On the Convergence of SARSA with Linear Function Approximation

Shangtong Zhang[1]   **Remi Tachet des Combes**[2]   **Romain Laroche**[3]

Abstract

SARSA, a classical on-policy control algorithm for reinforcement learning, is known to *chatter* when combined with linear function approximation: SARSA does not diverge but oscillates in a bounded region. However, little is known about how fast SARSA converges to that region and how large the region is. In this paper, we make progress towards this open problem by showing the convergence rate of projected SARSA to a bounded region. Importantly, the region is much smaller than the region that we project into, provided that the magnitude of the reward is not too large. Existing works regarding the convergence of linear SARSA to a fixed point all require the Lipschitz constant of SARSA's policy improvement operator to be sufficiently small; our analysis instead applies to arbitrary Lipschitz constants and thus characterizes the behavior of linear SARSA for a new regime.

1. Introduction

SARSA is a classical on-policy control algorithm for reinforcement learning (RL, Sutton & Barto (2018)) dating back to Rummery & Niranjan (1994). The key idea of SARSA is to update the estimate for action values with data generated by following an exploratory and greedy policy (e.g., an ϵ-greedy policy) derived from the estimate itself. In this paper, we refer to the operator used for deriving such a policy from the action value estimate as the *policy improvement operator*.

The study of SARSA begins in the tabular setting, where the action value estimates are stored in the form of a look-up table. For example, Singh et al. (2000) confirm the asymptotic convergence of SARSA to the optimal policy provided that the policies from the policy improvement

operator satisfy the "greedy in the limit with infinite exploration" condition. Tabular methods, however, are not preferred when the state space is large and generalization is required across states. One possible solution is linear function approximation, which approximates the action values via the inner product of state-action features and a learnable weight vector. The behavior of SARSA with linear function approximation (linear SARSA) is, however, less understood.

Gordon (1996) and Bertsekas & Tsitsiklis (1996) empirically observe that linear SARSA can chatter: the weight vector does not go to infinity (i.e., it does not diverge) but oscillates in a bounded region. Importantly, this chattering behavior remains even if a decaying learning rate is used. Gordon (2001) further proves that *trajectory-based* linear SARSA with an ϵ-greedy policy improvement operator converges to a bounded region *asymptotically*. Unlike standard linear SARSA, where the policy improvement operator is invoked *every step* to generate a new policy for action selection in the next step, trajectory-based linear SARSA generates a policy at the beginning of each episode and the policy remains fixed during the episode. Intuitively, within an episode, trajectory-based linear SARSA is just linear Temporal Difference (TD, Sutton (1988)) learning for evaluating action values. It, therefore, converges to an approximation of the action value function of the policy generated at the beginning of the episode (Tsitsiklis & Roy, 1996). Since the number of all possible ϵ-greedy policies is finite in a finite Markov Decision Process with a fixed ϵ, trajectory-based linear SARSA oscillates among the (approximate) action value functions of those ϵ-greedy policies, which form a bounded region. Later, Perkins & Precup (2002) prove the asymptotic convergence of *fitted* linear SARSA (a.k.a. model-free approximate policy iteration) to a fixed point. Similar to trajectory-based SARSA, fitted SARSA alternates between *thorough* TD learning for policy evaluation under a fixed policy and the application of the policy improvement operator. In other words, it involves bi-level optimization. Then assuming the Lipschitz constant of the policy improvement operator is sufficiently small such that the composition of the policy improvement operator and some other function becomes contractive, convergence of fitted SARSA is obtained thanks to Banach's fixed point theorem. Despite this progress, the asymptotic behavior of standard linear SARSA, which invokes the policy improve-

---

[1]Department of Computer Science, University of Virginia, United States [2]AlpacaML [3]Unemployed. Correspondence to: Shangtong Zhang <shangtong@virginia.edu>.

	Gordon	Perkins & Precup	Melo et al.	Zou et al.	Ours
convergence	to a region	to a point	to a point	to a point	to a region
per-step policy improvement	✗	✗	✓	✓	✓
any Lipschitz constant	✓	✓*	✗	✗	✓
convergence rate	✗	✗	✗	✓	✓

Table 1. Comparison with existing works. ✓* indicates that the corresponding property is not explicitly documented in the original work. "Per-step policy improvement" means that the policy improvement operator is applied every time step.

ment operator every time step, still remains unclear, as does a potential convergence rate. Understanding the behavior of linear SARSA is one of the four open theoretical questions in RL raised by Sutton (1999).

Several efforts have been made to analyze linear SARSA. Melo et al. (2008) prove the asymptotic convergence of linear SARSA. Zou et al. (2019) further provide a convergence rate of a projected linear SARSA, which uses an additional projection operator in the canonical linear SARSA update. Unlike Gordon (2001), the convergence in Melo et al. (2008); Zou et al. (2019) is to a fixed point instead of a bounded region. Although convergence to a fixed point is preferred, Melo et al. (2008); Zou et al. (2019) require that SARSA's policy improvement operator is Lipschitz continuous and the Lipschitz constant is sufficiently small. It remains an open problem *how linear SARSA behaves when the Lipschitz constant is large*.

This problem is of interest because to our knowledge, most meaningful empirical results using SARSA for control consider an ϵ-greedy policy or a softmax policy. The former behaves similarly to a Lipschitz continuous policy improvement operator with a very large Lipschitz constant. The latter is usually tuned such that its Lipschitz constant is reasonably large. We refer the reader to Section 2 for more discussion about those two classes of policy improvement operators and now name a few notable empirical results. In the tabular setting, Sutton & Barto (2018) use an ϵ-greedy policy with $\epsilon = 0.1$ in the Windy GridWorld. With linear function approximation, Rummery & Niranjan (1994) use a softmax policy in a robot control problem with the temperature decaying from 0.05 to 0.01 such that the softmax policy has a reasonably large Lipschitz constant. Liang et al. (2015) use an ϵ-greedy policy with $\epsilon = 0.01$ to play all Atari games (Bellemare et al., 2013). In deep RL, Mnih et al. (2016) use ϵ-greedy policies with decaying ϵ in their asynchronous methods for playing Atari games. Besides the aforementioned interest from the empirical side, this problem has also been recognized as an important theoretical open problem in Perkins & Precup (2002); Zou et al. (2019).

In this paper, we make contributions to this open problem. In

particular, we study the projected linear SARSA (Zou et al., 2019) and show that it converges to a bounded region regardless of the magnitude of the Lipschitz constant of the policy improvement operator. Importantly, the bounded region is much smaller than the region we project into provided that the magnitude of rewards is not too large. The differences between our work and existing works are summarized in Table 1.

2. Background

In this paper, all vectors are column. We use $\langle x, y \rangle \doteq x^\top y$ to denote the standard inner product in Euclidean spaces. For a positive definite matrix D, we use $\|x\|_D \doteq \sqrt{x^\top D x}$ to denote the vector norm induced by D. We overload $\|\cdot\|_D$ to also denote the induced matrix norm. We write $\|\cdot\|$ as shorthand for $\|\cdot\|_I$, where I is the identity matrix, i.e., $\|\cdot\|$ denotes the standard ℓ_2 norm. When it does not cause confusion, we use vectors and functions interchangeably. For example, if f is a function $\mathcal{S} \to \mathbb{R}$, we also use f to denote the vector in $\mathbb{R}^{|\mathcal{S}|}$ whose s-indexed element is $f(s)$.

We consider an infinite horizon Markov Decision Process (MDP, Puterman (2014)) with a finite state space $\mathcal{S}$, a finite action space $\mathcal{A}$, a transition kernel $p : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$, a reward function $r : \mathcal{S} \times \mathcal{A} \to [-r_{max}, r_{max}]$, a discount factor γ, and an initial distribution $p_0 : \mathcal{S} \to [0, 1]$. At time step $t = 0$, an initial state S_0 is sampled from $p_0(\cdot)$. At time step t, an agent at a state S_t takes an action $A_t \sim \pi(\cdot|S_t)$ according to a policy $\pi : \mathcal{A} \times \mathcal{S} \to [0, 1]$. The agent then receives 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 defined as $G_t \doteq \sum_{i=0}^{\infty} \gamma^i R_{t+i+1}$, which allows us to define the action value function as

$$q_\pi(s, a) \doteq \mathbb{E}\left[G_t | S_t = s, A_t = a, \pi, p\right].$$

The action value function q_π is closely related to the Bellman operator $\mathcal{T}_\pi$, which is defined as $\mathcal{T}_\pi q \doteq r + \gamma P_\pi q$, where $P_\pi \in \mathbb{R}^{|\mathcal{S} \times \mathcal{A}| \times |\mathcal{S} \times \mathcal{A}|}$ is the state-action pair transition matrix, i.e., $P_\pi((s, a), (s', a')) \doteq p(a'|s, a)\pi(a'|s')$. In particular, q_π is the only vector $q \in \mathbb{R}^{|\mathcal{S} \times \mathcal{A}|}$ satisfying $q = \mathcal{T}_\pi q$. The goal of control is to find an optimal policy π_*

such that $\forall \pi, s, a, q_{\pi_*}(s, a) \geq q_\pi(s, a)$. All optimal policies share the same action value function, which is referred to as q_*. One classical approach for finding q_* is SARSA, which updates an estimate $q \in \mathbb{R}^{|\mathcal{S} \times \mathcal{A}|}$ iteratively as

$$A_{t+1} \sim \pi_q(\cdot|S_{t+1}), \tag{1}$$
$$\delta_t \leftarrow R_{t+1} + \gamma q(S_{t+1}, A_{t+1}) - q(S_t, A_t)$$
$$q(S_t, A_t) \leftarrow q(S_t, A_t) + \alpha_t \delta_t,$$

where $\{\alpha_t\}$ is a sequence of learning rates and π_q denotes that the policy π is parameterized by the action value estimate q. A commonly used π_q is an ϵ-greedy policy, i.e.,

$$\pi_q(a|s) = \begin{cases} \frac{\epsilon}{|\mathcal{A}|}, & a \neq \arg\max_{b \in \mathcal{A}} q(s, b) \\ 1 - \epsilon + \frac{\epsilon}{|\mathcal{A}|}, & \text{otherwise} \end{cases}, \tag{2}$$

where $\epsilon \in [0, 1]$ is a hyperparameter. Another common example is an ϵ-softmax policy, i.e.,

$$\pi_q(a|s) = \frac{\epsilon}{|\mathcal{A}|} + (1 - \epsilon) \frac{\exp(q(s,a)/\iota)}{\sum_b \exp(q(s,b)/\iota)}, \tag{3}$$

where $\iota \in (0, \infty)$ is the temperature of the softmax function. This π_q is exactly the policy improvement operator discussed in Section 1: it maps an action value estimate q to a new policy; it is "improvement" in that it usually has greedification over the action value estimate to some extent. The ϵ-softmax policy π_q in (3) is Lipschitz continuous in q with the Lipschitz constant being $\frac{1-\epsilon}{\iota}$. When the temperature ι approaches 0, the ϵ-softmax policy approaches the ϵ-greedy policy. Therefore, despite the ϵ-greedy policy is not even continuous, it is the limit of a sequence of Lipschitz continuous policies with the Lipschitz constants approaching infinity. We, therefore, argue that an ϵ-greedy policy would empiricially behave similarly to a Lipschitz continuous policy with a very large Lipschitz constant.

So far we have considered only time-homogeneous policies. One can also consider time-inhomogeneous policies, e.g., a policy $\pi_{q,t}(a|s)$ that depends on both the action value estimate q and the time step t. Singh et al. (2000) show that if the time-inhomogeneous policies $\pi_{q,t}$ satisfy the "greedy in the limit with infinite exploration" (GLIE) condition then the iterates generated by (1) converge to q_* almost surely.

