\theta \beta_{t-\tau_{\alpha_t},t-1})}{A^2 \xi l_{cs}^2} \left(u_{cm}^2 A^2 \left\| w_t - w_{\theta_t}^* \right\|_m^2 + C^2 \right).$$

Proof

$$\begin{aligned} T_{334} =& \left\langle \nabla M(w_{t-\tau_{\alpha_t}} - w^*_{\theta_{t-\tau_{\alpha_t}}}), \bar{F}_{\theta_{t-\tau_{\alpha_t}}}(w_{t-\tau_{\alpha_t}}) - \bar{F}_{\theta_t}(w_{t-\tau_{\alpha_t}}) \right\rangle \\ \leq & \left\| \nabla M(w_{t-\tau_{\alpha_t}} - w^*_{\theta_{t-\tau_{\alpha_t}}}) \right\|_s^* \left\| \bar{F}_{\theta_t}(w_{t-\tau_{\alpha_t}}) - \bar{F}_{\theta_{t-\tau_{\alpha_t}}}(w_{t-\tau_{\alpha_t}}) \right\|_s \\ \leq & \frac{2L(1+L_w L_\theta \beta_{t-\tau_{\alpha_t},t-1})}{\xi l_{cs}^2} \left(\left\| w^*_{\theta_t} \right\|_c + \frac{B}{A} + \left\| w_t - w^*_{\theta_t} \right\|_c + 1 \right) \\ & \times L_F'' L_\theta \beta_{t-\tau_{\alpha_t},t-1} \left(\left\| w_{t-\tau_{\alpha_t}} \right\|_c + U_F'' \right) \quad \text{(Using (79) and Assumption 3.4)} \quad . \end{aligned}$$

Using (82) completes the proof. ∎

E.9 Proof of Lemma 20

Lemma 59 *(Bound of T_4)*

$$\mathbb{E}\left[T_4\right] = 0.$$

Proof

$$\begin{aligned} & \mathbb{E}\left[T_4\right] \\ =& \mathbb{E}\left[\left\langle \nabla M(w_t - w^*_{\theta_t}), \epsilon_t \right\rangle\right] \\ =& \mathbb{E}\left[\mathbb{E}\left[\left\langle \nabla M(w_t - w^*_{\theta_t}), \epsilon_t \right\rangle \mid \mathcal{F}_t\right]\right] \quad \text{(Tower law of expectation)} \\ =& \mathbb{E}\left[\left\langle \nabla M(w_t - w^*_{\theta_t}), \mathbb{E}\left[\epsilon_t \mid \mathcal{F}_t\right] \right\rangle\right] \quad \text{(Conditional independence)} \\ =& 0 \quad \text{(Assumption 3.5)} \end{aligned}$$

∎

E.10 Proof of Lemma 21

Lemma 60 *(Bound of T_5)*

$$T_5 \leq \frac{2L}{\xi l_{cs}^2} \left(A^2 u_{cm}^2 \left\| w_t - w^*_{\theta_t} \right\|_m^2 + C^2 \right).$$

Proof

$$
\begin{aligned}
T_5 =& \frac{L}{\xi} \|F_{\theta_t}(w_t, Y_t) - w_t + \epsilon_t\|_s^2 \\
\leq& \frac{L}{\xi l_{cs}^2} \|F_{\theta_t}(w_t, Y_t) - w_t + \epsilon_t\|_c^2 \\
\leq& \frac{L^2}{\xi l_{cs}} \left(\|F_{\theta_t}(w_t, Y_t)\|_c + \|w_t\|_c + \|\epsilon_t\|_c\right)^2 \\
\leq& \frac{L}{\xi l_{cs}^2} \left(U_F + (L_F + 1)\|w_t\|_c + U_\epsilon \|w_t\|_c + U'_\epsilon\right)^2 \quad \text{(Lemma 61 and Assumption 3.5)} \\
\leq& \frac{L}{\xi l_{cs}^2} \left(B + A\|w_t\|_c\right)^2 \\
\leq& \frac{L}{\xi l_{cs}^2} \left(B + A\left\|w_t - w_{\theta_t}^*\right\|_c + A\left\|w_{\theta_t}^*\right\|_c\right)^2 \\
\leq& \frac{2L}{\xi l_{cs}^2} \left(A^2 u_{cm}^2 \left\|w_t - w_{\theta_t}^*\right\|_m^2 + C^2\right)
\end{aligned}
$$

■

Lemma 61 *For any time step t, almost surely,*

$$
\|F_{\theta_t}(w, y)\|_c \leq U_F + L_F \|w\|_c
$$

Proof Assumption 3.4 implies that

$$
\begin{aligned}
\|F_{\theta_t}(w, y)\|_c - \|F_{\theta_t}(0, y)\|_c &\leq \|F_{\theta_t}(0, y) - F_{\theta_t}(w, y)\|_c \\
&\leq L_F \|w - 0\|_c,
\end{aligned}
$$

which completes the proof. ■

Lemma 62

$$
\begin{aligned}
&\left\| \mathbb{E}\left[F_{\theta_{t-\tau_{\alpha_t}}}(w_{t-\tau_{\alpha_t}}, Y_t) - F_{\theta_{t-\tau_{\alpha_t}}}(w_{t-\tau_{\alpha_t}}, \tilde{Y}_t) \mid \begin{smallmatrix} w_{t-\tau_{\alpha_t}} \\ \theta_{t-\tau_{\alpha_t}} \\ Y_{t-\tau_{\alpha_t}} \end{smallmatrix} \right] \right\| \\
\leq& 2|\mathcal{Y}| L_P L_\theta \sum_{j=t-\tau_{\alpha_t}}^{t-1} \beta_{t-\tau_{\alpha_t}, j} (B + A\left\|w_t - w_{\theta_t}^*\right\|_c + A\left\|w_{\theta_t}^*\right\|_c)
\end{aligned}
$$

Proof In this proof, all Pr and $\mathbb{E}$ are implicitly conditioned on $w_{t-\tau_{\alpha_t}}, \theta_{t-\tau_{\alpha_t}}, Y_{t-\tau_{\alpha_t}}$. We use Θ_t to denote the set of all possible θ_t.

$$\begin{aligned}
&\Pr\big(Y_t = y'\big) \\
=&\sum_y \int_{\Theta_t} \Pr\big(Y_t = y', Y_{t-1} = y, \theta_t = z\big) dz \\
=&\sum_y \int_{\Theta_t} \Pr\big(Y_t = y' \mid Y_{t-1} = y, \theta_t = z\big) \Pr(Y_{t-1} = y, \theta_t = z) dz \\
=&\sum_y \int_{\Theta_t} P_z(y, y') \Pr(Y_{t-1} = y) \Pr(\theta_t = z | Y_{t-1} = y) dz
\end{aligned}$$

$$\begin{aligned}
&\Pr\Big(\tilde{Y}_t = y'\Big) \\
=&\sum_y \Pr\Big(\tilde{Y}_{t-1} = y\Big) P_{\theta_{t-\tau_{\alpha_t}}}(y, y') \\
=&\sum_y \Pr\Big(\tilde{Y}_{t-1} = y\Big) P_{\theta_{t-\tau_{\alpha_t}}}(y, y') \int_{\Theta_t} \Pr(\theta_t = z | Y_{t-1} = y) dz \\
=&\sum_y \int_{\Theta_t} \Pr\Big(\tilde{Y}_{t-1} = y\Big) P_{\theta_{t-\tau_{\alpha_t}}}(y, y') \Pr(\theta_t = z | Y_{t-1} = y) dz
\end{aligned}$$

Consequently,

$$\begin{aligned}
&\sum_{y'} \Big|\Pr\big(Y_t = y'\big) - \Pr\Big(\tilde{Y}_t = y'\Big)\Big| \\
\leq& \sum_{y, y'} \int_{\Theta_t} \Big|\Pr(Y_{t-1} = y) P_z(y, y') - \Pr\Big(\tilde{Y}_{t-1} = y\Big) P_{\theta_{t-\tau_{\alpha_t}}}(y, y')\Big| \Pr(\theta_t = z \mid Y_{t-1} = y) dz.
\end{aligned}$$

Since for any $z \in \Theta_t$,

$$\begin{aligned}
&\Big|\Pr(Y_{t-1} = y) P_z(y, y') - \Pr\Big(\tilde{Y}_{t-1} = y\Big) P_{\theta_{t-\tau_{\alpha_t}}}(y, y')\Big| \\
\leq& \Big|\Pr(Y_{t-1} = y) P_z(y, y') - \Pr\Big(\tilde{Y}_{t-1} = y\Big) P_z(y, y')\Big| \\
&+ \Big|\Pr\Big(\tilde{Y}_{t-1} = y\Big) P_z(y, y') - \Pr\Big(\tilde{Y}_{t-1} = y\Big) P_{\theta_{t-\tau_{\alpha_t}}}(y, y')\Big| \\
\leq& \Big|\Pr(Y_{t-1} = y) - \Pr\Big(\tilde{Y}_{t-1} = y\Big)\Big| P_z(y, y') + L_P L_\theta \beta_{t-\tau_{\alpha_t}, t-1} \Pr\Big(\tilde{Y}_{t-1} = y\Big),
\end{aligned}$$

we have

$$\begin{aligned}
&\sum_{y'} \Big|\Pr\big(Y_t = y'\big) - \Pr\Big(\tilde{Y}_t = y'\Big)\Big| \\
\leq& \sum_y \Big|\Pr(Y_{t-1} = y) - \Pr\Big(\tilde{Y}_{t-1} = y\Big)\Big| + |\mathcal{Y}| L_P L_\theta \beta_{t-\tau_{\alpha_t}, t-1}.
\end{aligned}$$