It is, however, not always practical to use a look-up table for storing our action value estimates, especially when the state space is large or generalization is required across states. One natural solution is linear function approximation, where the action value estimate $q(s, a)$ is parameterized as $x(s, a)^\top w$. Here $x : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^K$ is the feature function which maps a state-action pair to a K-dimensional vector and $w \in \mathbb{R}^K$ is the learnable weight vector. We use $X \in \mathbb{R}^{|\mathcal{S} \times \mathcal{A}| \times K}$ to denote the feature matrix, whose (s, a)-indexed row is $x(s, a)^\top$. We use as shorthand

$$\pi_w \doteq \pi_{Xw}, \ x_t \doteq x(S_t, A_t), \ x_{max} \doteq \max_{s,a} \|x(s, a)\|.$$

Algorithm 1 SARSA with linear function approximation

> Initialize w_0 such that $\|w_0\| \leq C_\Gamma$
> $S_0 \sim p_0(\cdot), A_0 \sim \pi_{w_0}(\cdot|S_0)$
> $t \leftarrow 0$
> **while** true **do**
>   Execute A_t, get R_{t+1}, S_{t+1}
>   Sample $A_{t+1} \sim \pi_{w_t}(\cdot|S_{t+1})$
> $\delta_t \leftarrow R_{t+1} + \gamma x_{t+1}^\top w_t - x_t^\top w_t$
> $w_{t+1} \leftarrow \Gamma(w_t + \alpha_t \delta_t x_t)$
> $t \leftarrow t + 1$
> **end while**

SARSA (Algorithm 1) is a commonly used algorithm for learning w. In Algorithm 1, $\Gamma : \mathbb{R}^K \to \mathbb{R}^K$ is a projection operator onto a ball of radius C_Γ, i.e.,

$$\Gamma(w) \doteq \begin{cases} w, & \|w\| \leq C_\Gamma, \\ C_\Gamma \frac{w}{\|w\|}, & \|w\| > C_\Gamma \end{cases}. \tag{4}$$

For now we consider the setting where $C_\Gamma = \infty$, i.e., Γ is an identity mapping. If the iterates $\{w_t\}$ generated by SARSA converged to some vector w_*, the expected update at w_* would have to diminish, i.e.,

$$\mathbb{E}_{S_t, A_t \sim d_{\pi_{w_*}}} \left[(R_{t+1} + \gamma x_{t+1}^\top w_* - x_t^\top w_*) x_t \right] = 0, \tag{5}$$

where for a policy π, we use $d_\pi \in \mathbb{R}^{|\mathcal{S} \times \mathcal{A}|}$ to denote the stationary state action pair distribution of the chain in $\mathcal{S} \times \mathcal{A}$ induced by π (assuming it exists). We can equivalently write (5) in a matrix form as

$$X^\top D_{\pi_{w_*}} (r + \gamma P_{\pi_{w_*}} X w_* - X w_*) = 0, \tag{6}$$

where for a policy π, we use $D_\pi \in \mathbb{R}^{|\mathcal{S} \times \mathcal{A}| \times |\mathcal{S} \times \mathcal{A}|}$ to denote a diagonal matrix whose diagonal entry is d_π. By defining

$$A_\pi \doteq X^\top D_\pi(\gamma P_\pi - I)X, \ b_\pi \doteq X^\top D_\pi r, \tag{7}$$

(6) becomes $A_{\pi_{w_*}} w_* + b_{\pi_{w_*}} = 0$. It is known (see, e.g., Tsitsiklis & Roy (1996)) that A_π is negative definite under mild conditions. We define a projection operator Π_{D_π} mapping a vector in $\mathbb{R}^{|\mathcal{S} \times \mathcal{A}|}$ to the column space of X as

$$\Pi_{D_\pi} q \doteq \arg\min_{\hat{q} \in col(X)} \|\hat{q} - q\|_{D_\pi}^2,$$

where $col(X)$ denotes the column space of X. It can be computed that $\Pi_{D_\pi} = X(X^\top D_\pi X)^{-1} X^\top D_\pi$ and it is known (see, e.g., De Farias & Van Roy (2000)) that (6) holds if and only if $\Pi_{D_{\pi_{w_*}}} \mathcal{T}_{\pi_{w_*}} X w_* = X w_*$. In other words, $X w_*$ is a fixed point of the operator $\mathcal{H}(q) \doteq \Pi_{D_{\pi_q}} \mathcal{T}_{\pi_q} q$. The operator $\mathcal{H}$ is referred to as the *approximate policy iteration* operator and SARSA is an incremental, stochastic method to find a fixed point of approximate policy iteration. Unfortunately, when $\pi_q(a|s)$ is not continuous in q (e.g., π_q

is an ϵ-greedy policy, c.f. (2)), $\mathcal{H}$ does not necessarily have a fixed point (De Farias & Van Roy, 2000). Conversely, when $\pi_q(a|s)$ is continuous in q, De Farias & Van Roy (2000) show that $\mathcal{H}$ has at least one fixed point.

Perkins & Precup (2002) assume π_q is Lipschitz continuous in q and study a form of fitted SARSA, which is a model-free variant of approximate policy iteration. At the k-th iteration, Perkins & Precup (2002) first invoke TD for learning the action value function of π_k, which converges to $w_k = A_{\pi_k}^{-1} b_{\pi_k}$ after infinitely many steps. Then the policy for the $(k+1)$-th iteration is obtained via invoking the policy improvement operator, i.e., $\pi_{k+1} \doteq \pi_{w_k}$. Perkins & Precup (2002) show that

$$\left\| \pi_{A_{\hat{\pi}}^{-1} b_{\hat{\pi}}} - \pi_{A_{\pi'}^{-1} b_{\pi'}} \right\| \leq \mathcal{O}\left(L_\pi\right) \|\hat{\pi} - \pi'\|, \qquad (8)$$

where the policies $\hat{\pi}$ and π' should be interpreted as vectors in $\mathbb{R}^{|\mathcal{S} \times \mathcal{A}|}$ whose (s,a)-indexed element is $\pi(a|s)$ when computing $\|\hat{\pi} - \pi'\|$ and L_π denotes the Lipschitz constant of the policy improvement operator π_w, i.e., $\forall s, a$,

$$|\pi_{w_1}(a|s) - \pi_{w_2}(a|s)| \leq L_\pi \|w_1 - w_2\|.$$

Consequently, if L_π is sufficiently small, the function $x \to \pi_{A_x^{-1} b_x}$, which maps π_k to π_{k+1}, becomes a contraction. Banach's fixed point theorem then confirms the convergence of fitted SARSA. From the definition of A_π and b_π in (7), it is easy to see that (8) can also be expressed as

$$\left\| \pi_{A_{\hat{\pi}}^{-1} b_{\hat{\pi}}} - \pi_{A_{\pi'}^{-1} b_{\pi'}} \right\| \leq \mathcal{O}\left(L_\pi \|r\|\right) \|\hat{\pi} - \pi'\|$$

because r is a multiplier in the definition of b_π. Hence, for any L_π, if the magnitude of the reward $\|r\|$ is small enough, the function $x \to \pi_{A_x^{-1} b_x}$ is contractive and fitted SARSA remains convergent. Nevertheless, Perkins & Precup (2002) share the same spirit as Gordon (2001) by holding the policy fixed for sufficiently (possibly infinitely) many steps to wait for the policy evaluation to complete.

When it comes to standard linear SARSA that updates the policy every time step, Melo et al. (2008) consider, for a fixed point w_*,

$$C_{w_*} \doteq \sup_w \left\| A_{\pi_w} - A_{\pi_{w_*}} \right\| + \sup_{w \neq w_*} \frac{\|b_{\pi_w} - b_{\pi_{w_*}}\|}{\|w - w_*\|}.$$

They show that $C_{w_*} = \mathcal{O}\left(L_\pi\right)$ and if L_π is small enough such that

$$A_{\pi_{w_*}} + C_{w_*} I \qquad (9)$$

is negative definite, SARSA converge to w_*. The convergence of SARSA in Perkins & Precup (2002); Melo et al. (2008) does not require the projection operator (i.e., $C_\Gamma = \infty$) but is only asymptotic, Zou et al. (2019) further

provide a convergence rate of SARSA using some $C_\Gamma < \infty$, assuming L_π is small enough such that

$$A_{\pi_{w_*}} + \mathcal{O}\left(L_\pi \left(r_{max} + 2x_{max} C_\Gamma\right)\right) I \qquad (10)$$

is negative definite. It is easy to see that if L_π is not small enough, neither (9) nor (10) can be guaranteed to be negative definite no matter how small $\|r\|$ is. This is because the $\sup_w \left\| A_{\pi_w} - A_{\pi_{w_*}} \right\|$ term in C_{w_*} and the $2x_{max} C_\Gamma$ term in (10) are independent of r. In other words, unlike Perkins & Precup (2002), where the requirement for a sufficiently small L_π is not necessary when the magnitude of the reward is enough, a sufficiently small L_π is an essential requirement for the analysis of Melo et al. (2008); Zou et al. (2019). We, however, note that requring a finite C_Γ and a sufficiently small L_π can be restrictive. To see this, consider the ϵ-softmax policy in (3). On the one hand, for this parameterization, we have $L_\pi = \frac{1-\epsilon}{\iota}$. For L_π to be sufficiently small, we have to ensure the temperature ι to be sufficiently large. On the other hand, a finte C_Γ ensures that the action value estimate $q(s,a)$ inside $\exp(\cdot)$ is bounded. Combining the two facts together, it is easy to see that the ϵ-softmax policy cannot be much different from a uniformly random policy. Or more formally speaking, the probability of any action is at most

$$\frac{\epsilon}{|\mathcal{A}|} + \frac{1 - \epsilon}{1 + (|\mathcal{A}| - 1)\exp\left(-\frac{2x_{max} C_\Gamma}{\iota}\right)}.$$

When ι is too large, the above probability approaches $\frac{1}{|\mathcal{A}|}$, making it hard to ensure sufficient exploitation. The behavior of SARSA with a large L_π, however, remains an open problem.

3. Stochastic Approximation with Rapidly Changing Markov Chains

To prepare us for the analysis of SARSA, we show, in this section, a convergence rate (to a bounded region) of a generic stochastic approximation algorithm. More precisely, we consider the following iterative updates:

$$w_{t+1} \doteq \Gamma\left(w_t + \alpha_t \left(F_{\theta_t}(w_t, Y_t) - w_t\right)\right). \qquad (11)$$

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 a 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 Γ is the projection operator defined in (4). Importantly, we consider the setting where

$$\forall t, \theta_t \equiv w_t.$$

In other words, there is only a single iterate in our setting. To ease presentation, we use $\{w_t\}$ and $\{\theta_t\}$ to denote the

same quantity. This emphasizes their different roles as the iterates of interest and as the controller of the transition kernel.

Our analysis is a natural extension of Chen et al. (2021) and Zhang et al. (2022) but has significant differences to theirs. Chen et al. (2021) consider a time-homogeneous Markov chain (i.e., $\forall t, \theta_t \equiv \theta_0$). Consequently, their results are naturally applicable to policy evaluation problems. Zhang et al. (2022) consider a time-inhomogeneous Markov chain, where the iterates $\{w_t\}$ are *different* from the sequence $\{\theta_t\}$. More importantly, Zhang et al. (2022) assume that $\{\theta_t\}$ changes sufficiently slowly, i.e., there exists another sequence $\{\beta_t\}$ such that

$$\|\theta_{t+1} - \theta_t\| = \mathcal{O}\left(\beta_t\right), \quad \lim_{t \to \infty} \frac{\beta_t}{\alpha_t} = 0.$$

This is the classical two-timescale setting (see, e.g., Borkar (2009)) and their analysis naturally applies to actor-critic algorithms (Konda & Tsitsiklis, 1999) with $\{w_t\}$ and $\{\theta_t\}$ interpreted as critic and actor respectively. We instead consider the setting where $\forall t, \theta_t = w_t$. In other words, the time-inhomogeneous Markov chain we consider changes *rapidly*, which is the main challenge of our analysis. As a consequence, we introduce the projection operator Γ, not required in Chen et al. (2021); Zhang et al. (2022). The price is that we only show convergence to a bounded region while Chen et al. (2021); Zhang et al. (2022) show convergence to points. Convergence to a bounded region is, however, sufficient for our purpose of understanding the behavior of SARSA since it matches what practitioners have observed. Furthermore, we believe our analysis might be applicable to other RL algorithms and might also have independent interest beyond RL. We now state our assumptions.

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.

Those two assumptions are identical to those of Zhang et al. (2022). Assumption 3.1 states that the random process $\{Y_t\}$ is a time-inhomogeneous Markov chain. Assumption 3.2 states the ergodicity of the Markov chains. Assumption 3.2 is also used in the analysis of RL algorithms in both on-policy (Marbach & Tsitsiklis, 2001) and off-policy (Zhang et al., 2022; 2021) settings. We later show how SARSA(λ) can trivially fulfill Assumption 3.2. The uniform ergodicity in Assumption 3.2 immediately implies uniform mixing.

Lemma 3.3. *(Lemma 1 of Zhang et al. (2022)) Let Assumption 3.2 hold. Then, there exist constants $C_M > 0$ and $\tau \in (0, 1)$, independent of θ, such that for any $n > 0$,*

$$\sup_{y, \theta} \sum_{y'} |P_\theta^n(y, y') - d_\theta(y')| \le C_M \tau^n.$$

As noted in Zhang et al. (2022), the uniform mixing property in Lemma 3.3 is usually a direct technical assumption in previous works (e.g., Zou et al. (2019); Wu et al. (2020)).

Assumption 3.4. (Uniform pseudo-contraction) Let

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

$$f_\theta^\alpha(w) \doteq w + \alpha \left(\bar{F}_\theta(w) - w \right).$$

Then,

(i) For any θ, $\bar{F}_\theta$ has a unique fixed point, i.e., there exists a unique w_θ^* such that

$$\bar{F}_\theta(w_\theta^*) = w_\theta^*.$$

(ii) There exists a constant $\bar{\alpha} > 0$ such that for all $\alpha \in (0, \bar{\alpha})$, f_θ^α is a uniform pseudo-contraction, i.e., there exists a constant $\kappa_\alpha \in (0, 1)$ (depending on α), such that for all θ, w,

$$\|f_\theta^\alpha(w) - w_\theta^*\| \le \kappa_\alpha \|w - w_\theta^*\|.$$

Assumption 3.4 is another difference from Chen et al. (2021); Zhang et al. (2022). Namely, Chen et al. (2021); Zhang et al. (2022) require $\bar{F}_\theta$ to be a contraction while we only require f_θ^α to be a pseudo-contraction. It is easy to see that the contraction of $\bar{F}_\theta$ immediately implies the pseudo-contraction of f_θ^α but not in the opposite direction. In other words, our assumption is weaker.

Assumption 3.5. (Continuity and boundedness) There exist constants $L_F, L_F', L_F'', U_F, U_F', U_F'', L_w, U_w, L_P$ such that for any $w, w', y, y', \theta, \theta'$,

(i). $\|F_\theta(w, y) - F_\theta(w', y)\| \le L_F \|w - w'\|$

(ii). $\|F_\theta(w, y) - F_{\theta'}(w, y)\| \le L_F' \|\theta - \theta'\|(\|w\| + U_F')$

(iii). $\|F_\theta(0, y)\| \le U_F$

(iv). $\left\| \bar{F}_\theta(w) - \bar{F}_{\theta'}(w) \right\| \le L_F'' \|\theta - \theta'\|(\|w\| + U_F'')$

(v). $\|w_\theta^* - w_{\theta'}^*\| \le L_w \|\theta - \theta'\|$

(vi). $\sup_\theta \|w_\theta^*\| \le U_w$

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

Assumption 3.5 states some regularity conditions for the functions we consider and is identical to that of Zhang et al. (2022).

Assumption 3.6. (Projection)

(i). $\|w_0\| \leq C_\Gamma, U_w \leq C_\Gamma.$

(ii). For any θ, w, y, we have

$$P_\theta = P_{\Gamma(\theta)}, \quad F_\theta(w, y) = F_{\Gamma(\theta)}(w, y),$$
$$w_\theta^* = w_{\Gamma(\theta)}^*.$$

Assumption 3.6 requires that some of the functions we consider are invariant to the projection operator. We will later show that SARSA(λ) trivially satisfies this assumption.

Assumption 3.7. The learning rates $\{\alpha_t\}$ have the form

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

where $c_\alpha > 0, \epsilon_\alpha \in (0, 1], t_0 > 0$ are constants to be tuned.

Assumption 3.7 is just one of many possible forms of learning rates; we use this particular one to ease presentation. Importantly, the learning rates $\{\alpha_t\}$ here do *not* verify the Robbins-Monro's condition (Robbins & Monro, 1951) when $\epsilon_\alpha \leq 0.5$, neither do the learning rates in Wu et al. (2020); Chen et al. (2021).

We now present our analysis. Given the sequences $\{\theta_t\}$ (i.e., $\{w_t\}$) and $\{Y_t\}$ in (11), we define an auxiliary sequence $\{u_t\}$ as

$$u_0 \doteq w_0,$$
$$u_{t+1} \doteq \Gamma(u_t) + \alpha_t(F_{\theta_t}(\Gamma(u_t), Y_t) - \Gamma(u_t)). \quad (12)$$

Lemma 3.8. *Let Assumption 3.6 hold. Then for any t,* $w_t = \Gamma(u_t)$.

Proof. It follows immediately from induction. $\square$

Intuitively, $\{u_t\}$ is simply the pre-projection version of $\{w_t\}$. We are interested in $\{u_t\}$ because it has the following nice property.

Theorem 3.9. *Let Assumptions 3.1 - 3.7 hold. If t_0 is sufficiently large, then the iterates $\{u_t\}$ generated by (12) satisfy*

$$\mathbb{E}\left[\left\|u_{t+1} - w_{\theta_{t+1}}^*\right\|^2\right]$$
$$\leq \left(1 - 2\left(1 - \kappa_{\alpha_t} - \mathcal{O}\left(\alpha_t^2 \log^2 \alpha_t\right)\right)\right) \mathbb{E}\left[\left\|\Gamma(u_t) - w_{\theta_t}^*\right\|^2\right]$$
$$+ 2L_w L_\theta \alpha_t \mathbb{E}\left[\left\|\Gamma(u_t) - w_{\theta_t}^*\right\|\right] + \mathcal{O}\left(\alpha_t^2 \log^2 \alpha_t\right),$$

where $L_\theta \doteq U_F + (L_F + 1)C_\Gamma$.

See Section A.1 for the proof of Theorem 3.9, where the constants hidden by $\mathcal{O}(\cdot)$ and how large t_0 is are also explicitly documented. Theorem 3.9 gives a recursive form of some error terms. We, however, cannot go further unless we have the domain knowledge of κ_α.

Corollary 3.10. *Let Assumptions 3.1 - 3.7 hold. Assume $\kappa_\alpha = \sqrt{1 - \eta\alpha}$ for some positive constant $\eta > 0$. If t_0 is sufficiently large, then the iterates $\{w_t\}$ generated by (11) satisfy*

$$\mathbb{E}\left[\left\|w_t - w_{w_t}^*\right\|^2\right] = \frac{72 L_w^2 L_\theta^2}{\eta^2}$$
$$+ \begin{cases} \mathcal{O}\left(t^{-\frac{\eta c_\alpha}{3}} \log^2 t\right), & \epsilon_\alpha = 1, \eta c_\alpha \in (0, 3) \\ \mathcal{O}\left(\frac{\log^3 t}{t}\right), & \epsilon_\alpha = 1, \eta c_\alpha = 3 \\ \mathcal{O}\left(\frac{\log^2 t}{t}\right), & \epsilon_\alpha = 1, \eta c_\alpha \in (3, \infty) \\ \mathcal{O}\left(\frac{\log^2 t}{t^{\epsilon_\alpha}}\right), & \epsilon_\alpha \in (0, 1) \end{cases}.$$

See Section A.2 for the proof of Corollary 3.10, as well as the constants hidden by $\mathcal{O}(\cdot)$ and how large t_0 is. Due to the projection operator, we have the bound

$$\mathbb{E}\left[\left\|w_t - w_{w_t}^*\right\|^2\right] \leq 4C_\Gamma^2.$$

So Corollary 3.10 is informative only if

$$\frac{72 L_w^2 L_\theta^2}{\eta^2} \leq 4C_\Gamma^2.$$

This is where we need more domain knowledge and the analysis in the next section provides an example.

4. SARSA with Linear Function Approximation

We first analyze SARSA and expected SARSA with the following assumptions.

Assumption 4.1. (Lipschitz continuity) There exists $L_\pi > 0$ such that $\forall w, w', a, s$,

$$\|\pi_w(a|s) - \pi_{w'}(a|s)\| \leq L_\pi \|w - w'\|.$$

Assumption 4.2. (Uniform ergodicity) Let $\bar{\Lambda}_\pi$ be the closure of $\{\pi_w \mid w \in \mathbb{R}^{|\mathcal{S} \times \mathcal{A}|}\}$. For any $\pi \in \bar{\Lambda}_\pi$, the chain induced by π is ergodic and $\pi(a|s) > 0$.

Assumption 4.3. (Linear independence) The feature matrix X has full column rank.

Assumption 4.1 is also used in Perkins & Precup (2002); Melo et al. (2008); Zou et al. (2019). As noted by Zhang et al. (2022), Assumption 4.2 is easy to fulfill especially when the chain induced by a uniformly random policy is ergodic, which we believe is a fairly weak assumption. An example policy satisfying those two assumptions is the ϵ-softmax policy in (3) with any $\epsilon \in (0, 1]$, provided that the chain induced by a uniformly random policy is ergodic. Assumption 4.3 is standard in the literature regarding RL with linear function approximation (see, e.g., Tsitsiklis & Roy (1996)).

As discussed in (7), if SARSA, as well as expected SARSA, converged to some vector w_*, that vector would verify

$$w_* = A_{\pi_{w_*}}^{-1} b_{\pi_{w_*}}.$$

This inspires us to define the error function

$$e(w)$$
$$\doteq \left\| w - \left(X^\top D_{\pi_{\Gamma(w)}} \left(\gamma P_{\pi_{\Gamma(w)}} - I \right) X \right)^{-1} X^\top D_{\pi_{\Gamma(w)}} r \right\|^2$$

to study the behavior of SARSA. Here we have included the projection operator Γ in the definition of $e(w)$ because we use a finite C_Γ in Algorithm 1.

Theorem 4.4. *Let Assumptions 3.7 and 4.1 - 4.3 hold. Assume $\|X\| = 1$ and r_{max} is not so large such that*

$$L_w \doteq \mathcal{O}\left(L_\pi r_{max} \right) < 1.$$

Assume C_Γ is large enough such that

$$U_w \doteq \mathcal{O}\left(r_{max} \right) \leq C_\Gamma.$$

Let t_0 be sufficiently large. Then the iterates $\{w_t\}$ generated by Algorithm 1 satisfy

$$\mathbb{E}\left[\|w_t - w_*\| \right] = \frac{6\sqrt{2} L_w \left(1 + 4 C_\Gamma \right)}{\eta(1 - L_w)}$$
$$+ \begin{cases} \mathcal{O}\left(t^{-\frac{\eta c_\alpha}{6}} \log t \right), & \epsilon_\alpha = 1, \eta c_\alpha \in (0, 3) \\ \mathcal{O}\left(t^{-\frac{1}{2}} \log^{\frac{3}{2}} t \right), & \epsilon_\alpha = 1, \eta c_\alpha = 3 \\ \mathcal{O}\left(t^{-\frac{1}{2}} \log t \right), & \epsilon_\alpha = 1, \eta c_\alpha \in (3, \infty) \\ \mathcal{O}\left(t^{-\frac{\epsilon_\alpha}{2}} \log t \right), & \epsilon_\alpha \in (0, 1) \end{cases},$$

where η is a positive constant and w_* is the unique vector such that $e(w_*) = 0$.

We prove Theorem 4.4 mainly by invoking Corollary 3.10. See Section B.1 for more details. The exact expressions of L_w, U_w, η are detailed in (23), (24), (22) in the proof, all of which are independent of C_Γ.

Significance of Theorem 4.4. Theorem 4.4 ensures that the iterates $\{w_t\}$ converge, in expectation, to a ball of size

$$R_* \doteq \frac{6\sqrt{2} L_w (1 + 4 C_\Gamma)}{\eta(1 - L_w)},$$

centered at w_*. Due to the use of projection, one can also trivially claim that the iterates always stay in a ball of size $2C_\Gamma$, centered at w_*. Thus Theorem 4.4 is informative only if

$$\frac{R_*}{2 C_\Gamma} = \frac{24\sqrt{2} L_w}{\eta(1 - L_w)} \times \frac{1 + 4 C_\Gamma}{4 C_\Gamma} < 1.$$

The term $\frac{1 + 4 C_\Gamma}{4 C_\Gamma}$ monotonically decreases when the size of the ball for projection increases and eventually converges

to 1. So this ratio is essentially determined by the first term $\frac{24\sqrt{2} L_w}{\eta(1 - L_w)}$, which can be arbitrarily small as long as L_w is small. In other words, suppose L_w is small enough, no matter how large the ball used for projection is (practitioners usually use a very large ball for projection), the iterates generated by linear SARSA asymptotically only visit a small portion of the ball. The exact portion is determined by the relative magnitude of L_w and other properties of the problem and can be arbitrarily small when L_w is small enough. Here we want to compare with Q-learning (Watkins, 1989) with linear function approximation. As demonstrated in Baird's counterexample (Baird, 1995), the Q-learning iterates can diverge to infinity. This essentially means that if we apply a ball for projection in linear Q-learning, the iterates can asymptotically visit every part of the ball. By contrast, Theorem 4.4 proves that the iterates generated by linear SARSA visit asymptotically only a possibly small portion of the ball, on some problems. In those problems, Theorem 4.4, to our best knowledge, is the first to characterize the fundamentally different behaviors between linear SARSA and linear Q-learning. We regard this as the most important contribution of this work. This difference demonstrates the challenge in off-policy learning compared with on-policy learning.

Limitation of Theorem 4.4. That being said, there are a few things that Theorem 4.4 does not offer.

First, Theorem 4.4 is solely about the magnitude of the iterates and is not about the performance of the corresponding policy. In other words, this work does *not* fully address the question raised by Sutton (1999). Addressing that question requires to understand (a) how linear SARSA iterates behave asymptotically and (b) why such behavior generates good performance. This work contributes to the former but does *not* contribute to the latter.

Second, the ball of size R_* specified in Theorem 4.4 is small only when compared with the the ball for projection on some problems. If we compare R_* with some other quantities of the problem, it is indeed very large. On some problems where L_w is not small enough, the ball is also quite large. In other words, even for (a), we only address it in some problems with a very coarse bound. On problems where L_w does not meet our condition, Theorem 4.4 simply does not apply. Theorem 4.4 is *not* meant to be a general result that applies to all problems. We, however, argue that this is still, to our knowledge, the best result regarding the question in Sutton (1999).

Third, even if we step back to the tabular setting, Theorem 4.4 is still convergence to only a ball instead of a fixed point. This indicates the fundamental limit of the techniques employed therein. This might seem disappointing at first glance but less so when taking a holistic view of the history

of SARSA. The best result regarding tabular SARSA for control, to our knowledge, is Singh et al. (2000), which proves that if the ϵ decays properly in the ϵ-greedy policy, tabular SARSA converges to an optimal policy. That being said, to implement the decay, it is required to maintain a counter of the state-action visitations and the policy itself is non-Markovian in that it is a function of the current time step. When it comes to the canonical Markovian policies that depend only on current states, we still know nothing about the behavior of tabular SARSA for control.

5. Related Work

Comparison with Zou et al. (2019). The most similar result to our work is Zou et al. (2019). Assuming L_π is small enough, Zou et al. (2019) give a finite sample analysis of the convergence of $\{w_t\}$ to w_*. Theorem 4.4 requires L_w to be sufficiently small. This holds when either L_π or r_{max} is sufficiently small. In the case where L_π is sufficiently small, our result is indeed weaker than Zou et al. (2019) because Zou et al. (2019) show convergence to a point while we show convergence to only a region. Thus the only scenario that our result is preferred over Zou et al. (2019) is when L_π is large but r_{max} is small. In this scenario, our result still apply but Zou et al. (2019) do *not* apply. The reason is that in Theorem 4.4, our condition is related to the product $L_\pi r_{max}$. So the role of L_π and r_{max} is interchangeable and a small L_π is only a sufficient condition for the product to be small. However, Zou et al. (2019) essentially consider a condition in the form of $L_\pi r_{max} + L_\pi C_\Gamma$ (see (10) or Assumption 2 in Zou et al. (2019)). For this summation to be sufficiently small, a small L_π is a necessary condition. This means although our convergence is weaker than Zou et al. (2019), our result applies to much more problems than Zou et al. (2019). In particular, in the case where L_π is large but r_{max} is small, our result applies but Zou et al. (2019) do not. More importantly, due to the existence of $L_\pi C_\Gamma$, L_π in Zou et al. (2019) is at most the order of $\mathcal{O}\left(\frac{1}{C_\Gamma}\right)$. Since practitioners usually use a very large, possibly infinite, ball for projection, the L_π in Zou et al. (2019) has to be really small. By contrast, our $L_w = \mathcal{O}\left(L_\pi r_{max}\right)$ does *not* have any dependence on C_Γ. In other words, if C_Γ is really large, our result applies to much more L_π than Zou et al. (2019). We regard the broader settings as an improvement, and thus a contribution, over Zou et al. (2019). To our best knowledge, we do not know how to make Zou et al. (2019) work in such broader settings.