Applying the above inequality recursively yields

$$\sum_{y'} \left| \Pr\big(Y_t = y'\big) - \Pr\Big(\tilde{Y}_t = y'\Big) \right| \leq |\mathcal{Y}| L_P L_\theta \sum_{j=t-\tau_{\alpha_t}}^{t-1} \beta_{t-\tau_{\alpha_t}, j}. \tag{83}$$

Consequently,

$$\begin{aligned}
&\left\| \mathbb{E}\left[F_{\theta_{t-\tau_{\alpha_t}}}(w_{t-\tau_{\alpha_t}}, Y_t) - F_{\theta_{t-\tau_{\alpha_t}}}(w_{t-\tau_{\alpha_t}}, \tilde{Y}_t) \right] \right\|_c \\
=& \left\| \sum_y \left(\Pr(Y_t = y) - \Pr\Big(\tilde{Y}_t = y\Big) \right) F_{\theta_{t-\tau_{\alpha_t}}}(w_{t-\tau_{\alpha_t}}, y) \right\|_c \\
\leq& \max_y \left\| F_{\theta_{t-\tau_{\alpha_t}}}(w_{t-\tau_{\alpha_t}}, y) \right\|_c |\mathcal{Y}| L_P L_\theta \sum_{j=t-\tau_{\alpha_t}}^{t-1} \beta_{t-\tau_{\alpha_t}, j} \\
\leq& 2|\mathcal{Y}| L_P L_\theta \sum_{j=t-\tau_{\alpha_t}}^{t-1} \beta_{t-\tau_{\alpha_t}, j} (B + A\left\| w_t - w^*_{\theta_t} \right\|_c + A\left\| w^*_{\theta_t} \right\|_c) \quad \text{(Using (81))},
\end{aligned}$$

which completes the proof. ■

E.11 Proof of Lemma 11

Lemma 63 *For sufficiently large* t_0*,*

$$\tau_{\alpha_t} = \mathcal{O}(\log(t+t_0)), \quad \alpha_{t-\tau_{\alpha_t}, t-1} = \mathcal{O}\left(\frac{\log(t+t_0)}{(t+t_0)^{\epsilon_\alpha}} \right),$$
$$\beta_{t-\tau_{\alpha_t}, t-1} = \mathcal{O}\left(\frac{\log(t+t_0)}{(t+t_0)^{\epsilon_\alpha}} \right), \quad \frac{\alpha_t \alpha_{t-\tau_{\alpha_t}, t-1}}{\beta_t} = \mathcal{O}\left(\frac{\log(t+t_0)}{(t+t_0)^{2\epsilon_\alpha - \epsilon_\beta}} \right).$$

Proof By the definition of τ_{α_t} in (25), it is easy to see

$$\tau_{\alpha_t} = \left\lceil \frac{\log \alpha_t - \log C_0}{\log \tau} \right\rceil = \mathcal{O}(\log(t+t_0)),$$

where $\lceil \cdot \rceil$ is the ceiling function. Consequently,

$$\alpha_{t-\tau_{\alpha_t}, t-1} \leq \tau_{\alpha_t} \alpha_{t-\tau_{\alpha_t}} = \mathcal{O}\left(\frac{\log(t+t_0)}{(t+t_0-\tau_{\alpha_t})^{\epsilon_\alpha}} \right) = \mathcal{O}\left(\frac{\log(t+t_0)}{(t+t_0)^{\epsilon_\alpha}} \right),$$

implying

$$\frac{\alpha_t \alpha_{t-\tau_{\alpha_t}, t-1}}{\beta_t} = \mathcal{O}\left(\frac{\log(t+t_0)}{(t+t_0)^{2\epsilon_\alpha - \epsilon_\beta}} \right).$$

Assumption 3.6 ensures $\beta_t < \alpha_t$ holds for all t. Consequently,

$$\beta_{t-\tau_{\alpha_t}, t-1} < \alpha_{t-\tau_{\alpha_t}, t-1} = \mathcal{O}\left(\frac{\log(t+t_0)}{(t+t_0)^{\epsilon_\alpha}} \right),$$

which completes the proof. ■

E.12 Proof of Lemma 23

Lemma 64 *There exists a constant C_{t_0,w_0} such that for all $t \le t_0$,*

$$\mathbb{E}\left[\left\|w_t - w^*_{\theta_t}\right\|_m^2\right] \le C_{t_0,w_0}.$$

Proof According to (9), we have

$$\begin{aligned}\|w_{t+1}\|_c \le & \|w_t\|_c + \alpha_t\left(\|F_{\theta_t}(w_t, Y_t)\|_c + \|w_t\|_c + \|\epsilon_t\|_c\right)\\ \le & \|w_t\|_c + \alpha_t\left(U_F + L_F\|w_t\|_c + \|w_t\|_c + U_\epsilon\|w_t\|_c + U'_\epsilon\right)\\ & \text{(Lemma 61 and Assumption 3.5).}\end{aligned}$$

Consequently, it is easy to see that there exists a constant C_{t_0,w_0} such that for all $t \le t_0$,

$$\mathbb{E}\left[\left\|w_t - w^*_{\theta_t}\right\|_m^2\right] \le C_{t_0,w_0}.$$

■

E.13 Proof of Lemma 27

Lemma 65 *(Bound of M_{11}) There exists a constant $\chi_{11} > 0$ such that,*

$$M_{11} \ge \chi_{11}\|\nabla J_{\lambda_t}(\theta_t)\|^2.$$

Proof

$$\begin{aligned}
& M_{11}\\
=& \sum_{s,a}\left(d_{\mu_{\theta_t}}(s)\pi_{\theta_t}(a|s)\nabla\log\pi_{\theta_t}(a|s)q_{\pi_{\theta_t}}(s,a) + \frac{\lambda_t}{|\mathcal{A}|}d_{\mu_{\theta_t}}(s)\nabla\log\pi_{\theta_t}(a|s)\right)^\top \nabla J_{\lambda_t}(\theta_t)\\
=& \sum_{s',a'}\sum_{s,a}\left(d_{\mu_{\theta_t}}(s)\frac{\mathrm{d}\pi_{\theta_t}(a|s)}{\mathrm{d}\theta_{s',a'}}q_{\pi_{\theta_t}}(s,a) + \frac{\lambda_t}{|\mathcal{A}|}d_{\mu_{\theta_t}}(s)\frac{\mathrm{d}\log\pi_{\theta_t}(a|s)}{\mathrm{d}\theta_{s',a'}}\right)\frac{\mathrm{d}J_{\lambda_t}(\theta_t)}{\mathrm{d}\theta_{s',a'}}\\
=& \sum_{s',a'}\left(d_{\mu_{\theta_t}}(s')\pi_{\theta_t}(a'|s')\mathrm{Adv}_{\pi_{\theta_t}}(s',a') + \frac{\lambda_t}{|\mathcal{A}|}d_{\mu_{\theta_t}}(s')(1 - |\mathcal{A}|\pi_{\theta_t}(a'|s'))\right)\frac{\mathrm{d}J_{\lambda_t}(\theta_t)}{\mathrm{d}\theta_{s',a'}}\\
& \text{(Lemma 50)}\\
=& \sum_{s,a}\left(d_{\mu_{\theta_t}}(s)\pi_{\theta_t}(a|s)\mathrm{Adv}_{\pi_{\theta_t}}(s,a) + \lambda_t d_{\mu_{\theta_t}}(s)(\frac{1}{|\mathcal{A}|} - \pi_{\theta_t}(a|s))\right)\\
& \times\left(\underbrace{\frac{1}{1-\gamma}d_{\pi_{\theta_t},\gamma}(s)\pi_{\theta_t}(a|s)\mathrm{Adv}_{\pi_{\theta_t}}(s,a)}_{M_{111}} + \underbrace{\frac{\lambda_t}{|\mathcal{S}|}(\frac{1}{|\mathcal{A}|} - \pi_{\theta_t}(a|s))}_{M_{112}}\right)
\end{aligned}$$