A concurrent work (Gopalan & Thoppe, 2022) prove that the iterates in their linear SARSA is bounded almost surely. In particular, they do not have projection but their result is only asymptotic without finite sample analysis. The most significant restriction is that they require i.i.d. samples, i.e., at each time step, the state S_t is assumed to be sampled from the stationary distribution $d_{\pi_{w_t}}$ of the current policy π_{w_t}. Since the policy w_t changes rapidly every time step, we argue that such i.i.d. samples are hard to obtain in practice.

Our results regarding the finite sample analysis of the general stochastic approximation algorithm in Section 3 rely on the pseudo contraction property and follow from Chen et al. (2021); Zhang et al. (2022). Another family of convergent results regarding stochastic approximation algorithms is usually based on the analysis of the corresponding ordinary differential equations (see, e.g., Benveniste et al. (1990); Kushner & Yin (2003); Borkar (2009)). See Chen et al. (2021) and references therein for more details.

SARSA is an extension of TD for control. The convergence of linear TD, which aims at estimating the value of a *fixed* policy, is an active research area, see, e.g., Tsitsiklis & Roy (1996); Dalal et al. (2018); Lakshminarayanan & Szepesvári (2018); Bhandari et al. (2018); Srikant & Ying (2019). Analyzing SARSA is more challenging than TD because the policy SARSA considers *changes* every step.

SARSA is an incremental and stochastic way to implement approximate policy iteration. Other variants of approximate policy iteration include Lagoudakis & Parr (2003); Antos et al. (2008); Farahmand et al. (2010); Lazaric et al. (2012; 2016).
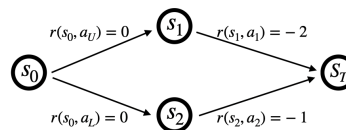
6. Experiments

Figure 1. A diagnostic MDP from Gordon (1996). The state s_0 is the initial state with two actions a_U and a_L available, both of which yield a 0 reward. At s_1, only one action a_1 is available, which yields a reward -2. At s_2, the action a_2 yields a reward -1. Both a_1 and a_2 leads to the terminal state S_T.

We use a diagnostic MDP from Gordon (1996) (Figure 1) to illustrate the chattering of linear SARSA. Gordon (1996) tested the ϵ-greedy policy (2), which is not continuous. We further test the ϵ-softmax policy (3), whose Lipschitz constant is inversely proportional to the temperature ι. When ι approaches 0, the ϵ-softmax policy approaches the ϵ-greedy policy. We run Algorithm 1 in this MDP with $C_\Gamma = \infty$, i.e., there is no projection. Following Gordon (1996), we set $\epsilon = 0.1, \gamma = 1.0$, and $\alpha_t = 0.01 \,\forall t$. As discussed in Gordon (1996), using a smaller discount factor or a decaying learning rate only slows down the chattering but the chattering always occurs. Following Gordon (1996), we use
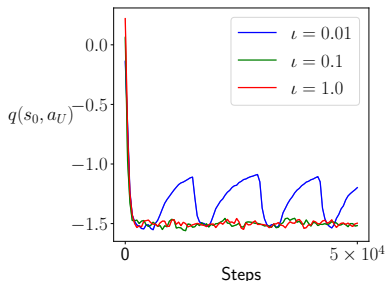
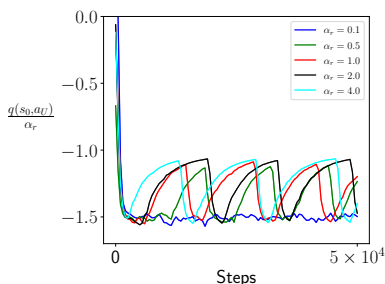*Figure 2.* The action value of a_U during training under different temperatures.

Figure 3. The α_r-normalized action value of a_U during training with a fixed temperature $\iota = 0.01$. The reward of the MDP in Figure 1 is scaled via α_r, e.g., the reward for the action a_1 is now $-2\alpha_r$.

the following feature function:

$$x(s_0, a_U) = [1, 0, 0]^\top, \ x(s_0, a_L) = [0, 1, 0]^\top,$$
$$x(s_1, a_1) = x(s_2, a_2) = [0, 0, 1]^\top.$$

In other words, it is essentially state aggregation.

As shown in Figure 2, when the temperature is small (i.e., $\iota = 0.01$), linear SARSA chatters. We further fix ι to be 0.01 and test reward with different magnitudes. To this end, we multiply the reward with a multiplier α_r. We stress that this is just a simple way to get MDPs with rewards of different magnitudes. It does *not* mean that one should artificially scale the reward down when Theorem 4.4 does not apply. As shown in Figure 3, the chattering behavior disappears with $\alpha_r = 0.1$. This suggests that our results might be improved such that when the magnitude of the rewards is small enough we can also achieve convergence to a fixed point, instead of a bounded region. We, however, leave this for future work. When we set $\alpha_r = 4.0$, the iterates still only chatter but do not diverge. This suggests that our requirement for r_{max} might be only sufficient and not necessary. We, however, leave the development of a necessary condition for future work. All the curves in Figures 2 and 3 are from a single run. Due to the randomness in the policy and the initialization of the weight, we find the peaks and valleys can sometimes average each other out when we average over multiple runs.

7. Conclusion

The behavior of linear SARSA is a long-standing open problem in the RL community. Despite the progress made in this work, there are still many open problems regarding the behavior of linear SARSA. To name a few: how does linear SARSA behave if the policy improvement operator is merely continuous but not Lipschitz continuous? How does linear SARSA behave if both the Lipschitz constant and the magnitude of the rewards are not small? Can we get a convergence rate without using any projection? We hope this work can draw more attention to the convergence of linear SARSA, arguably one of the most fundamental RL algorithms.

Acknowledgements

References

Antos, A., Szepesvári, C., and Munos, R. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Mach. Learn.*, 2008.

Baird, L. C. Residual algorithms: Reinforcement learning with function approximation. In *Machine Learning, Proceedings of the Twelfth International Conference on Machine Learning, Tahoe City, California, USA, July 9-12, 1995*, 1995.

Bellemare, M. G., Naddaf, Y., Veness, J., and Bowling, M. The arcade learning environment: An evaluation platform for general agents. *J. Artif. Intell. Res.*, 2013.

Benveniste, A., Métivier, M., and Priouret, P. *Adaptive Algorithms and Stochastic Approximations*, volume 22 of *Applications of Mathematics*. Springer, 1990. ISBN 978-3-642-75896-6. doi: 10.1007/978-3-642-75894-2. URL https://doi.org/10.1007/978-3-642-75894-2.

Bertsekas, D. P. and Tsitsiklis, J. N. *Neuro-Dynamic Programming*. Athena Scientific Belmont, MA, 1996.

Bhandari, J., Russo, D., and Singal, R. A finite time analysis of temporal difference learning with linear function approximation. In *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, 2018.

Borkar, V. S. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer, 2009.

Chen, Z., Maguluri, S. T., Shakkottai, S., and Shanmugam, K. A lyapunov theory for finite-sample guarantees of asynchronous q-learning and td-learning variants. *arXiv preprint arXiv:2102.01567*, 2021.

Dalal, G., Szörényi, B., Thoppe, G., and Mannor, S. Finite sample analyses for td (0) with function approximation. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

De Farias, D. P. and Van Roy, B. On the existence of fixed points for approximate value iteration and temporal-difference learning. *Journal of Optimization theory and Applications*, 105(3):589–608, 2000.

Farahmand, A. M., Munos, R., and Szepesvári, C. Error propagation for approximate policy and value iteration. In *Advances in Neural Information Processing Systems*, 2010.

Gopalan, A. and Thoppe, G. Approximate q-learning and sarsa (0) under the *epsilon*-greedy policy: a differential inclusion analysis. *arXiv preprint arXiv:2205.13617*, 2022.

Gordon, G. J. Chattering in sarsa (lambda)-a cmu learning lab internal report. 1996.

Gordon, G. J. Reinforcement learning with function approximation converges to a region. In *Advances in neural information processing systems*, pp. 1040–1046, 2001.

Konda, V. R. and Tsitsiklis, J. N. Actor-critic algorithms. In *Advances in Neural Information Processing Systems 12, [NIPS Conference, Denver, Colorado, USA, November 29 - December 4, 1999]*, 1999.

Kushner, H. and Yin, G. G. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

Lagoudakis, M. G. and Parr, R. Least-squares policy iteration. *J. Mach. Learn. Res.*, 2003.

Lakshminarayanan, C. and Szepesvári, C. Linear stochastic approximation: How far does constant step-size and iterate averaging go? In *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, 2018.

Lazaric, A., Ghavamzadeh, M., and Munos, R. Finite-sample analysis of least-squares policy iteration. *Journal of Machine Learning Research*, 13:3041–3074, 2012.

Lazaric, A., Ghavamzadeh, M., and Munos, R. Analysis of classification-based policy iteration algorithms. 2016.

Liang, Y., Machado, M. C., Talvitie, E., and Bowling, M. State of the art control of atari games using shallow reinforcement learning. *arXiv preprint arXiv:1512.01563*, 2015.

Marbach, P. and Tsitsiklis, J. N. Simulation-based optimization of markov reward processes. *IEEE Trans. Autom. Control.*, 2001.

Melo, F. S., Meyn, S. P., and Ribeiro, M. I. An analysis of reinforcement learning with function approximation. In *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, 2008.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T. P., Harley, T., Silver, D., and Kavukcuoglu, K. Asynchronous methods for deep reinforcement learning. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, 2016.

Perkins, T. J. and Precup, D. A convergent form of approximate policy iteration. In *Advances in Neural Information Processing Systems 15 [Neural Information Processing Systems, NIPS 2002, December 9-14, 2002, Vancouver, British Columbia, Canada]*, 2002.

Puterman, M. L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Robbins, H. and Monro, S. A stochastic approximation method. *The Annals of Mathematical Statistics*, 1951.

Rummery, G. A. and Niranjan, M. *On-line Q-learning using connectionist systems*. University of Cambridge, Department of Engineering Cambridge, UK, 1994.

Singh, S. P., Jaakkola, T. S., Littman, M. L., and Szepesvári, C. Convergence results for single-step on-policy reinforcement-learning algorithms. *Mach. Learn.*, 2000.

Srikant, R. and Ying, L. Finite-time error bounds for linear stochastic approximation andtd learning. In *Conference on Learning Theory*, pp. 2803–2830. PMLR, 2019.

Sutton, R. S. Learning to predict by the methods of temporal differences. *Mach. Learn.*, 1988.

Sutton, R. S. Open theoretical questions in reinforcement learning. In *European Conference on Computational Learning Theory*, pp. 11–17. Springer, 1999.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction (2nd Edition)*. MIT press, 2018.

Tsitsiklis, J. N. and Roy, B. V. Analysis of temporal-diffference learning with function approximation. In *Advances in Neural Information Processing Systems 9, NIPS, Denver, CO, USA, December 2-5, 1996*, 1996.

Watkins, C. J. C. H. *Learning from delayed rewards*. PhD thesis, King's College, Cambridge, 1989.

Wu, Y., Zhang, W., Xu, P., and Gu, Q. A finite-time analysis of two time-scale actor-critic methods. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Zhang, S., Yao, H., and Whiteson, S. Breaking the deadly triad with a target network. *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, 2021.

Zhang, S., des Combes, R. T., and Laroche, R. Global optimality and finite sample analysis of softmax off-policy actor critic under state distribution mismatch. *Journal of Machine Learning Research*, 2022.

Zou, S., Xu, T., and Liang, Y. Finite-sample analysis for SARSA with linear function approximation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 2019.

A. Proofs of Section 3

A.1. Proof of Theorem 3.9

Theorem 3.9. *Let Assumptions 3.1 - 3.7 hold. If t_0 is sufficiently large, then the iterates $\{u_t\}$ generated by (12) satisfy*

$$
\mathbb{E}\left[\left\|u_{t+1} - w^*_{\theta_{t+1}}\right\|^2\right]
$$
$$
\leq \left(1 - 2\left(1 - \kappa_{\alpha_t} - \mathcal{O}\left(\alpha_t^2 \log^2 \alpha_t\right)\right)\right) \mathbb{E}\left[\left\|\Gamma(u_t) - w^*_{\theta_t}\right\|^2\right]
$$
$$
+ 2L_w L_\theta \alpha_t \mathbb{E}\left[\left\|\Gamma(u_t) - w^*_{\theta_t}\right\|\right] + \mathcal{O}\left(\alpha_t^2 \log^2 \alpha_t\right),
$$

where $L_\theta \doteq U_F + (L_F + 1)C_\Gamma$.

Proof. We consider a Lyapunov function

$$
M(x) \doteq \frac{1}{2}\|x\|^2.
$$

It is well-known that for any x, x',

$$
M(x') \leq M(x) + \langle \nabla M(x), x' - x \rangle + \frac{1}{2}\|x - x'\|^2.
$$

Using $x' = u_{t+1} - w^*_{\theta_{t+1}}$ and $x = \Gamma(u_t) - w^*_{\theta_t}$ in the above inequality and recalling the update (12)

$$
u_{t+1} = \Gamma(u_t) + \alpha_t(F_{\theta_t}(\Gamma(u_t), Y_t) - \Gamma(u_t))
$$
$$
= f^{\alpha_t}_{\theta_t}(\Gamma(u_t)) + \alpha_t\left(F_{\theta_t}(\Gamma(u_t), Y_t) - \bar{F}_{\theta_t}(\Gamma(u_t))\right)
$$

yield

$$
\frac{1}{2}\left\|u_{t+1} - w^*_{\theta_{t+1}}\right\|^2 \tag{13}
$$
$$
\leq \frac{1}{2}\left\|\Gamma(u_t) - w^*_{\theta_t}\right\|^2 + \left\langle \Gamma(u_t) - w^*_{\theta_t}, u_{t+1} - \Gamma(u_t) + w^*_{\theta_t} - w^*_{\theta_{t+1}} \right\rangle
$$
$$
+ \frac{1}{2}\left\|u_{t+1} - \Gamma(u_t) + w^*_{\theta_t} - w^*_{\theta_{t+1}}\right\|^2
$$
$$
= \frac{1}{2}\left\|\Gamma(u_t) - w^*_{\theta_t}\right\|^2
$$
$$
+ \underbrace{\left\langle \Gamma(u_t) - w^*_{\theta_t}, w^*_{\theta_t} - w^*_{\theta_{t+1}} \right\rangle}_{T_1}
$$
$$
+ \underbrace{\left\langle \Gamma(u_t) - w^*_{\theta_t}, f^{\alpha_t}_{\theta_t}(\Gamma(u_t)) - \Gamma(u_t) \right\rangle}_{T_2}
$$
$$
+ \alpha_t \underbrace{\left\langle \Gamma(u_t) - w^*_{\theta_t}, F_{\theta_t}(\Gamma(u_t), Y_t) - \bar{F}_{\theta_t}(\Gamma(u_t)) \right\rangle}_{T_3}
$$
$$
+ \alpha_t^2 \underbrace{\left\|F_{\theta_t}(\Gamma(u_t), Y_t) - \Gamma(u_t)\right\|^2}_{T_5}
$$
$$
+ \underbrace{\left\|w^*_{\theta_t} - w^*_{\theta_{t+1}}\right\|^2}_{T_6}.
$$

Here we do not have T_4 because the counterpart in Zhang et al. (2022) is now 0. To further decompose T_3, we define a function τ_α of α as

$$
\tau_\alpha \doteq \min\{n \geq 0 \mid C_M \tau^n \leq \alpha\},
$$

where the constants C_M and τ are given in Lemma 3.3. In particular, τ_{α_t} denotes the number of steps the chain needs to mix to an accuracy of α_t. It is easy to see

$$\tau_\alpha = \mathcal{O}\left(-\log\alpha\right), \ \lim_{\alpha\to 0}\alpha\tau_\alpha = 0. \tag{14}$$

We now decompose T_3 as

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

We further decompose T_{33} as

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

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 $P_{\theta_{t-\tau_{\alpha_t}}}, P_{\theta_{k-\tau_{\alpha_t}+1}}, \dots$.

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

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

$$A \doteq 2L_F + 1, \quad B \doteq U_F, \quad C \doteq AU_w + B + A + A(1 + U_F' + U_F''), \ \alpha_{i:j} \doteq \sum_{t=i}^{j}\alpha_t. \tag{15}$$

According to (14), we can select a sufficiently large t_0 such that

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

holds for all t. This condition is crucial for Lemma C.2, which plays an important role in the following bounds.

Lemma A.1. *(Bound of T_1)*

$$T_1 \leq L_w L_\theta \alpha_t \big\| \Gamma(u_t) - w_{\theta_t}^* \big\|.$$

The proof of Lemma A.1 is provided in Section D.1.

Lemma A.2. *(Bound of T_2)*

$$T_2 \leq -(1 - \kappa_{\alpha_t}) \big\| \Gamma(u_t) - w_{\theta_t}^* \big\|^2.$$

The proof of Lemma A.2 is provided in Section D.2.

Lemma A.3. *(Bound of T_{31})*

$$T_{31} \leq 8(L_w L_\theta + 1)\alpha_{t-\tau_{\alpha_t}, t-1} \left(A^2 \big\| \Gamma(u_t) - w_{\theta_t}^* \big\|^2 + C^2 \right).$$

The proof of Lemma A.3 is provided in Section D.3.

Lemma A.4. *(Bound of T_{32})*

$$T_{32} \leq 16\alpha_{t-\tau_{\alpha_t}, t-1}(1 + L_w L_\theta \alpha_{t-\tau_{\alpha_t}, t-1}) \left(A^2 \big\| \Gamma(u_t) - w_{\theta_t}^* \big\|^2 + C^2 \right).$$

The proof of Lemma A.4 is provided in Section D.4.

Lemma A.5. *(Bound of T_{331})*

$$\mathbb{E}\left[T_{331}\right] \leq \frac{8\alpha_t(1 + L_w L_\theta \alpha_{t-\tau_{\alpha_t}, t-1})}{A} \left(A^2 \mathbb{E}\left[\big\| \Gamma(u_t) - w_{\theta_t}^* \big\|^2 \right] + C^2 \right).$$

The proof of Lemma A.5 is provided in Section D.5.

Lemma A.6. *(Bound of T_{332})*

$$\mathbb{E}\left[T_{332}\right] \leq \frac{8|\mathcal{Y}|L_P L_\theta \sum_{j=t-\tau_{\alpha_t}}^{t-1} \alpha_{t-\tau_{\alpha_t}, j}(1 + L_w L_\theta \alpha_{t-\tau_{\alpha_t}, t-1})}{A} \left(A^2 \mathbb{E}\left[\big\| \Gamma(u_t) - w_{\theta_t}^* \big\|^2 \right] + C^2 \right).$$

The proof of Lemma A.6 is provided in Section D.6.

Lemma A.7. *(Bound of T_{333})*

$$T_{333} \leq \frac{4L_F' L_\theta \alpha_{t-\tau_{\alpha_t}, t-1}(1 + L_w L_\theta \alpha_{t-\tau_{\alpha_t}, t-1})}{A^2} \left(A^2 \big\| \Gamma(u_t) - w_{\theta_t}^* \big\|^2 + C^2 \right).$$

The proof of Lemma A.7 is provided in Section D.7.

Lemma A.8. *(Bound of T_{334})*

$$T_{334} \leq \frac{4L_F'' L_\theta \alpha_{t-\tau_{\alpha_t}, t-1}(1 + L_w L_\theta \alpha_{t-\tau_{\alpha_t}, t-1})}{A^2} \left(A^2 \big\| \Gamma(u_t) - w_{\theta_t}^* \big\|^2 + C^2 \right).$$

The proof of Lemma A.8 is provided in Section D.8.

Lemma A.9. *(Bound of T_5)*

$$T_5 \leq 2 \left(A^2 \big\| \Gamma(u_t) - w_{\theta_t}^* \big\|^2 + C^2 \right).$$

The proof of Lemma A.9 is provided in Section D.9.

Lemma A.10. *(Bound of T_6)*

$$T_6 = \left\| w_{\theta_t}^* - w_{\theta_{t+1}}^* \right\|^2 \le L_w^2 L_\theta^2 \alpha_t^2.$$

Lemma A.10 follows immediately from Lemma C.5.

We now assemble the bounds in Lemmas A.1 - A.10 back into (13). Define

$$L_{\alpha,t} \doteq \left(\sum_{j=t-\tau_{\alpha_t}}^t \alpha_{t-\tau_{\alpha_t},j} \right) (1 + L_w L_\theta \max\{1, \alpha_{t-\tau_{\alpha_t},t}\}),$$

$$C_0 \doteq \max\{A^2, C^2\} \max\{16, L_w^2 L_\theta^2, 8|\mathcal{Y}|L_P L_\theta, 4L_F' L_\theta, 4L_F'' L_\theta\},$$

Using $A > 1$ and Lemmas A.3 - A.10, it is easy to see

$$\mathbb{E}[T_3] \le C_0 L_{\alpha,t} \left(\mathbb{E}\left[\left\| \Gamma(u_t) - w_{\theta_t}^* \right\|^2\right] + 1 \right),$$

$$\alpha_t \mathbb{E}[T_3] \le C_0 \alpha_t L_{\alpha,t} \left(\mathbb{E}\left[\left\| \Gamma(u_t) - w_{\theta_t}^* \right\|^2\right] + 1 \right),$$

$$\mathbb{E}[T_5] \le C_0 \left(\mathbb{E}\left[\left\| \Gamma(u_t) - w_{\theta_t}^* \right\|^2\right] + 1 \right),$$

$$\alpha_t^2 \mathbb{E}[T_5] \le C_0 \alpha_t^2 \left(\mathbb{E}\left[\left\| \Gamma(u_t) - w_{\theta_t}^* \right\|^2\right] + 1 \right),$$

$$\mathbb{E}[T_6] \le C_0 \alpha_t^2.$$

Then we have

$$\frac{1}{2}\mathbb{E}\left[\left\| u_{t+1} - w_{\theta_{t+1}}^* \right\|^2\right]$$

$$\le \frac{1}{2}\mathbb{E}\left[\left\| \Gamma(u_t) - w_{\theta_t}^* \right\|^2\right] + L_w L_\theta \alpha_t \mathbb{E}\left[\left\| \Gamma(u_t) - w_{\theta_t}^* \right\|\right] - (1 - \kappa_{\alpha_t})\mathbb{E}\left[\left\| \Gamma(u_t) - w_{\theta_t}^* \right\|^2\right]$$

$$+ C_0 \alpha_t L_{\alpha,t} \left(\mathbb{E}\left[\left\| \Gamma(u_t) - w_{\theta_t}^* \right\|^2\right] + 1 \right) + C_0 \alpha_t^2 \left(\mathbb{E}\left[\left\| \Gamma(u_t) - w_{\theta_t}^* \right\|^2\right] + 1 \right)$$

$$+ C_0 \alpha_t^2,$$

implying

$$\mathbb{E}\left[\left\| u_{t+1} - w_{\theta_{t+1}}^* \right\|^2\right] \le \left(1 - 2\left(1 - \kappa_{\alpha_t} - C_0 \alpha_t L_{\alpha,t} - C_0 \alpha_t^2 \right) \right) \mathbb{E}\left[\left\| \Gamma(u_t) - w_{\theta_t}^* \right\|^2\right]$$

$$+ 2L_w L_\theta \alpha_t \mathbb{E}\left[\left\| \Gamma(u_t) - w_{\theta_t}^* \right\|\right] + 2C_0 \alpha_t L_{\alpha,t} + 4C_0 \alpha_t^2.$$

Observing that

$$L_{\alpha,t} = \mathcal{O}\left(\alpha_t \log^2 \alpha_t \right)$$

then completes the proof. $\square$

A.2. Proof of Corollary 3.10

Corollary A.11. *Let Assumptions 3.1 - 3.7 hold. Assume $\kappa_\alpha = \sqrt{1 - \eta\alpha}$ for some positive constant $\eta > 0$. If t_0 is sufficiently large, then the iterates $\{w_t\}$ generated by (11) satisfy*

$$\mathbb{E}\left[\left\| w_t - w_{w_t}^* \right\|^2\right] = \frac{72 L_w^2 L_\theta^2}{\eta^2}$$

$$+ \begin{cases} \mathcal{O}\left(t^{-\frac{\eta c_\alpha}{3}} \log^2 t \right), & \epsilon_\alpha = 1, \eta c_\alpha \in (0,3) \\ \mathcal{O}\left(\frac{\log^3 t}{t} \right), & \epsilon_\alpha = 1, \eta c_\alpha = 3 \\ \mathcal{O}\left(\frac{\log^2 t}{t} \right), & \epsilon_\alpha = 1, \eta c_\alpha \in (3,\infty) \\ \mathcal{O}\left(\frac{\log^2 t}{t^{\epsilon_\alpha}} \right), & \epsilon_\alpha \in (0,1) \end{cases}.$$

Proof. According to Theorem 3.9, we have

$$\mathbb{E}\left[\left\|u_{t+1} - w_{\theta_{t+1}}^*\right\|^2\right] \leq \left(1 - 2\left(1 - \kappa_{\alpha_t} - C_0\alpha_t L_{\alpha,t} - C_0\alpha_t^2\right)\right)\mathbb{E}\left[\left\|\Gamma(u_t) - w_{\theta_t}^*\right\|^2\right]$$
$$+ 2L_w L_\theta \alpha_t \mathbb{E}\left[\left\|\Gamma(u_t) - w_{\theta_t}^*\right\|\right] + 2C_0\alpha_t L_{\alpha,t} + 4C_0\alpha_t^2.$$

Since $L_{\alpha,t} = \mathcal{O}\left(\alpha_t \log^2 \alpha_t\right)$, we conclude that there exists a constant $C_1 > 0$ such that

$$\mathbb{E}\left[\left\|u_{t+1} - w_{\theta_{t+1}}^*\right\|^2\right] \leq \left(1 - 2\left(1 - \sqrt{1 - \eta\alpha_t} - C_1\alpha_t^2 \log^2 \alpha_t\right)\right)\mathbb{E}\left[\left\|\Gamma(u_t) - w_{\theta_t}^*\right\|^2\right]$$
$$+ 2L_w L_\theta \alpha_t \mathbb{E}\left[\left\|\Gamma(u_t) - w_{\theta_t}^*\right\|\right] + C_1\alpha_t^2 \log^2 \alpha_t.$$

When t_0 is sufficiently large, we have $\forall t$,

$$1 - 2\left(1 - \sqrt{1 - \eta\alpha_t} - C_1\alpha_t^2 \log^2 \alpha_t\right) > 0.$$