(Lemma 50)

$$
\begin{aligned}
&= \sum_{s,a} \left(\frac{d_{\mu_{\theta_t}}(s)(1-\gamma)}{d_{\pi_{\theta_t},\gamma}(s)} M_{111} + d_{\mu_{\theta_t}}(s)|\mathcal{S}|M_{112} \right)(M_{111} + M_{112}) \\
&= \sum_{s,a} \frac{d_{\mu_{\theta_t}}(s)(1-\gamma)}{d_{\pi_{\theta_t},\gamma}(s)} M_{111}^2 + d_{\mu_{\theta_t}}(s)|\mathcal{S}|M_{112}^2 + \left(\frac{d_{\mu_{\theta_t}}(s)(1-\gamma)}{d_{\pi_{\theta_t},\gamma}(s)} + d_{\mu_{\theta_t}}(s)|\mathcal{S}| \right) M_{111}M_{112} \\
&\geq \sum_{s,a} \underbrace{\chi_{11} M_{111}^2 + \chi_{12} M_{112}^2 + \left(\frac{d_{\mu_{\theta_t}}(s)(1-\gamma)}{d_{\pi_{\theta_t},\gamma}(s)} + d_{\mu_{\theta_t}}(s)|\mathcal{S}| \right) M_{111}M_{112}}_{M_{113}}
\end{aligned}
$$

where

$$
\begin{aligned}
\chi_{11} &\doteq \inf_{\theta,s} \frac{d_{\mu_\theta}(s)(1-\gamma)}{d_{\pi_\theta,\gamma}(s)}, \\
\chi_{12} &\doteq \inf_{\theta,s} d_{\mu_\theta}(s)|\mathcal{S}|.
\end{aligned}
$$

Assumption 4.4, the continuity of d_{μ_θ} w.r.t. θ (Lemma 48), and the extreme value theorem ensures that

$$
\chi_{11} > 0, \, \chi_{12} > 0.
$$

If $M_{111}M_{112} < 0$, then

$$
M_{113} \geq \chi_{11} M_{111}^2 + \chi_{12} M_{112}^2 \geq \frac{\min\{\chi_{11}, \chi_{12}\}}{2} (M_{111} + M_{112})^2.
$$

If $M_{111}M_{112} \geq 0$, then

$$
M_{113} \geq \chi_{11} M_{111}^2 + \chi_{12} M_{112}^2 + (\chi_{11} + \chi_{12}) M_{111}M_{112} \geq \min\{\chi_{11}, \chi_{12}\} (M_{111} + M_{112})^2
$$

Let

$$
\chi_1 \doteq \frac{\min\{\chi_{11}, \chi_{12}\}}{2} > 0,
$$

then we always have

$$
M_{113} \geq \chi_1 (M_{111} + M_{112})^2,
$$

implying

$$
\begin{aligned}
M_{11} &\geq \chi_1 \sum_{s,a} (M_{111} + M_{112})^2 \\
&= \chi_1 \|\nabla J_{\lambda_t}(\theta_t)\|^2 \quad \text{(Lemma 50)} \quad ,
\end{aligned}
$$

which completes the proof. ∎

E.14 Proof of Lemma 28

Lemma 66 *(Bound of M_{121}) There exist constants $L^*_{\Lambda'} > 0$ such that*

$$\|M_{121}\| \leq \frac{L^*_{\Lambda'} L_\theta}{l_{2,p}} \beta_{t-\tau_{\beta_t},t-1}.$$

Proof We first study the Lipschitz continuity of $\Lambda(\theta, y, \eta)$ defined in (48). As shown in the verification of Assumption 3.4 (v) in Section B.2, q_{π_θ} is Lipschitz continuous in θ and bounded. According to Lemma 50, it is easy to see $\nabla \log \pi_\theta(a|s)$ is also Lipschitz continuous in θ and bounded. Assumption 4.4 ensures that $\inf_{\theta,a,s} \mu_\theta(a|s) > 0$, hence it is easy to see $\frac{\pi_\theta(a|s)}{\mu_\theta(a|s)}$ is also Lipschitz continuous and bounded from above. We, therefore, conclude via Lemma 44 that there exist continuous functions $L_\Lambda(\eta)$ and $U_\Lambda(\eta)$ such that for any y,

$$\begin{aligned}\left\|\Lambda(\theta, y, \eta) - \Lambda(\theta', y, \eta)\right\| &\leq L_\Lambda(\eta)\left\|\theta - \theta'\right\|,\\ \sup_\theta \|\Lambda(\theta, y, \eta)\| &\leq U_\Lambda(\eta).\end{aligned}$$

We now study the Lipschitz continuity of $\bar{\Lambda}(\theta, \eta)$ defined in (48). Lemma 48 confirms the Lipschitz continuity of d_{μ_θ}. Consequently, Lemma 44 implies that there exist continuous functions $L_{\bar{\Lambda}}(\eta)$ and $U_{\bar{\Lambda}}(\eta)$ such that

$$\begin{aligned}\left\|\bar{\Lambda}(\theta, \eta) - \bar{\Lambda}(\theta', \eta)\right\| &\leq L_{\bar{\Lambda}}(\eta)\left\|\theta - \theta'\right\|,\\ \sup_\theta \left\|\bar{\Lambda}(\theta, y, \eta)\right\| &\leq U_{\bar{\Lambda}}(\eta).\end{aligned}$$

We now study the Lipschitz continuity of $\Lambda'(\theta, y, \eta)$ defined in (49). Since $J_\eta(\theta)$ is $L_J + \eta L_{\mathrm{KL}}$ smooth, Lemma 45 implies that $L_J + \eta L_{\mathrm{KL}}$ is a Lipschitz constant of $\nabla J_\eta(\theta)$. From Lemma 50, it is easy to see the upper bound of $\nabla J_\eta(\theta)$ is also a continuous function of η. Consequently, Lemma 44 implies there exist continuous functions $L_{\Lambda'}(\eta)$ and $U_{\Lambda'}(\eta)$ such that for all y,

$$\begin{aligned}\left\|\Lambda'(\theta, y, \eta) - \Lambda'(\theta', y, \eta)\right\| &\leq L_{\Lambda'}(\eta)\left\|\theta - \theta'\right\|,\\ \sup_\theta \left\|\Lambda'(\theta, y, \eta)\right\| &\leq U_{\Lambda'}(\eta).\end{aligned} \tag{84}$$

Hence

$$\begin{aligned}\|M_{121}\| &= \left\|\Lambda'(\theta_t, Y_t, \lambda_t) - \Lambda'(\theta_{t-\tau_{\beta_t}}, Y_t, \lambda_t)\right\|\\ &\leq L_{\Lambda'}(\lambda_t)\left\|\theta_t - \theta_{t-\tau_{\beta_t}}\right\|\\ &\leq \frac{L_{\Lambda'}(\lambda_t) L_\theta}{l_{2,p}} \beta_{t-\tau_{\beta_t},t-1} \quad \text{(Using (43))} \quad .\end{aligned}$$

Since $\lambda_t \in [0, \lambda]$, $L_{\Lambda'}(\eta)$ is a continuous function and well defined in $[0, \lambda]$, the extreme value theorem asserts that $L_{\Lambda'}(\eta)$ obtains its maximum in $[0, \lambda]$, say, e.g., $L^*_{\Lambda'}$. Then

$$\|M_{121}\| \leq \frac{L^*_{\Lambda'} L_\theta}{l_{2,p}} \beta_{t-\tau_{\beta_t},t-1}.$$

■

E.15 Proof of Lemma 29

Lemma 67 *(Bound of M_{122}) There exists a constant $U^*_{\Lambda'} > 0$ such that*

$$\|\mathbb{E}\left[M_{122}\right]\| \leq U^*_{\Lambda'}|\mathcal{S}||\mathcal{A}|L_\mu L_\theta \sum_{j=t-\tau_{\beta_t}}^{t-1} \beta_{t-\tau_{\beta_t},j}.$$

Proof

$$\begin{aligned}
&\|\mathbb{E}\left[M_{122}\right]\| \\
=&\left\|\mathbb{E}\left[\mathbb{E}\left[M_{122} \mid \theta_{t-\tau_{\beta_t}}, Y_{t-\tau_{\beta_t}}\right]\right]\right\| \\
\leq&\mathbb{E}\left[\left\|\mathbb{E}\left[M_{122} \mid \theta_{t-\tau_{\beta_t}}, Y_{t-\tau_{\beta_t}}\right]\right\|\right].
\end{aligned}$$

We now bound the inner expectation. In the rest of the proof, all Pr and $\mathbb{E}$ are implicitly conditioned on $\theta_{t-\tau_{\beta_t}}$ and $Y_{t-\tau_{\beta_t}}$.

$$\begin{aligned}
&\|\mathbb{E}\left[M_{122}\right]\| \\
=&\left\|\mathbb{E}\left[\Lambda'(\theta_{t-\tau_{\beta_t}}, Y_t, \lambda_t) - \Lambda'(\theta_{t-\tau_{\beta_t}}, \tilde{Y}_t, \lambda_t)\right]\right\| \\
=&\left\|\sum_y \left(\Pr\left(\tilde{Y}_t = y\right) - \Pr(Y_t = y)\right)\Lambda'(\theta_{t-\tau_{\beta_t}}, y, \lambda_t)\right\| \\
\leq&\max_y \left\|\Lambda'(\theta_{t-\tau_{\beta_t}}, y, \lambda_t)\right\| \sum_y \left|\Pr\left(\tilde{Y}_t = y\right) - \Pr(Y_t = y)\right| \\
\leq&U_{\Lambda'}(\lambda_t)\sum_y \left|\Pr\left(\tilde{Y}_t = y\right) - \Pr(Y_t = y)\right| \quad \text{(Using (84))} \\
\leq&U_{\Lambda'}(\lambda_t)|\mathcal{S}||\mathcal{A}|L_\mu L_\theta \sum_{j=t-\tau_{\beta_t}}^{t-1} \beta_{t-\tau_{\beta_t},j} \quad \text{(Similar to (83) with } L_\theta \text{ defined in (43))} \quad .
\end{aligned}$$