Using

$$\left\|\Gamma(u_t) - w_{\theta_t}^*\right\| = \left\|\Gamma(u_t) - \Gamma(w_{\theta_t}^*)\right\| \leq \left\|u_t - w_{\theta_t}^*\right\|$$

then yields

$$\mathbb{E}\left[\left\|u_{t+1} - w_{\theta_{t+1}}^*\right\|^2\right]$$
$$\leq \left(1 - 2\left(1 - \sqrt{1 - \eta\alpha_t} - C_1\alpha_t^2 \log^2 \alpha_t\right)\right)\mathbb{E}\left[\left\|u_t - w_{\theta_t}^*\right\|^2\right]$$
$$+ 2L_w L_\theta \alpha_t \mathbb{E}\left[\left\|u_t - w_{\theta_t}^*\right\|\right] + C_1\alpha_t^2 \log^2 \alpha_t$$
$$\leq \left(1 - 2\left(1 - \sqrt{1 - \eta\alpha_t} - C_1\alpha_t^2 \log^2 \alpha_t\right)\right)\mathbb{E}\left[\left\|u_t - w_{\theta_t}^*\right\|^2\right]$$
$$+ 2L_w L_\theta \alpha_t \sqrt{\mathbb{E}\left[\left\|u_t - w_{\theta_t}^*\right\|^2\right]} + C_1\alpha_t^2 \log^2 \alpha_t \quad \text{(Jensen's inequality)} \quad .$$

Since

$$\lim_{\alpha \to 0} \frac{1 - \sqrt{1 - \eta\alpha}}{\frac{\eta}{2}\alpha} = 1,$$

we conclude that when t_0 is sufficiently large, $\forall t$,

$$1 - \sqrt{1 - \eta\alpha_t} - C_1\alpha_t^2 \log^2 \alpha_t \geq \frac{\eta}{3}\alpha_t - C_1\alpha_t^2 \log^2 \alpha_t^2 \geq \frac{\eta}{4}\alpha_t.$$

With

$$z_t \doteq \sqrt{\mathbb{E}\left[\left\|u_t - w_{\theta_t}^*\right\|^2\right]},$$

we then get

$$z_{t+1}^2 \leq \left(1 - \frac{\eta}{2}\alpha_t\right)z_t^2 + 2L_w L_\theta \alpha_t z_t + C_1\alpha_t^2 \log^2 \alpha_t.$$

If

$$\left(1 - \frac{\eta}{2}\alpha_t\right)z_t^2 + 2L_w L_\theta \alpha_t z_t \leq \left(1 - \frac{\eta}{3}\alpha_t\right)z_t^2,$$
$$\iff \frac{12L_w L_\theta}{\eta} \leq z_t, \tag{16}$$

we have

$$z_{t+1}^2 \le (1 - \frac{\eta}{3}\alpha_t)z_t^2 + C_1\alpha_t^2 \log^2 \alpha_t.$$

If

$$\frac{12L_w L_\theta}{\eta} \ge z_t, \tag{17}$$

we have

$$z_{t+1}^2 \le \left(1 - \frac{\eta}{2}\alpha_t\right) z_t^2 + 2L_w L_\theta \alpha_t z_t + C_1\alpha_t^2 \log^2 \alpha_t$$
$$\le \left(1 - \frac{\eta}{2}\alpha_t\right) z_t^2 + \frac{24L_w^2 L_\theta^2}{\eta}\alpha_t + C_1\alpha_t^2 \log^2 \alpha_t.$$

Since for any time t, one of (16) and (17) must hold, we always have

$$z_{t+1}^2 \le \left(1 - \frac{\eta}{3}\alpha_t\right) z_t^2 + \frac{24L_w^2 L_\theta^2}{\eta}\alpha_t + C_1\alpha_t^2 \log^2 \alpha_t. \tag{18}$$

Telescoping the above inequality from t_0 to t yields

$$z_t^2 \le \underbrace{\prod_{i=t_0}^{t-1} \left(1 - \frac{\eta}{3}\alpha_i\right) z_{t_0}^2}_{E_1} + \frac{24L_w^2 L_\theta^2}{\eta} \underbrace{\sum_{i=t_0}^{t-1} \prod_{j=i+1}^{t-1} \left(1 - \frac{\eta}{3}\alpha_j\right) \alpha_i}_{E_2}$$
$$+ C_1 \underbrace{\sum_{i=t_0}^{t-1} \prod_{j=i+1}^{t-1} \left(1 - \frac{\eta}{3}\alpha_j\right) \alpha_i^2 \log^2 \alpha_i}_{E_3},$$

where we adopt the convention that $\prod_{x=i}^j (\cdot) = 1$ if $i > j$. For E_1, using $1 + x \le \exp x$ yields

$$E_1 \le \exp\left(-\frac{\eta}{3}\sum_{i=t_0}^{t-1}\alpha_i\right) \le \exp\left(-\frac{\eta}{3}\int_{x=t_0}^t \frac{c_\alpha}{(t_0 + x)^{\epsilon_\alpha}}dx\right) \tag{19}$$
$$= \begin{cases} \left(\frac{2t_0}{t_0+t}\right)^{\frac{\eta c_\alpha}{3}}, & \epsilon_\alpha = 1 \\ \exp\left(\frac{\eta c_\alpha}{3(1-\epsilon_\alpha)}\left((2t_0)^{1-\epsilon_\alpha} - (t_0 + t)^{1-\epsilon_\alpha}\right)\right), & \epsilon_\alpha \in (0, 1) \end{cases}.$$

For E_2, define

$$B_t \doteq \sum_{i=0}^t \prod_{j=i+1}^t \left(1 - \frac{\eta}{3}\alpha_j\right) \alpha_i.$$

Then we have

$$B_t = \alpha_t + (1 - \frac{\eta}{3}\alpha_t)B_{t-1}.$$

When t_0 is sufficiently large such that

$$1 - \frac{\eta}{3}\alpha_t > 0,$$

it is easy to see

$$B_{t-1} \le \frac{3}{\eta} \implies B_t \le \frac{3}{\eta}.$$

As $B_0 = \alpha_0$, we have $B_0 < \frac{3}{\eta}$ for sufficiently large t_0. We, therefore, conclude by induction that $\forall t$,

$$B_t \leq \frac{3}{\eta}.$$

Consequently,

$$E_2 \leq B_{t-1} \leq \frac{3}{\eta}. \tag{20}$$

For E_3, we have

$$E_3 \leq \log^2 \alpha_t \underbrace{\sum_{i=t_0}^{t-1} \prod_{j=i+1}^{t-1} \left(1 - \frac{\eta}{3}\alpha_j\right) \alpha_i^2}_{E_4}. \tag{21}$$

If $\epsilon_\alpha = 1$, we have

$$
\begin{aligned}
E_4 &\leq \sum_{i=t_0}^{t-1} \exp\left(-\frac{\eta}{3}\int_{x=i+1}^{t} \frac{c_\alpha}{(t_0+x)^{\epsilon_\alpha}}dx\right) \frac{c_\alpha^2}{(t_0+i)^{2\epsilon_\alpha}} \\
&= \sum_{i=t_0}^{t-1} \left(\frac{t_0+i+1}{t+t_0}\right)^{\frac{\eta c_\alpha}{3}} \frac{c_\alpha^2}{(t_0+i)^2} \\
&= \sum_{i=t_0}^{t-1} \left(\frac{t_0+i+1}{t+t_0}\right)^{\frac{\eta c_\alpha}{3}} \frac{c_\alpha^2}{(t_0+i+1)^2} \left(\frac{t_0+i+1}{t_0+i}\right)^2 \\
&\leq \frac{4c_\alpha^2}{(t+t_0)^{\frac{\eta c_\alpha}{3}}} \sum_{i=t_0}^{t-1} \frac{1}{(t_0+i+1)^{2-\frac{\eta c_\alpha}{3}}} \\
&= \begin{cases} \mathcal{O}\left(t^{-\frac{\eta c_\alpha}{3}}\right), & \eta c_\alpha \in (0,3) \\ \mathcal{O}\left(\frac{\log t}{t}\right), & \eta c_\alpha = 3 \\ \mathcal{O}\left(\frac{1}{t}\right), & \eta c_\alpha \in (3,\infty) \end{cases}.
\end{aligned}
$$

If $\epsilon_\alpha \in (0,1)$, when t_0 is sufficiently large, we can use induction (see, e.g., Section A.3.7 of Chen et al. (2021)) to show that

$$E_4 = \mathcal{O}\left(\frac{1}{t^{\epsilon_\alpha}}\right).$$

Putting the bounds in (19), (20), and (21) back into (18) yields

$$z_t^2 = \frac{72L_w^2 L_\theta^2}{\eta^2} + \begin{cases} \mathcal{O}\left(t^{-\frac{\eta c_\alpha}{3}}\log^2 t\right), & \epsilon_\alpha = 1, \eta c_\alpha \in (0,3) \\ \mathcal{O}\left(\frac{\log^3 t}{t}\right), & \epsilon_\alpha = 1, \eta c_\alpha = 3 \\ \mathcal{O}\left(\frac{\log^2 t}{t}\right), & \epsilon_\alpha = 1, \eta c_\alpha \in (3,\infty) \\ \mathcal{O}\left(\frac{\log^2 t}{t^{\epsilon_\alpha}}\right), & \epsilon_\alpha \in (0,1) \end{cases}.$$

Here we have used the fact that E_3 always dominates E_1 for any $\epsilon_\alpha, c_\alpha$. Using

$$\mathbb{E}\left[\left\|w_t - w_{w_t}^*\right\|^2\right] = \mathbb{E}\left[\left\|\Gamma(u_t) - \Gamma(w_{w_t}^*)\right\|^2\right] \leq \mathbb{E}\left[\left\|u_t - w_{w_t}^*\right\|^2\right] = z_t^2$$

then completes the proof. $\qquad \square$

B. Proofs of Section 4

B.1. Proof of Theorem 4.4

Theorem 4.4. *Let Assumptions 3.7 and 4.1 - 4.3 hold. Assume $\|X\| = 1$ and r_{max} is not so large such that*

$$L_w \doteq \mathcal{O}\left(L_\pi r_{max}\right) < 1.$$

Assume C_Γ is large enough such that

$$U_w \doteq \mathcal{O}\left(r_{max}\right) \le C_\Gamma.$$

Let t_0 be sufficiently large. Then the iterates $\{w_t\}$ generated by Algorithm 1 satisfy

$$\mathbb{E}\left[\|w_t - w_*\|\right] = \frac{6\sqrt{2}L_w\left(1 + 4C_\Gamma\right)}{\eta(1 - L_w)}$$
$$+ \begin{cases} \mathcal{O}\left(t^{-\frac{\eta c_\alpha}{6}} \log t\right), & \epsilon_\alpha = 1, \eta c_\alpha \in (0, 3) \\ \mathcal{O}\left(t^{-\frac{1}{2}} \log^{\frac{3}{2}} t\right), & \epsilon_\alpha = 1, \eta c_\alpha = 3 \\ \mathcal{O}\left(t^{-\frac{1}{2}} \log t\right), & \epsilon_\alpha = 1, \eta c_\alpha \in (3, \infty) \\ \mathcal{O}\left(t^{-\frac{\epsilon_\alpha}{2}} \log t\right), & \epsilon_\alpha \in (0, 1) \end{cases},$$

where η is a positive constant and w_* is the unique vector such that $e(w_*) = 0$.

Proof. To start with, define

$$\mathcal{Y} \doteq \{(s, a, s', a') \mid s \in \mathcal{S}, a \in \mathcal{A}, s' \in \mathcal{S}, p(s'|s, a) > 0\},$$
$$Y_t \doteq (S_t, A_t, S_{t+1}, A_{t+1}),$$
$$y \doteq (s, a, s', a'),$$
$$\tilde{\pi}_\theta(a|s) \doteq \pi_{\Gamma(\theta)}(a|s),$$
$$P_\theta((s_1, a_1, s_1', a_1'), (s_2, a_2, s_2', a_2')) \doteq \begin{cases} 0 & (s_1', a_1') \ne (s_2, a_2) \\ p(s_2'|s_2, a_2)\tilde{\pi}_\theta(a_2'|s_2') & (s_1', a_1') = (s_2, a_2) \end{cases},$$
$$F_\theta(w, y) \doteq \left(r(s, a) + \gamma x(s', a')^\top w - x(s, a)^\top w\right) x(s, a) + w.$$

Here our $F_\theta(w, y)$ is actually independent of θ.

The update of $\{w_t\}$ in Algorithm 1 with $\lambda = 0$ can then be expressed as

$$w_{t+1} = \Gamma\left(w_t + \alpha_t\left(F_{\theta_t}(w_t, Y_t) - w_t\right)\right).$$

According to the action selection rule for A_{t+1} 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 4.2. In particular, for any θ, the invariant distribution of the chain induced by P_θ is $d_{\tilde{\pi}_\theta}(s)\tilde{\pi}_\theta(a|s)p(s'|s, a)\tilde{\pi}_\theta(a'|s')$.

For Assumption 3.4, it is easy to see

$$\bar{F}_\theta(w) = X^\top D_{\tilde{\pi}_\theta}(\gamma P_{\tilde{\pi}_\theta} - I)Xw + X^\top D_{\tilde{\pi}_\theta}r + w,$$
$$f_\theta^\alpha(w) = w + \alpha\left(X^\top D_{\tilde{\pi}_\theta}(\gamma P_{\tilde{\pi}_\theta} - I)Xw + X^\top D_{\tilde{\pi}_\theta}r\right).$$

Define

$$w_\theta^* \doteq -\left(X^\top D_{\tilde{\pi}_\theta}(\gamma P_{\tilde{\pi}_\theta} - I)X\right)^{-1} X^\top D_{\tilde{\pi}_\theta}r.$$

It is then easy to see that w_θ^* is the unique fixed point of $\bar{F}_\theta(w)$. The uniform pseudo-contraction is verified by Lemma C.6. In particular, we have

$$
\begin{aligned}
\kappa_\alpha &\doteq \sqrt{1 - \eta\alpha}, \\
\eta &\doteq (1 - \gamma) \inf_\theta \lambda_{min}\left(X^\top D_{\pi_\theta} X\right) > 0,
\end{aligned}
\tag{22}
$$

where $\lambda_{min}(\cdot)$ denotes the minimum eigenvalue of a symmetric positive definite matrix.

We now verify Assumption 3.5. To verify Assumption 3.5 (i), we have

$$
\begin{aligned}
&\|F_\theta(w, y) - F_\theta(w', y)\| \\
\leq &\left\|\gamma x(s,a)x(s',a')^\top - x(s,a)x(s,a)^\top + I\right\|\|w - w'\| \\
\leq &\underbrace{\left((1+\gamma)x_{max}^2 + 1\right)}_{L_F}\|w - w'\|.
\end{aligned}
$$

Assumption 3.5 (ii) immediately holds since our $F_\theta(w, y)$ is independent of θ.

To verify Assumption 3.5 (iii), we have

$$
\|F_\theta(0, y)\| = \|r(s,a)x(s,a)\| \leq \underbrace{r_{max}x_{max}}_{U_F}.
$$

To verify Assumption 3.5 (iv), we have

$$
\begin{aligned}
&\left\|\bar{F}_\theta(w) - \bar{F}_{\theta'}(w)\right\|_\infty \\
= &\left\|X^\top\left(D_{\tilde{\pi}_\theta}(\gamma P_{\tilde{\pi}_\theta} - I) - D_{\tilde{\pi}_{\theta'}}(\gamma P_{\tilde{\pi}_{\theta'}} - I)\right)Xw + X^\top\left(D_{\tilde{\pi}_\theta} - D_{\tilde{\pi}_{\theta'}}\right)r\right\|_\infty \\
\leq &\|X\|_\infty^2\left\|D_{\tilde{\pi}_\theta}(\gamma P_{\tilde{\pi}_\theta} - I) - D_{\tilde{\pi}_{\theta'}}(\gamma P_{\tilde{\pi}_{\theta'}} - I)\right\|_\infty\|w\|_\infty + \|X\|_\infty\left\|D_{\tilde{\pi}_\theta} - D_{\tilde{\pi}_{\theta'}}\right\|_\infty\|r\|_\infty.
\end{aligned}
$$

Lemma C.3 asserts that D_{π_θ} is Lipschitz continuous in θ. We then conclude, by Lemma C.1, that there exist positive constants $L_{DP} > 0, L_D > 0$ such that

$$
\begin{aligned}
\left\|D_{\pi_\theta}(\gamma P_{\pi_\theta} - I) - D_{\pi_{\theta'}}(\gamma P_{\pi_{\theta'}} - I)\right\|_\infty &\leq L_{DP}L_\pi\|\theta - \theta\|_\infty, \\
\left\|D_{\pi_\theta} - D_{\pi_{\theta'}}\right\|_\infty &\leq L_D L_\pi\|\theta - \theta'\|_\infty.
\end{aligned}
$$

Importantly, L_{DP} and L_D do not depend on C_Γ. It is then easy to see that

$$
\begin{aligned}
\left\|\bar{F}_\theta(w) - \bar{F}_{\theta'}(w)\right\|_\infty &\leq \left(\|X\|_\infty^2 L_{DP}L_\pi\|w\|_\infty + \|X\|_\infty\|r\|_\infty L_D L_\pi\right)\|\Gamma(\theta) - \Gamma(\theta')\|_\infty \\
&\leq \left(\|X\|_\infty^2 L_{DP}L_\pi\|w\| + \|X\|_\infty\|r\|_\infty L_D L_\pi\right)\|\Gamma(\theta) - \Gamma(\theta')\| \\
&\leq \left(\|X\|_\infty^2 L_{DP}L_\pi\|w\| + \|X\|_\infty\|r\|_\infty L_D L_\pi\right)\|\theta - \theta'\|.
\end{aligned}
$$

It follow immediately that

$$
\left\|\bar{F}_\theta(w) - \bar{F}_{\theta'}(w)\right\| \leq \sqrt{K}\left(\|X\|_\infty^2 L_{DP}L_\pi\|w\| + \|X\|_\infty\|r\|_\infty L_D L_\pi\right)\|\theta - \theta'\|.
$$

To verify Assumption 3.5 (v), we first use Lemma C.4 to get

$$
\begin{aligned}
&\left\|\left(X^\top D_{\tilde{\pi}_\theta}(\gamma P_{\tilde{\pi}_\theta} - I)X\right)^{-1} - \left(X^\top D_{\tilde{\pi}_{\theta'}}(\gamma P_{\tilde{\pi}_{\theta'}} - I)X\right)^{-1}\right\|_\infty \\
\leq &\left\|\left(X^\top D_{\tilde{\pi}_\theta}(\gamma P_{\tilde{\pi}_\theta} - I)X\right)^{-1}\right\|_\infty\left\|\left(X^\top D_{\tilde{\pi}_{\theta'}}(\gamma P_{\tilde{\pi}_{\theta'}} - I)X\right)^{-1}\right\|_\infty \\
&\times \left\|X^\top D_{\tilde{\pi}_\theta}(\gamma P_{\tilde{\pi}_\theta} - I)X - X^\top D_{\tilde{\pi}_{\theta'}}(\gamma P_{\tilde{\pi}_{\theta'}} - I)X\right\|_\infty.
\end{aligned}
$$

Thanks to Assumption 4.2, for any θ,

$$\left(X^\top D_{\pi_\theta}(\gamma P_{\pi_\theta} - I)X\right)^{-1}$$

is well-defined. Since $\bar{\Lambda}_\pi$ is a compact set, we conclude, by the extreme value theorem, that there exists a constant $U_{inv} > 0$, independent of C_Γ, such that

$$\sup_\theta \left\| \left(X^\top D_{\pi_\theta}(\gamma P_{\pi_\theta} - I)X\right)^{-1} \right\|_\infty < U_{inv}.$$

Recalling that $\tilde{\pi}_\theta(a|s) = \pi_{\Gamma(\theta)}(a|s)$ then yields

$$\sup_\theta \left\| \left(X^\top D_{\tilde{\pi}_\theta}(\gamma P_{\tilde{\pi}_\theta} - I)X\right)^{-1} \right\|_\infty < U_{inv}.$$

It then follows immediately that

$$\left\| \left(X^\top D_{\tilde{\pi}_\theta}(\gamma P_{\tilde{\pi}_\theta} - I)X\right)^{-1} - \left(X^\top D_{\tilde{\pi}_{\theta'}}(\gamma P_{\tilde{\pi}_{\theta'}} - I)X\right)^{-1} \right\|_\infty \leq U_{inv}^2 \|X\|_\infty^2 L_{DP} L_\pi \|\theta - \theta'\|_\infty.$$

It is also easy to see that

$$\left\| X^\top D_{\tilde{\pi}_\theta} r \right\|_\infty \leq \|X\|_\infty \|r\|_\infty,$$
$$\left\| X^\top D_{\tilde{\pi}_\theta} r - X^\top D_{\tilde{\pi}_{\theta'}} r \right\|_\infty \leq \|X\|_\infty L_D L_\pi \|r\|_\infty.$$

Using Lemma C.1 again yields

$$\|w_\theta^* - w_{\theta'}^*\|_\infty \leq \left(U_{inv}^2 \|X\|_\infty^2 L_{DP} \|X\|_\infty + U_{inv} \|X\|_\infty L_D \right) L_\pi \|r\|_\infty \|\theta - \theta'\|_\infty.$$

It follows immediately that

$$\|w_\theta^* - w_{\theta'}^*\| \leq \underbrace{\sqrt{K} \left(U_{inv}^2 \|X\|_\infty^2 L_{DP} \|X\|_\infty + U_{inv} \|X\|_\infty L_D \right) L_\pi \|r\|_\infty}_{L_w} \|\theta - \theta'\|. \tag{23}$$

To verify Assumption 3.5 (vi), we have

$$\sup_\theta \|w_\theta^*\|_\infty \leq U_{inv} \|X\|_\infty \|r\|_\infty.$$

It follows immediately that

$$\sup_\theta \|w_\theta^*\| \leq \underbrace{\sqrt{K} U_{inv} \|X\|_\infty \|r\|_\infty}_{U_w}. \tag{24}$$

Assumption 3.5 (vii) follows immediately from Assumption 4.1.

We now verify Assumption 3.6. Assumption 3.6 (i) is fulfilled by our selection of C_Γ. It is easy to see

$$\tilde{\pi}_\theta(a|s) = \tilde{\pi}_{\Gamma(\theta)}(a|s),$$

Assumption 3.6 (ii) then follows immediately.

With Assumptions 3.1 - 3.6 satisfied, we conclude by Corollary 3.10 that the iterates $\{w_t\}$ generated by Algorithm 1 with $\lambda = 0$ satisfy

$$\mathbb{E}\left[\|w_t - w_{w_t}^*\|^2\right] = \frac{72 L_w^2 L_\theta^2}{\eta^2} + \begin{cases} \mathcal{O}\left(t^{-\frac{\eta c_\alpha}{3}} \log^2 t\right), & \epsilon_\alpha = 1, \eta c_\alpha \in (0,3) \\ \mathcal{O}\left(\frac{\log^3 t}{t}\right), & \epsilon_\alpha = 1, \eta c_\alpha = 3 \\ \mathcal{O}\left(\frac{\log^2 t}{t}\right), & \epsilon_\alpha = 1, \eta c_\alpha \in (3, \infty) \\ \mathcal{O}\left(\frac{\log^2 t}{t^{\epsilon_\alpha}}\right), & \epsilon_\alpha \in (0,1) \end{cases}.$$

where

$$
\begin{aligned}
L_\theta &\doteq U_F + (L_F + 1)C_\Gamma = \left(r_{max}x_{max} + \left((1+\gamma)x_{max}^2 + 2\right)C_\Gamma\right) \\
&\leq 1 + 4C_\Gamma.
\end{aligned}
$$

Consequently,

$$
\begin{aligned}
\mathbb{E}\left[\left\|w_t - w_{w_t}^*\right\|\right] &\leq \sqrt{\mathbb{E}\left[\left\|w_t - w_{w_t}^*\right\|^2\right]} \\
&= \frac{6\sqrt{2}L_w L_\theta}{\eta} + \begin{cases} \mathcal{O}\left(t^{-\frac{\eta c_\alpha}{6}}\log t\right), & \epsilon_\alpha = 1, \eta c_\alpha \in (0,3) \\ \mathcal{O}\left(t^{-\frac{1}{2}}\log^{\frac{3}{2}}t\right), & \epsilon_\alpha = 1, \eta c_\alpha = 3 \\ \mathcal{O}\left(t^{-\frac{1}{2}}\log t\right), & \epsilon_\alpha = 1, \eta c_\alpha \in (3,\infty) \\ \mathcal{O}\left(t^{-\frac{\epsilon_\alpha}{2}}\log t\right), & \epsilon_\alpha \in (0,1) \end{cases}.
\end{aligned}
$$

If

$$
\|r\|_\infty < \frac{1}{\sqrt{K}\left(U_{inv}^2\|X\|_\infty^2 L_{DP}\|X\|_\infty + U_{inv}\|X\|_\infty L_D\right)},
$$

we get

$$
L_w < 1.
$$

Since

$$
\begin{aligned}
&\mathbb{E}\left[\left\|w_t - w_*\right\|\right] \\
=&\mathbb{E}\left[\left\|w_t - w_{w_*}^*\right\|\right] \\
\leq&\mathbb{E}\left[\left\|w_t - w_{w_t}^*\right\|\right] + \mathbb{E}\left[\left\|w_{w_t}^* - w_{w_*}^*\right\|\right] \\
\leq&\mathbb{E}\left[\left\|w_t - w_{w_t}^*\right\|\right] + L_w\mathbb{E}\left[\left\|w_t - w_*\right\|\right],
\end{aligned}
$$

we conclude that

$$
\begin{aligned}
\mathbb{E}\left[\left\|w_t - w_*\right\|\right] &\leq \frac{1}{1-L_w}\mathbb{E}\left[\left\|w_t - w_{w_t}^*\right\|\right] \\
&= \frac{6\sqrt{2}L_w\left(1+4C_\Gamma\right)}{\eta(1-L_w)} + \begin{cases} \mathcal{O}\left(t^{-\frac{\eta c_\alpha}{6}}\log t\right), & \epsilon_\alpha = 1, \eta c_\alpha \in (0,3) \\ \mathcal{O}\left(t^{-\frac{1}{2}}\log^{\frac{3}{2}}t\right), & \epsilon_\alpha = 1, \eta c_\alpha = 3 \\ \mathcal{O}\left(t^{-\frac{1}{2}}\log t\right), & \epsilon_\alpha = 1, \eta c_\alpha \in (3,\infty) \\ \mathcal{O}\left(t^{-\frac{\epsilon_\alpha}{2}}\log t\right), & \epsilon_\alpha \in (0,1) \end{cases},
\end{aligned}
$$

which completes the proof.