Since $\lambda_t \in [0, \lambda]$ and the continuous function $U_{\Lambda'}(\eta)$ obtains its maximum, say, e.g., $U^*_{\Lambda'}$, in the compact set $[0, \lambda]$, we have

$$\|\mathbb{E}\left[M_{122}\right]\| \leq U^*_{\Lambda'}|\mathcal{S}||\mathcal{A}|L_\mu L_\theta \sum_{j=t-\tau_{\beta_t}}^{t-1} \beta_{t-\tau_{\beta_t},j},$$

which completes the proof. ∎

E.16 Proof of Lemma 30

Lemma 68 *(Bound of M_{123})*

$$\|\mathbb{E}\left[M_{123}\right]\| \leq U^*_{\Lambda'}\beta_t.$$

Proof

$$
\begin{aligned}
&\left\|\mathbb{E}\left[M_{123}\right]\right\| \\
=&\left\|\mathbb{E}\left[\Lambda'(\theta_{t-\tau_{\beta_t}}, \tilde{Y}_t, \lambda_t)\right]\right\| \\
=&\left\|\mathbb{E}\left[\mathbb{E}\left[\Lambda'(\theta_{t-\tau_{\beta_t}}, \tilde{Y}_t, \lambda_t) \mid \theta_{t-\tau_{\beta_t}}, Y_{t-\tau_{\beta_t}}\right]\right]\right\| \\
\leq&\mathbb{E}\left[\left\|\mathbb{E}\left[\Lambda'(\theta_{t-\tau_{\beta_t}}, \tilde{Y}_t, \lambda_t) \mid \theta_{t-\tau_{\beta_t}}, Y_{t-\tau_{\beta_t}}\right]\right\|\right].
\end{aligned}
$$

We now bound the inner expectation. In the rest of the proof, all Pr and $\mathbb{E}$ are implicitly conditioned on $\theta_{t-\tau_{\beta_t}}$ and $Y_{t-\tau_{\beta_t}}$. Since $\tilde{Y}_t = (\tilde{S}_t, \tilde{A}_t)$ and

$$
\sum_{s,a} d_{\mu_{\theta_{t-\tau_{\beta_t}}}}(s)\mu_{\theta_{t-\tau_{\beta_t}}}(a|s)\Lambda'(\theta_{t-\tau_{\beta_t}}, (s,a), \lambda_t) = 0,
$$

we have

$$
\begin{aligned}
&\left\|\mathbb{E}\left[\Lambda'(\theta_{t-\tau_{\beta_t}}, \tilde{Y}_t, \lambda_t)\right]\right\| \\
=&\left\|\sum_{s,a}\left(\Pr\left(\tilde{S}_t = s, \tilde{A}_t = a\right) - d_{\mu_{\theta_{t-\tau_{\beta_t}}}}(s)\mu_{\theta_{t-\tau_{\beta_t}}}(a|s)\right)\Lambda'(\theta_{t-\tau_{\beta_t}}, (s,a), \lambda_t)\right\| \\
\leq&\sup_{s,a,\theta}\left\|\Lambda'(\theta, (s,a), \lambda_t)\right\|\sum_{s,a}\left|\Pr\left(\tilde{S}_t = s, \tilde{A}_t = a\right) - d_{\mu_{\theta_{t-\tau_{\beta_t}}}}(s)\mu_{\theta_{t-\tau_{\beta_t}}}(a|s)\right| \\
\leq&U^*_{\Lambda'}\beta_t \quad \text{(Using (50))} \quad ,
\end{aligned}
$$

which completes the proof. ■

E.17 Proof of Lemma 31

Lemma 69 *(Bound of M_{13}) There exists a constant $\rho_{max} > 0$ such that*

$$
\left\|\mathbb{E}\left[M_{13}\right]\right\| \leq 2\rho_{max}\sqrt{|\mathcal{S}\times\mathcal{A}|}\sqrt{\mathbb{E}\left[\left\|q_t - q_{\pi_{\theta_t}}\right\|_\infty^2\right]}\sqrt{\mathbb{E}\left[\left\|\nabla J_{\lambda_t}(\theta_t)\right\|^2\right]}.
$$

Proof

$$
\begin{aligned}
&\left\|\mathbb{E}\left[M_{13}\right]\right\| \\
=&\left\|\mathbb{E}\left[\left\langle\nabla J_{\lambda_t}(\theta_t), \rho_t\nabla\log\pi_{\theta_t}(A_t|S_t)\left(\Pi\left(q_t(S_t,A_t)\right) - q_{\pi_{\theta_t}}(S_t,A_t)\right)\right\rangle\right]\right\| \\
\leq&\sum_{s,a}\mathbb{E}\left[\left|\frac{\mathrm{d}J_{\lambda_t}(\theta_t)}{\mathrm{d}\theta_{s,a}}\rho_t\frac{\mathrm{d}\log\pi_{\theta_t}(A_t|S_t)}{\mathrm{d}\theta_{s,a}}\left(\Pi\left(q_t(S_t,A_t)\right) - q_{\pi_{\theta_t}}(S_t,A_t)\right)\right|\right] \\
\leq&\sum_{s,a}\sqrt{\mathbb{E}\left[\left(\frac{\mathrm{d}J_{\lambda_t}(\theta_t)}{\mathrm{d}\theta_{s,a}}\right)^2\right]\mathbb{E}\left[\left(\rho_t\frac{\mathrm{d}\log\pi_{\theta_t}(A_t|S_t)}{\mathrm{d}\theta_{s,a}}\right)^2\left(\Pi\left(q_t(S_t,A_t)\right) - q_{\pi_{\theta_t}}(S_t,A_t)\right)^2\right]},
\end{aligned}
$$

(Cauchy-Schwarz inequality)

Lemma 50 implies that

$$\left|\frac{\mathrm{d}\log\pi_\theta(a|s)}{\mathrm{d}\theta_{s',a'}}\right| < 2.$$

Assumption 4.4 implies that

$$\rho_{max} \doteq \sup_{\theta,s,a}\frac{\pi_\theta(a|s)}{\mu_\theta(a|s)} < \infty.$$

Hence

$$\begin{aligned}
&\|\mathbb{E}\left[M_{13}\right]\| \\
\leq& 2\rho_{max}\sum_{s,a}\sqrt{\mathbb{E}\left[\left(\frac{\mathrm{d}J_{\lambda_t}(\theta_t)}{\mathrm{d}\theta_{s,a}}\right)^2\right]\mathbb{E}\left[\left(\Pi\left(q_t(S_t,A_t)\right) - q_{\pi_{\theta_t}}(S_t,A_t)\right)^2\right]} \\
=& 2\rho_{max}\sum_{s,a}\sqrt{\mathbb{E}\left[\left(\frac{\mathrm{d}J_{\lambda_t}(\theta_t)}{\mathrm{d}\theta_{s,a}}\right)^2\right]\mathbb{E}\left[\left(\Pi\left(q_t(S_t,A_t)\right) - \Pi\left(q_{\pi_{\theta_t}}(S_t,A_t)\right)\right)^2\right]} \\
\leq& 2\rho_{max}\sum_{s,a}\sqrt{\mathbb{E}\left[\left(\frac{\mathrm{d}J_{\lambda_t}(\theta_t)}{\mathrm{d}\theta_{s,a}}\right)^2\right]\mathbb{E}\left[\left(q_t(S_t,A_t) - q_{\pi_{\theta_t}}(S_t,A_t)\right)^2\right]}
\end{aligned}$$

(Projection to a convex set is nonexpansive)

$$\begin{aligned}
\leq& 2\rho_{max}\sum_{s,a}\sqrt{\mathbb{E}\left[\left(\frac{\mathrm{d}J_{\lambda_t}(\theta_t)}{\mathrm{d}\theta_{s,a}}\right)^2\right]\mathbb{E}\left[\left\|q_t - q_{\pi_{\theta_t}}\right\|_\infty^2\right]} \\
\leq& 2\rho_{max}\sqrt{\mathbb{E}\left[\left\|q_t - q_{\pi_{\theta_t}}\right\|_\infty^2\right]}\sum_{s,a}\left(\sqrt{\mathbb{E}\left[\left(\frac{\mathrm{d}J_{\lambda_t}(\theta_t)}{\mathrm{d}\theta_{s,a}}\right)^2\right]}\times 1\right) \\
\leq& 2\rho_{max}\sqrt{|\mathcal{S}\times\mathcal{A}|}\sqrt{\mathbb{E}\left[\left\|q_t - q_{\pi_{\theta_t}}\right\|_\infty^2\right]}\sqrt{\mathbb{E}\left[\|\nabla J_{\lambda_t}(\theta_t)\|^2\right]},
\end{aligned}$$

(Cauchy-Schwarz inequality)

which completes the proof. ■

E.18 Proof of Lemma 32

Lemma 70 *There exists a constant $U_{J,\lambda}$ such that for all t,*

$$|\mathbb{E}\left[J_{\lambda_t}(\theta_t)\right]| \leq U_{J,\lambda},\ |\mathbb{E}\left[J_{\lambda_t}(\theta_{t+1})\right]| \leq U_{J,\lambda}.$$

Proof Lemma 50 implies that

$$\left|\frac{\mathrm{d}J_{\lambda_t}(\theta_t)}{\mathrm{d}\theta_{s,a}}\right| \leq \left|\frac{1}{1-\gamma}d_{\pi_\theta,\gamma,p_0}(s)\pi_\theta(a|s)\mathrm{Adv}_{\pi_\theta}(s,a)\right| + \frac{\lambda_t}{|\mathcal{S}|}\left|\pi_\theta(a|s) - \frac{1}{|\mathcal{A}|}\right|.$$

Since $\lambda_t < \lambda$, we conclude that there exists a constant χ_6 (depending on λ) such that $\forall t, \theta$

$$\left\|\nabla J_{\lambda_t}(\theta)\right\|^2 \leq \chi_6.$$

Then (52) and Proposition 4 imply that there exists some constant $\chi_7 > 0$ such that

$$\begin{aligned}\mathbb{E}\left[J_{\lambda_t}(\theta_{t+1})\right] \geq& \mathbb{E}\left[J_{\lambda_t}(\theta_t)\right] + \beta_t\chi_{11}\mathbb{E}\left[\left\|\nabla J_{\lambda_t}(\theta_t)\right\|^2\right] \\ &- \underbrace{\left(\beta_t\chi_{12}\frac{\log^2(t+t_0)}{(t+t_0)^{\epsilon_\beta}} + \beta_t\chi_7 t^{-\frac{\epsilon_q}{2}}\sqrt{\chi_6} + \beta_t\chi_2\frac{1}{(t+t_0)^{\epsilon_\beta}}\right)}_{z_t}.\end{aligned}$$

Hence

$$\begin{aligned}&\mathbb{E}\left[J_{\lambda_{t+1}}(\theta_{t+1})\right] \\ \geq& \mathbb{E}\left[J_{\lambda_t}(\theta_t)\right] + \beta_t\chi_{11}\mathbb{E}\left[\left\|\nabla J_{\lambda_t}(\theta_t)\right\|^2\right] + \mathbb{E}\left[J_{\lambda_{t+1}}(\theta_{t+1})\right] - \mathbb{E}\left[J_{\lambda_t}(\theta_{t+1})\right] - z_t \\ =& \mathbb{E}\left[J_{\lambda_t}(\theta_t)\right] + \beta_t\chi_{11}\mathbb{E}\left[\left\|\nabla J_{\lambda_t}(\theta_t)\right\|^2\right] + (\lambda_t - \lambda_{t+1})\mathbb{E}_{s\sim\mathcal{U}_\mathcal{S}}\left[\mathrm{KL}\left(\mathcal{U}_\mathcal{A}||\pi_{\theta_{t+1}}(\cdot|s)\right)\right] - z_t \\ & \hfill \text{(Using (44))} \\ \geq& \mathbb{E}\left[J_{\lambda_t}(\theta_t)\right] - z_t \quad \text{(Using } \lambda_t > \lambda_{t+1}\text{)} \quad .\end{aligned}$$

Telescoping the above inequality yields

$$\mathbb{E}\left[J_{\lambda_t}(\theta_t)\right] \geq \mathbb{E}\left[J_{\lambda_0}(\theta_0)\right] - \sum_{k=0}^{t} z_k \geq \mathbb{E}\left[J_{\lambda_0}(\theta_0)\right] - \sum_{k=0}^{\infty} z_k.$$

Since $\epsilon_\beta > 0.5$, we have

$$\sum_{t=0}^{\infty}\beta_t\frac{\log^2(t+t_0)}{(t+t_0)^{\epsilon_\beta}} = \sum_{t=0}^{\infty}\frac{\beta\log^2(t+t_0)}{(t+t_0)^{2\epsilon_\beta}} < \infty,$$
$$\sum_{t=0}^{\infty}\beta_t\frac{1}{(t+t_0)^{\epsilon_\beta}} = \sum_{t=0}^{\infty}\frac{\beta}{(t+t_0)^{2\epsilon_\beta}} < \infty.$$

Since $\epsilon_q > 2(1-\epsilon_\beta)$, we have

$$\sum_{t=0}^{\infty}\beta_t t^{-\frac{\epsilon_q}{2}} < \sum_{t=0}\frac{\beta}{t^{\epsilon_\beta+\frac{\epsilon_q}{2}}} < \infty.$$

We, therefore, conclude that

$$\sum_{t=0}^{\infty} z_t < \infty,$$

implying $\mathbb{E}\left[J_{\lambda_{t+1}}(\theta_{t+1})\right]$ is bounded from the below by some constant. By (44),

$$\mathbb{E}\left[J_{\lambda_{t+1}}(\theta_{t+1})\right] \leq \frac{r_{max}}{1-\gamma},$$

we, therefore, conclude that $\left|\mathbb{E}\left[J_{\lambda_{t+1}}(\theta_{t+1})\right]\right|$ is bounded by some constant. Similarly, we have

$$\begin{aligned}
&\mathbb{E}\left[J_{\lambda_t}(\theta_{t+1})\right] \\
\geq&\mathbb{E}\left[J_{\lambda_{t-1}}(\theta_t)\right] + \beta_t \chi_{11} \mathbb{E}\left[\|\nabla J_{\lambda_t}(\theta_t)\|^2\right] + \mathbb{E}\left[J_{\lambda_t}(\theta_t)\right] - \mathbb{E}\left[J_{\lambda_{t-1}}(\theta_t)\right] - z_t \\
=&\mathbb{E}\left[J_{\lambda_{t-1}}(\theta_t)\right] + \beta_t \chi_{11} \mathbb{E}\left[\|\nabla J_{\lambda_t}(\theta_t)\|^2\right] + (\lambda_{t-1} - \lambda_t)\mathbb{E}_{s\sim\mathcal{U}_\mathcal{S}}\left[\mathrm{KL}\left(\mathcal{U}_\mathcal{A}||\pi_{\theta_t}(\cdot|s)\right)\right] - z_t \\
\geq&\mathbb{E}\left[J_{\lambda_{t-1}}(\theta_t)\right] - z_t \quad (\text{Using } \lambda_{t-1} > \lambda_t) \quad .
\end{aligned}$$

Hence $|\mathbb{E}\left[J_{\lambda_t}(\theta_{t+1})\right]|$ is also bounded, which completes the proof. ■

E.19 Proof of Lemma 33

Lemma 71

$$\mathbb{E}\left[\sum_{k=\lceil \frac{t}{2} \rceil}^{t} \frac{1}{\beta_k}\left(J_{\lambda_k}(\theta_{k+1}) - J_{\lambda_k}(\theta_k)\right)\right] \leq \frac{2U_{J,\lambda}}{\beta}(t+t_0)^{\epsilon_\beta}$$

Proof

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{k=\lceil\frac{t}{2}\rceil}^{t}\left(\frac{1}{\beta_k}J_{\lambda_k}(\theta_{k+1})-\frac{1}{\beta_k}J_{\lambda_k}(\theta_k)\right)\right]\\
=&\mathbb{E}\left[\sum_{k=\lceil\frac{t}{2}\rceil}^{t}\left(\frac{1}{\beta_{k-1}}J_{\lambda_{k-1}}(\theta_k)-\frac{1}{\beta_k}J_{\lambda_k}(\theta_k)\right)+\frac{1}{\beta_t}J_{\lambda_t}(\theta_{t+1})-\frac{1}{\beta_{\lceil\frac{t}{2}\rceil-1}}J_{\lambda_{\lceil\frac{t}{2}\rceil-1}}(\theta_{\lceil\frac{t}{2}\rceil})\right]\\
=&\mathbb{E}\left[\sum_{k=\lceil\frac{t}{2}\rceil}^{t}\left(\frac{1}{\beta_{k-1}}J_{\lambda_{k-1}}(\theta_k)-\frac{1}{\beta_{k-1}}J_{\lambda_k}(\theta_k)+\frac{1}{\beta_{k-1}}J_{\lambda_k}(\theta_k)-\frac{1}{\beta_k}J_{\lambda_k}(\theta_k)\right)\right]\\
&+\mathbb{E}\left[\frac{1}{\beta_t}J_{\lambda_t}(\theta_{t+1})-\frac{1}{\beta_{\lceil\frac{t}{2}\rceil-1}}J_{\lambda_{\lceil\frac{t}{2}\rceil-1}}(\theta_{\lceil\frac{t}{2}\rceil})\right]\\
=&\mathbb{E}\left[\sum_{k=\lceil\frac{t}{2}\rceil}^{t}\left(\frac{1}{\beta_{k-1}}(\lambda_k-\lambda_{k-1})\mathbb{E}_{s\sim\mathcal{U}_{\mathcal{S}}}\left[\mathrm{KL}\left(\mathcal{U}_{\mathcal{A}}||\pi_{\theta_k}(\cdot|s)\right)\right]+\left(\frac{1}{\beta_{k-1}}-\frac{1}{\beta_k}\right)J_{\lambda_k}(\theta_k)\right)\right]\\
&+\mathbb{E}\left[\frac{1}{\beta_t}J_{\lambda_t}(\theta_{t+1})-\frac{1}{\beta_{\lceil\frac{t}{2}\rceil-1}}J_{\lambda_{\lceil\frac{t}{2}\rceil-1}}(\theta_{\lceil\frac{t}{2}\rceil})\right]\\
\leq&\mathbb{E}\left[\sum_{k=\lceil\frac{t}{2}\rceil}^{t}\left(\frac{1}{\beta_{k-1}}-\frac{1}{\beta_k}\right)J_{\lambda_k}(\theta_k)+\frac{1}{\beta_t}J_{\lambda_t}(\theta_{t+1})-\frac{1}{\beta_{\lceil\frac{t}{2}\rceil-1}}J_{\lambda_{\lceil\frac{t}{2}\rceil-1}}(\theta_{\lceil\frac{t}{2}\rceil})\right]\quad(\text{Using }\lambda_k<\lambda_{k-1})\quad,
\end{aligned}
$$

$$
\begin{aligned}
\leq&\sum_{k=\lceil\frac{t}{2}\rceil}^{t}\left(\frac{1}{\beta_k}-\frac{1}{\beta_{k-1}}\right)U_{J,\lambda}+\frac{1}{\beta_t}U_{J,\lambda}+\frac{1}{\beta_{\lceil\frac{t}{2}\rceil-1}}U_{J,\lambda}\quad(\text{Using }\beta_{k-1}>\beta_k\text{ and Lemma 32})\\
=&\frac{U_{J,\lambda}}{\beta_t}-\frac{U_{J,\lambda}}{\beta_{\lceil\frac{t}{2}\rceil-1}}+\frac{1}{\beta_t}U_{J,\lambda}+\frac{1}{\beta_{\lceil\frac{t}{2}\rceil-1}}U_{J,\lambda}\\
=&\frac{2U_{J,\lambda}}{\beta}(t+t_0)^{\epsilon_\beta}
\end{aligned}
$$

∎

E.20 Proof of Lemma 36

Lemma 72 *(Bound of $\tilde{M}_{11}$) There exists a constant $\chi_{11}>0$ such that,*

$$\tilde{M}_{11}\geq\chi_{11}\left\|\nabla\tilde{J}_{\lambda_t}(\theta_t)\right\|^2.$$

Proof

$$
\begin{aligned}
&\tilde{M}_{11} \\
=&\sum_{s,a} \left(d_{\mu_{\theta_t}}(s)\pi_{\theta_t}(a|s)\nabla \log \pi_{\theta_t}(a|s) \left(q_{\pi_{\theta_t},\lambda_t}(s,a) - \lambda_t \log \pi_{\theta_t}(a|s) \right) \right)^\top \nabla \tilde{J}_{\lambda_t}(\theta_t) \\
=&\sum_{s',a'} \sum_{s,a} \left(d_{\mu_{\theta_t}}(s) \frac{\mathrm{d}\pi_{\theta_t}(a|s)}{\mathrm{d}\theta_{s',a'}} \left(q_{\pi_{\theta_t},\lambda_t}(s,a) - \lambda_t \log \pi_{\theta_t}(a|s) \right) \right) \frac{\mathrm{d}\tilde{J}_{\lambda_t}(\theta_t)}{\mathrm{d}\theta_{s',a'}} \\
=&\sum_{s',a'} \left(d_{\mu_{\theta_t}}(s')\pi_{\theta_t}(a'|s')\tilde{\mathrm{Adv}}_{\pi_{\theta_t}}(s',a') \right) \frac{\mathrm{d}\tilde{J}_{\lambda_t}(\theta_t)}{\mathrm{d}\theta_{s',a'}} \quad \text{(Lemma 50)} \\
=&\sum_{s',a'} \frac{(1-\gamma)d_{\mu_{\theta_t}}(s')}{d_{\pi_{\theta_t},\gamma}(s')} \left(\frac{\mathrm{d}\tilde{J}_{\lambda_t}(\theta_t)}{\mathrm{d}\theta_{s',a'}} \right)^2 \quad \text{(Lemma 50)} \\
\geq& \inf_{\theta,s} \frac{(1-\gamma)d_{\mu_\theta}(s)}{d_{\pi_\theta,\gamma}(s)} \left\| \nabla \tilde{J}_{\lambda_t}(\theta_t) \right\|^2.
\end{aligned}
$$

Assumption 4.4, the continuity of d_{μ_θ} w.r.t. θ (Lemma 48), and the extreme value theorem ensures that the above inf is strictly positive, which completes the proof. ∎

E.21 Proof of Lemma 37

Lemma 73 *(Bound of $\tilde{M}_{121}$) There exist constants $L^*_{\Lambda_1'} > 0$ such that*

$$
\left\| \tilde{M}_{121} \right\| \leq L^*_{\Lambda_1'} L_\theta \beta_{t-\tau_{\beta_t},t-1},
$$

where L_θ is defined in (59).

Proof We first study the Lipschitz continuity of $\Lambda_1(\theta, s, \eta)$ defined in (60). We have

$$
\begin{aligned}
&\Lambda_1(\theta,s,\eta) \\
=&\sum_a \pi_\theta(a|s)\nabla \log \pi_\theta(a|s) \left(\tilde{q}_{\pi_\theta,\eta}(s,a) - \eta \log \pi_\theta(a|s) \right) \\
=&\sum_a \nabla \pi_\theta(a|s)\tilde{q}_{\pi_\theta,\eta}(s,a) - \eta \sum_a \nabla \pi_\theta(a|s) \log \pi_\theta(a|s) \\
=&\sum_a \nabla \pi_\theta(a|s)\tilde{q}_{\pi_\theta,\eta}(s,a) - \eta \sum_a \nabla \pi_\theta(a|s) \log \pi_\theta(a|s) - \eta \nabla \sum_a \pi_\theta(a|s) \\
=&\sum_a \nabla \pi_\theta(a|s)\tilde{q}_{\pi_\theta,\eta}(s,a) - \eta \sum_a \nabla \pi_\theta(a|s) \log \pi_\theta(a|s) - \eta \sum_a \pi_\theta(a|s) \nabla \log \pi_\theta(a|s) \\
=&\sum_a \nabla \pi_\theta(a|s)\tilde{q}_{\pi_\theta,\eta}(s,a) + \eta \nabla \mathbb{H}\left(\pi_\theta(\cdot|s) \right).
\end{aligned}
$$

From (55), (56) and (57), it is easy to see that (1) $\tilde{v}_{\pi_\theta,\eta}(s)$, as well as $\tilde{q}_{\pi_\theta,\eta}(s,a)$, is Lipschitz continuous in θ with the Lipschitz constant being a continuous function of η; (2) $|\tilde{v}_{\pi_\theta,\eta}(s)|$, as well as $|\tilde{q}_{\pi_\theta,\eta}(s,a)|$, is bounded with the bound being a continuous function of η. From Lemmas 50 and 45, it is then easy to see both $\nabla \pi_\theta(a|s)$ and $\nabla \mathbb{H}\left(\pi_\theta(\cdot|s)\right)$ are bounded and Lipschitz continuous in θ. With Lemma 44, we, therefore, conclude that there exist continuous functions $L_{\Lambda_1}(\eta)$ and $U_{\Lambda_1}(\eta)$ such that for any s,

$$\begin{aligned} \left\|\Lambda_1(s,\theta,\eta) - \Lambda_1(s,\theta',\eta)\right\| &\leq L_{\Lambda_1}(\eta)\left\|\theta-\theta'\right\|, \\ \sup_\theta \left\|\Lambda_1(s,\theta,\eta)\right\| &\leq U_{\Lambda_1}(\eta). \end{aligned}$$

We now study the Lipschitz continuity of $\bar{\Lambda}_1(\theta,\eta)$ defined in (60). Lemma 48 confirms the Lipschitz continuity of d_{μ_θ}. Consequently, Lemma 44 implies that there exist continuous functions $L_{\bar{\Lambda}_1}(\eta)$ and $U_{\bar{\Lambda}_1}(\eta)$ such that

$$\begin{aligned} \left\|\bar{\Lambda}_1(\theta,\eta) - \bar{\Lambda}_1(\theta',\eta)\right\| &\leq L_{\bar{\Lambda}_1}(\eta)\left\|\theta-\theta'\right\|, \\ \sup_\theta \left\|\bar{\Lambda}_1(\theta,y,\eta)\right\| &\leq U_{\bar{\Lambda}_1}(\eta). \end{aligned}$$

We now study the Lipschitz continuity of $\Lambda_1'(\theta,y,\eta)$ defined in (61). Since $\tilde{J}_\eta(\theta)$ is $L_J + \eta L_H$ smooth, Lemma 45 implies that $L_J + \eta L_H$ is a Lipschitz constant of $\nabla \tilde{J}_\eta(\theta)$. From Lemma 50, it is easy to see the upper bound of $\tilde{\nabla} J_\eta(\theta)$ is also a continuous function of η. Consequently, Lemma 44 implies there exist continuous functions $L_{\Lambda_1'}(\eta)$ and $U_{\Lambda_1'}(\eta)$ such that for all y,

$$\begin{aligned} \left\|\Lambda_1'(\theta,y,\eta) - \Lambda_1'(\theta',y,\eta)\right\| &\leq L_{\Lambda_1'}(\eta)\left\|\theta-\theta'\right\|, \\ \sup_\theta \left\|\Lambda_1'(\theta,y,\eta)\right\| &\leq U_{\Lambda_1'}(\eta). \end{aligned}$$

Hence

$$\begin{aligned} \left\|\mathbb{E}\left[\tilde{M}_{121}\right]\right\| &= \left\|\Lambda_1'(\theta_t,S_t,\lambda_t) - \Lambda_1'(\theta_{t-\tau_{\beta_t}},S_t,\lambda_t)\right\| \\ &\leq L_{\Lambda_1'}(\lambda_t)\left\|\theta_t - \theta_{t-\tau_{\beta_t}}\right\| \\ &\leq L_{\Lambda_1'}(\lambda_t)L_\theta \beta_{t-\tau_{\beta_t},t-1} \quad \text{(Using (59))} \quad . \end{aligned}$$

Since $\lambda_t \in [0,\lambda]$, $L_{\Lambda_1'}(\eta)$ is a continuous function and well defined in $[0,\lambda]$, the extreme value theorem asserts that $L_{\Lambda_1'}(\eta)$ obtains its maximum in $[0,\lambda]$, say, e.g., $L^*_{\Lambda_1'}$. Then

$$\left\|\mathbb{E}\left[M_{121}\right]\right\| \leq L^*_{\Lambda_1'} L_\theta \beta_{t-\tau_{\beta_t},t-1},$$

which completes the proof. ■

E.22 Proof of Lemma 41

Lemma 74

$$\sum_{k=\lceil \frac{t}{2} \rceil}^{t} \frac{1}{\beta_k}\left(\tilde{J}_{\lambda_k}(\theta_{k+1}) - \tilde{J}_{\lambda_k}(\theta_k)\right) \leq \frac{3\lambda\beta\log|\mathcal{A}|}{1-\gamma} + \frac{2U_{\tilde{J}}}{\beta}(t+t_0)^{\epsilon_\beta}$$

Proof

$$
\begin{aligned}
&\sum_{k=\lceil \frac{t}{2}\rceil}^{t}\left(\frac{1}{\beta_k}\tilde{J}_{\lambda_k}(\theta_{k+1})-\frac{1}{\beta_k}\tilde{J}_{\lambda_k}(\theta_k)\right)\\
=&\sum_{k=\lceil \frac{t}{2}\rceil}^{t}\left(\frac{1}{\beta_{k-1}}\tilde{J}_{\lambda_{k-1}}(\theta_k)-\frac{1}{\beta_k}\tilde{J}_{\lambda_k}(\theta_k)\right)+\frac{1}{\beta_t}\tilde{J}_{\lambda_t}(\theta_{t+1})-\frac{1}{\beta_{\lceil \frac{t}{2}\rceil-1}}\tilde{J}_{\lambda_{\lceil \frac{t}{2}\rceil-1}}(\theta_{\lceil \frac{t}{2}\rceil})\\
=&\sum_{k=\lceil \frac{t}{2}\rceil}^{t}\left(\frac{1}{\beta_{k-1}}\tilde{J}_{\lambda_{k-1}}(\theta_k)-\frac{1}{\beta_{k-1}}\tilde{J}_{\lambda_k}(\theta_k)+\frac{1}{\beta_{k-1}}\tilde{J}_{\lambda_k}(\theta_k)-\frac{1}{\beta_k}\tilde{J}_{\lambda_k}(\theta_k)\right)\\
&+\frac{1}{\beta_t}\tilde{J}_{\lambda_t}(\theta_{t+1})-\frac{1}{\beta_{\lceil \frac{t}{2}\rceil-1}}\tilde{J}_{\lambda_{\lceil \frac{t}{2}\rceil-1}}(\theta_{\lceil \frac{t}{2}\rceil})\\
=&\sum_{k=\lceil \frac{t}{2}\rceil}^{t}\left(\frac{1}{\beta_{k-1}}(\lambda_{k-1}-\lambda_k)\frac{\sum_s d_{\pi_{\theta_k},\gamma}(s)\mathbb{H}\left(\pi_{\theta_k}(\cdot|s)\right)}{1-\gamma}+\left(\frac{1}{\beta_{k-1}}-\frac{1}{\beta_k}\right)\tilde{J}_{\lambda_k}(\theta_k)\right)\\
&+\frac{1}{\beta_t}\tilde{J}_{\lambda_t}(\theta_{t+1})-\frac{1}{\beta_{\lceil \frac{t}{2}\rceil-1}}\tilde{J}_{\lambda_{\lceil \frac{t}{2}\rceil-1}}(\theta_{\lceil \frac{t}{2}\rceil})\\
\leq&\sum_{k=\lceil \frac{t}{2}\rceil}^{t}\left(\frac{1}{\beta_{k-1}}\left(\frac{\lambda}{(k+t_0-1)^{\epsilon_\lambda}}-\frac{\lambda}{(k+t_0)^{\epsilon_\lambda}}\right)\frac{\log|\mathcal{A}|}{1-\gamma}+\left(\frac{1}{\beta_{k-1}}-\frac{1}{\beta_k}\right)\tilde{J}_{\lambda_k}(\theta_k)\right)\\
&+\frac{1}{\beta_t}\tilde{J}_{\lambda_t}(\theta_{t+1})-\frac{1}{\beta_{\lceil \frac{t}{2}\rceil-1}}\tilde{J}_{\lambda_{\lceil \frac{t}{2}\rceil-1}}(\theta_{\lceil \frac{t}{2}\rceil})\\
\overset{(i)}{\leq}&\sum_{k=\lceil \frac{t}{2}\rceil}^{t}\left(\frac{1}{\beta_{k-1}(k-1+t_0)^{1+\epsilon_\lambda}}\frac{\lambda\log|\mathcal{A}|}{1-\gamma}+\left(\frac{1}{\beta_{k-1}}-\frac{1}{\beta_k}\right)\tilde{J}_{\lambda_k}(\theta_k)\right)\\
&+\frac{1}{\beta_t}\tilde{J}_{\lambda_t}(\theta_{t+1})-\frac{1}{\beta_{\lceil \frac{t}{2}\rceil-1}}\tilde{J}_{\lambda_{\lceil \frac{t}{2}\rceil-1}}(\theta_{\lceil \frac{t}{2}\rceil}),\\
\overset{(ii)}{\leq}&\frac{3\lambda\beta\log|\mathcal{A}|}{1-\gamma}+\sum_{k=\lceil \frac{t}{2}\rceil}^{t}\left(\frac{1}{\beta_k}-\frac{1}{\beta_{k-1}}\right)U_{\tilde{J}}+\frac{1}{\beta_t}U_{\tilde{J}}+\frac{1}{\beta_{\lceil \frac{t}{2}\rceil-1}}U_{\tilde{J}}\\
=&\frac{3\lambda\beta\log|\mathcal{A}|}{1-\gamma}+\frac{U_{\tilde{J}}}{\beta_t}-\frac{U_{\tilde{J}}}{\beta_{\lceil \frac{t}{2}\rceil-1}}+\frac{1}{\beta_t}U_{\tilde{J}}+\frac{1}{\beta_{\lceil \frac{t}{2}\rceil-1}}U_{\tilde{J}}\\
=&\frac{3\lambda\beta\log|\mathcal{A}|}{1-\gamma}+\frac{2U_{\tilde{J}}}{\beta}(t+t_0)^{\epsilon_\beta},
\end{aligned}
$$

where (i) results from the inequality

$$
\begin{aligned}
\frac{1}{(t-1)^x}-\frac{1}{t^x}=\frac{t^x-(t-1)^x}{(t-1)^x t^x}&=\frac{t^x(t-1)^{1-x}-(t-1)}{(t-1)t^x}\\
&\leq\frac{t^x t^{1-x}-(t-1)}{(t-1)(t-1)^x}=\frac{1}{(t-1)^{1+x}}
\end{aligned}
$$

and (ii) results from the inequality

$$\sum_{t=1}^{\infty} \frac{1}{t^{1+\epsilon_\lambda+\epsilon_\beta}} \leq \sum_{t=1}^{\infty} \frac{1}{t^{1.5}} < 3.$$

■

References

Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. Optimality and approximation with policy gradient methods in markov decision processes. In *Proceedings of the Conference on Learning Theory*, 2020.

Kristopher De Asis, J. Fernando Hernandez-Garcia, G. Zacharias Holland, and Richard S. Sutton. Multi-step reinforcement learning: A unifying algorithm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Amir Beck. *First-order methods in optimization*. SIAM, 2017.

Albert Benveniste, Michel Métivier, and Pierre Priouret. *Adaptive Algorithms and Stochastic Approximations*. Springer, 1990.

Dimitri P Bertsekas and John N Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific Belmont, MA, 1996.

Shalabh Bhatnagar, Richard S Sutton, Mohammad Ghavamzadeh, and Mark Lee. Natural actor–critic algorithms. *Automatica*, 2009.

Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*. Springer, 2009.

Zaiwei Chen, Siva Theja Maguluri, Sanjay Shakkottai, and Karthikeyan Shanmugam. Finite-sample analysis of contractive stochastic approximation using smooth convex envelopes. *arXiv preprint arXiv:2002.00874*, 2020.

Zaiwei Chen, Siva Theja Maguluri, Sanjay Shakkottai, and Karthikeyan Shanmugam. A lyapunov theory for finite-sample guarantees of asynchronous q-learning and td-learning variants. *arXiv preprint arXiv:2102.01567*, 2021.

Zaiwei Chen, Sajad Khodadadian, and Siva Theja Maguluri. Finite-sample analysis of off-policy natural actor-critic with linear function approximation. *IEEE Control Systems Letters*, 2022.

Kamil Ciosek and Shimon Whiteson. Expected policy gradients for reinforcement learning. *Journal of Machine Learning Research*, 2020.

Bo Dai, Albert Shaw, Lihong Li, Lin Xiao, Niao He, Zhen Liu, Jianshu Chen, and Le Song. SBEED: convergent reinforcement learning with nonlinear function approximation. In *Proceedings of the International Conference on Machine Learning*, 2018.

Thomas Degris, Martha White, and Richard S. Sutton. Linear off-policy actor-critic. In *Proceedings of the International Conference on Machine Learning*, 2012.

Gabriel Dulac-Arnold, Daniel Mankowitz, and Todd Hester. Challenges of real-world reinforcement learning. *arXiv preprint arXiv:1904.12901*, 2019.

Lasse Espeholt, Hubert Soyer, Rémi Munos, Karen Simonyan, Volodymyr Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, Shane Legg, and Koray Kavukcuoglu. IMPALA: scalable distributed deep-rl with importance weighted actor-learner architectures. In *Proceedings of the International Conference on Machine Learning*, 2018.

Carles Gelada and Marc G. Bellemare. Off-policy deep reinforcement learning by bootstrapping the covariate shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the International Conference on Machine Learning*, 2018.

Assaf Hallak and Shie Mannor. Consistent on-line off-policy evaluation. In *Proceedings of the International Conference on Machine Learning*, 2017.

Jiawei Huang and Nan Jiang. On the convergence rate of off-policy policy optimization methods with density-ratio correction. *arXiv preprint arXiv:2106.00993*, 2021.

Sham M Kakade. A natural policy gradient. *Advances in Neural Information Processing Systems*, 2001.

Sajad Khodadadian, Zaiwei Chen, and Siva Theja Maguluri. Finite-sample analysis of off-policy natural actor-critic algorithm. In *Proceedings of the International Conference on Machine Learning*, 2021.

Sajad Khodadadian, Thinh T Doan, Siva Theja Maguluri, and Justin Romberg. Finite sample analysis of two-time-scale natural actor-critic algorithm. *IEEE Transactions on Automatic Control*, 2022.

Vijay R. Konda. *Actor-Critic Algorithms.* PhD thesis, Massachusetts Institute of Technology, 2002.

Vijay R. Konda and John N. Tsitsiklis. Actor-critic algorithms. In *Advances in Neural Information Processing Systems*, 1999.

Harshat Kumar, Alec Koppel, and Alejandro Ribeiro. On the sample complexity of actor-critic method for reinforcement learning with function approximation. *arXiv preprint arXiv:1910.08412*, 2019.

Harold Kushner and G George Yin. *Stochastic approximation and recursive algorithms and applications.* Springer Science & Business Media, 2003.

Romain Laroche and Remi Tachet. Dr Jekyll and Mr Hyde: the strange case of off-policy policy updates. In *Advances in Neural Information Processing Systems*, 2021.

David A Levin and Yuval Peres. *Markov chains and mixing times*. American Mathematical Soc., 2017.

Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909*, 2018.

Long Ji Lin. Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine Learning*, 1992.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, 2018.

Yanli Liu, Kaiqing Zhang, Tamer Basar, and Wotao Yin. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. In *Advances in Neural Information Processing Systems*, 2020.

Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Off-policy policy gradient with state distribution correction. *arXiv preprint arXiv:1904.08473*, 2019.

Peter Marbach and John N. Tsitsiklis. Simulation-based optimization of markov reward processes. *IEEE Transactions on Automatic Control*, 2001.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvári, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *Proceedings of the International Conference on Machine Learning*, 2020.

Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2016.

Ofir Nachum, Mohammad Norouzi, Kelvin Xu, and Dale Schuurmans. Bridging the gap between value and policy based reinforcement learning. In *Advances in Neural Information Processing Systems*, 2017.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Advances in Neural Information Processing Systems*, 2019.

Jan Peters and Stefan Schaal. Natural actor-critic. *Neurocomputing*, 2008.

Shuang Qiu, Zhuoran Yang, Jieping Ye, and Zhaoran Wang. On finite-time convergence of actor-critic algorithm. *IEEE Journal on Selected Areas in Information Theory*, 2021.

Gavin A Rummery and Mahesan Niranjan. *On-line Q-learning using connectionist systems*. University of Cambridge, Department of Engineering Cambridge, UK, 1994.

Simon Schmitt, Matteo Hessel, and Karen Simonyan. Off-policy actor-critic with shared experience replay. In *Proceedings of the International Conference on Machine Learning*, 2020.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Vedavyas Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy P. Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 2016.

Richard S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 1988.

Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction (2nd Edition)*. MIT press, 2018.

Richard S. Sutton, David A. McAllester, Satinder P. Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, 1999.

Richard S. Sutton, Hamid Reza Maei, Doina Precup, Shalabh Bhatnagar, David Silver, Csaba Szepesvári, and Eric Wiewiora. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the International Conference on Machine Learning*, 2009.

Richard S. Sutton, Joseph Modayil, Michael Delp, Thomas Degris, Patrick M. Pilarski, Adam White, and Doina Precup. Horde: a scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, 2011.

Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H. Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, Laurent Sifre, Trevor Cai, John P. Agapiou, Max Jaderberg, Alexander Sasha Vezhnevets, Rémi Leblond, Tobias Pohlen, Valentin Dalibard, David Budden, Yury Sulsky, James Molloy, Tom L. Paine, Çaglar Gülçehre, Ziyu Wang, Tobias Pfaff, Yuhuai Wu, Roman Ring, Dani Yogatama, Dario Wünsch, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy P. Lillicrap, Koray Kavukcuoglu, Demis Hassabis, Chris Apps, and David Silver. Grandmaster level in starcraft II using multi-agent reinforcement learning. *Nature*, 2019.

Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. In *Proceedings of the International Conference on Learning Representations*, 2019.

Yue Wang and Shaofeng Zou. Finite-sample analysis of greedy-gq with linear function approximation under markovian noise. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2020.

Ziyu Wang, Victor Bapst, Nicolas Heess, Volodymyr Mnih, Rémi Munos, Koray Kavukcuoglu, and Nando de Freitas. Sample efficient actor-critic with experience replay. In *Proceedings of the International Conference on Learning Representations*, 2017.

Martha White. Unifying task specification in reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2017.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 1992.

Ronald J Williams and Jing Peng. Function optimization using connectionist reinforcement learning algorithms. *Connection Science*, 1991.

Yue Wu, Weitong Zhang, Pan Xu, and Quanquan Gu. A finite-time analysis of two timescale actor-critic methods. In *Advances in Neural Information Processing Systems*, 2020.

Tengyu Xu, Zhe Wang, and Yingbin Liang. Improving sample complexity bounds for (natural) actor-critic algorithms. In *Advances in Neural Information Processing Systems*, 2020.

Tengyu Xu, Zhuoran Yang, Zhaoran Wang, and Yingbin Liang. Doubly robust off-policy actor-critic: Convergence and optimality. *arXiv preprint arXiv:2102.11866*, 2021.

Tom Zahavy, Zhongwen Xu, Vivek Veeriah, Matteo Hessel, Junhyuk Oh, Hado P van Hasselt, David Silver, and Satinder Singh. A self-tuning actor-critic algorithm. In *Advances in Neural Information Processing Systems*, 2020.

Kaiqing Zhang, Alec Koppel, Hao Zhu, and Tamer Basar. Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM Journal on Control and Optimization*, 2020a.

Shangtong Zhang, Bo Liu, and Shimon Whiteson. GradientDICE: Rethinking generalized offline estimation of stationary values. In *Proceedings of the International Conference on Machine Learning*, 2020b.

Shangtong Zhang, Bo Liu, Hengshuai Yao, and Shimon Whiteson. Provably convergent two-timescale off-policy actor-critic with function approximation. In *Proceedings of the International Conference on Machine Learning*, 2020c.

Shangtong Zhang, Hengshuai Yao, and Shimon Whiteson. Breaking the deadly triad with a target network. In *Proceedings of the International Conference on Machine Learning*, 2021.

Shaofeng Zou, Tengyu Xu, and Yingbin Liang. Finite-sample analysis for SARSA with linear function approximation. In *Advances in Neural Information Processing Systems*, 2019.