$\square$

C. Technical Lemmas

Lemma C.1. *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 $L_1 U_2 + L_2 U_1$ is a Lipschitz constant of $f(x) \doteq f_1(x)f_2(x)$.*

Proof.

$$
\begin{aligned}
&\|f_1(x)f_2(x) - f_1(y)f_2(y)\| \\
\leq&\|f_1(x)\|\|f_2(x) - f_2(y)\| + \|f_2(y)\|\|f_1(x) - f_1(y)\| \\
\leq&(U_1 L_2 + U_2 L_1)\|x - y\|.
\end{aligned}
$$

$\square$

Lemma C.2. *Given positive integers $t_1 < t_2$ satisfying*

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

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

$$\|\Gamma(u_t) - \Gamma(u_{t_1})\| \leq 2\alpha_{t_1, t_2-1}(A\|\Gamma(u_{t_1})\| + B), \tag{25}$$

$$\|\Gamma(u_t) - \Gamma(u_{t_1})\| \leq 4\alpha_{t_1, t_2-1}(A\|\Gamma(u_{t_2})\| + B), \tag{26}$$

$$\|\Gamma(u_t) - \Gamma(u_{t_1})\| \leq \min\left\{\|\Gamma(u_{t_1})\|, \|\Gamma(u_{t_2})\|\right\} + \frac{B}{A}. \tag{27}$$

Proof. Notice that

$$
\begin{aligned}
&\|\Gamma(u_{t+1})\| - \|\Gamma(u_t)\| \\
\leq& \|\Gamma(u_{t+1}) - \Gamma(u_t)\| \\
=& \|\Gamma(u_{t+1}) - \Gamma(\Gamma(u_t))\| \\
\leq& \|u_{t+1} - \Gamma(u_t)\| \\
=& \alpha_t \|F_{\theta_t}(\Gamma(u_t), Y_t) - \Gamma(u_t)\| \\
\leq& \alpha_t \left(\|F_{\theta_t}(\Gamma(u_t), Y_t)\| + \|\Gamma(u_t)\|\right) \\
\leq& \alpha_t (U_F + (L_F + 1)\|\Gamma(u_t)\|) \quad \text{(Lemma D.10)} \\
\leq& \alpha_t(A\|\Gamma(u_t)\| + B) \quad \text{(Using (15))}
\end{aligned} \tag{28}
$$

The rest of the proof follows from 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

$$\|\Gamma(u_{t+1})\| + \frac{B}{A} \leq (1 + \alpha_t A)\left(\|\Gamma(u_t)\| + \frac{B}{A}\right),$$

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

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

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

$$\|\Gamma(u_t)\| + \frac{B}{A} \leq (1 + 2A\alpha_{t_1, t-1})\left(\|\Gamma(u_{t_1})\| + \frac{B}{A}\right)$$

$$\implies \|\Gamma(u_t)\| \leq (1 + 2A\alpha_{t_1, t-1})\|\Gamma(u_{t_1})\| + 2B\alpha_{t_1, t-1},$$

which together with (28) yields that for any $t \in (t_1, t_2 - 1]$

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

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

$$\|\Gamma(u_t) - \Gamma(u_{t_1})\| \leq \sum_{j=t_1}^{t-1} \|\Gamma(w_{j+1}) - \Gamma(w_j)\| \leq \sum_{j=t_1}^{t-1} 2\alpha_j(A\|\Gamma(u_{t_1})\| + B)$$

$$= 2\alpha_{t_1, t-1}(A\|\Gamma(u_{t_1})\| + B) \leq 2\alpha_{t_1, t_2-1}(A\|\Gamma(u_{t_1})\| + B),$$

which completes the proof of (25). For (26), we have from the above inequality

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

implying

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

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

$$\begin{aligned}
\|\Gamma(u_t) - \Gamma(u_{t_1})\| \leq & 2\alpha_{t_1, t_2-1}(A\|\Gamma(u_{t_1})\| + B) \\
\leq & 2\alpha_{t_1, t_2-1}(A\|\Gamma(u_{t_1}) - \Gamma(u_{t_2})\| + A\|\Gamma(u_{t_2})\| + B) \\
\leq & 2\alpha_{t_1, t_2-1}(A4\alpha_{t_1, t_2-1}(A\|\Gamma(u_{t_2})\| + B) + A\|\Gamma(u_{t_2})\| + B) \\
\leq & 4\alpha_{t_1, t_2-1}(A\|\Gamma(u_{t_2})\| + B) \quad (\text{Using } \alpha_{t_1, t_2-1} \leq \frac{1}{4A}) \quad,
\end{aligned}$$

which completes the proof of (26). (25) implies

$$\|\Gamma(u_t) - \Gamma(u_{t_1})\| \leq \|\Gamma(u_{t_1})\| + \frac{B}{A},$$

(26) implies

$$\|\Gamma(u_t) - \Gamma(u_{t_1})\| \leq \|\Gamma(u_{t_2})\| + \frac{B}{A},$$

then (27) follows immediately, which completes the proof. □

Lemma C.3. *Let Assumptions 4.1 and 4.2 hold. Then there exists a constant L'_π such that $\forall \theta, \theta', a, s$,*

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

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

Lemma C.4. *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 C.5. *Recall that*

$$L_\theta = U_F + (L_F + 1)C_\Gamma,$$

then for any $j > i, y, y', w,$

$$
\begin{aligned}
\left\| w_{\theta_j}^* - w_{\theta_i}^* \right\| &\leq L_w L_\theta \alpha_{i,j-1}, \\
\left| P_{\theta_j}(y, y') - P_{\theta_i}(y, y') \right| &\leq L_P L_\theta \alpha_{i,j-1}, \\
\left\| F_{\theta_j}(w, y) - F_{\theta_i}(w, y) \right\| &\leq L_F' L_\theta \alpha_{i,j-1} \left(\|w\| + U_F' \right), \\
\left\| \bar{F}_{\theta_j}(w) - \bar{F}_{\theta_i}(w) \right\| &\leq L_F'' L_\theta \alpha_{i,j-1} \left(\|w\| + U_F'' \right).
\end{aligned}
$$

Proof.

$$
\begin{aligned}
\left\| w_{\theta_j}^* - w_{\theta_i}^* \right\| &\leq \sum_{k=i}^{j-1} \left\| w_{\theta_{k+1}}^* - w_{\theta_k}^* \right\| \\
&\leq \sum_{k=i}^{j-1} \left\| w_{\theta_{k+1}}^* - w_{\Gamma(\theta_k)}^* \right\| \quad \text{(Assumption 3.6)} \\
&\leq \sum_{k=i}^{j-1} L_w \| \theta_{k+1} - \Gamma(\theta_k) \| \\
&= \sum_{k=i}^{j-1} L_w \| w_{k+1} - \Gamma(w_k) \| \\
&= \sum_{k=i}^{j-1} L_w \| \Gamma(u_{k+1}) - \Gamma(\Gamma(u_k)) \| \quad \text{(Lemma 3.8)} \\
&\leq \sum_{k=i}^{j-1} L_w \| u_{k+1} - \Gamma(u_k) \| \\
&= \sum_{k=i}^{j-1} L_w \alpha_k \| F_{\theta_k}(\Gamma(u_k), Y_k) - \Gamma(u_k) \| \\
&\leq \sum_{k=i}^{j-1} L_w \alpha_k \left(U_F + L_F \| \Gamma(u_k) \| + \| \Gamma(u_k) \| \right) \quad \text{(Lemma D.10)} \\
&\leq \sum_{k=i}^{j-1} L_w \alpha_k \left(U_F + L_F C_\Gamma + C_\Gamma \right) \\
&= L_w L_\theta \alpha_{i,j-1}.
\end{aligned}
$$

Similarly we can get

$$
\left| P_{\theta_j}(y, y') - P_{\theta_i}(y, y') \right| \leq L_P L_\theta \alpha_{i,j-1}.
$$

Moreover,

$$
\begin{aligned}
\left\| F_{\theta_j}(w, y) - F_{\theta_i}(w, y) \right\| &\leq \sum_{k=i}^{j-1} \left\| F_{\theta_{k+1}}(w, y) - F_{\theta_k}(w, y) \right\| \\
&\leq \sum_{k=i}^{j-1} \left\| F_{\theta_{k+1}}(w, y) - F_{\Gamma(\theta_k)}(w, y) \right\| \quad \text{(Assumption 3.6)} \\
&\leq \sum_{k=i}^{j-1} L_F' \| \theta_{k+1} - \Gamma(\theta_k) \| \left(\|w\| + U_F' \right) \\
&\leq L_F' L_\theta \alpha_{i,j-1} \left(\|w\| + U_F' \right).
\end{aligned}
$$

Since $P_\theta = P_{\Gamma(\theta)}$, it is easy to see $d_\theta(y) = d_{\Gamma(\theta)}(y)$. Consequently, $\bar{F}_\theta(w) = \bar{F}_{\Gamma(\theta)}(w)$. We can then similarly get

$$\left\| \bar{F}_{\theta_j}(w) - \bar{F}_{\theta_i}(w) \right\| \le L_F'' L_\theta \alpha_{i,j-1} \left(\|w\| + U_F'' \right),$$

which completes the proof. $\qquad\square$

Lemma C.6. *(Lemma 5.4 of De Farias & Van Roy (2000)) There exists an $\bar{\alpha}$ such that for all $\alpha \in (0, \bar{\alpha})$ and all θ,*

$$\|f_\theta^\alpha(w) - w_\theta^*\| \le \kappa_\alpha \|w - w_\theta^*\|,$$

where

$$\kappa_\alpha \doteq \sqrt{1 - (1 - \gamma) \inf_\theta \lambda_{min} \left(X^\top D_{\pi_\theta} X \right) \alpha} < 1.$$

Here $\lambda_{min}(\cdot)$ denotes the minimum eigenvalue of a symmetric positive definite matrix.

Proof. The proof is due to De Farias & Van Roy (2000); we rewrite it in our notation for completeness. We first recall

$$f_\theta^\alpha(w) = w + \alpha \left(X^\top D_{\tilde{\pi}_\theta}(\gamma P_{\tilde{\pi}_\theta} - I)Xw + X^\top D_{\tilde{\pi}_\theta} r \right),$$
$$w_\theta^* = - \left(X^\top D_{\tilde{\pi}_\theta}(\gamma P_{\tilde{\pi}_\theta} - I)X \right)^{-1} X^\top D_{\tilde{\pi}_\theta} r.$$

Define

$$\begin{aligned}
g_\theta(w) &\doteq X^\top D_{\tilde{\pi}_\theta}(\gamma P_{\tilde{\pi}_\theta} - I)Xw + X^\top D_{\tilde{\pi}_\theta} r \\
&= X^\top D_{\tilde{\pi}_\theta} X \left(X^\top D_{\tilde{\pi}_\theta} X \right)^{-1} X^\top D_{\tilde{\pi}_\theta} \left(\mathcal{T}_{\tilde{\pi}_\theta} Xw - Xw \right) \\
&= X^\top D_{\tilde{\pi}_\theta} \Pi_{D_{\tilde{\pi}_\theta}} \mathcal{T}_{\tilde{\pi}_\theta} Xw - X^\top D_{\tilde{\pi}_\theta} Xw \\
&= X^\top D_{\tilde{\pi}_\theta} \left(\Pi_{D_{\tilde{\pi}_\theta}} \mathcal{T}_{\tilde{\pi}_\theta} Xw - Xw \right).
\end{aligned}$$

By the contraction property (see, e.g., Tsitsiklis & Roy (1996)),

$$\left\| \Pi_{D_{\tilde{\pi}_\theta}} \mathcal{T}_{\tilde{\pi}_\theta} Xw - Xw_\theta^* \right\|_{D_{\tilde{\pi}_\theta}} \le \gamma \|Xw - Xw_\theta^*\|_{D_{\tilde{\pi}_\theta}}.$$

Consequently,

$$\begin{aligned}
&(w - w_\theta^*)^\top g_\theta(s) \\
&= (Xw - Xw_\theta^*)^\top D_{\tilde{\pi}_\theta} \left(\Pi_{D_{\tilde{\pi}_\theta}} \mathcal{T}_{\tilde{\pi}_\theta} Xw - Xw \right) \\
&= (Xw - Xw_\theta^*)^\top D_{\tilde{\pi}_\theta} \left(\Pi_{D_{\tilde{\pi}_\theta}} \mathcal{T}_{\tilde{\pi}_\theta} Xw - Xw_\theta^* + Xw_\theta^* - Xw \right) \\
&\le \|Xw - Xw_\theta^*\|_{D_{\tilde{\pi}_\theta}} \left\| \Pi_{D_{\tilde{\pi}_\theta}} \mathcal{T}_{\tilde{\pi}_\theta} Xw - Xw_\theta^* \right\|_{D_{\tilde{\pi}_\theta}} - \|Xw - Xw_\theta^*\|_{D_{\tilde{\pi}_\theta}}^2
\end{aligned}$$

$$\text{(Cauthy-Schwarz inequality)}$$

$$\begin{aligned}
&\le (\gamma - 1)\|Xw - Xw_\theta^*\|_{D_{\tilde{\pi}_\theta}}^2 \\
&= (\gamma - 1)(w - w_\theta^*)^\top \left(X^\top D_{\tilde{\pi}_\theta} X \right) (w - w_\theta^*).
\end{aligned}$$

Since $X^\top D_{\pi_\theta} X$ is symmetric and positive define, eigenvalues are continuous in the elements of the matrix, $\bar{\Lambda}_\pi$ is compact, we conclude, by the extreme value theorem, that

$$C_1 \doteq \inf_\theta \lambda_{min} \left(X^\top D_{\pi_\theta} X \right) > 0.$$

Consequently, for any y and θ,

$$y^\top X^\top D_{\pi_\theta} X y \ge C_1 \|y\|^2,$$

implying

$$y^\top X^\top D_{\tilde{\pi}_\theta} X y \geq C_1 \|y\|^2.$$

It follows immediately that

$$(w - w_\theta^*)^\top g_\theta(w) \leq -(1 - \gamma) C_1 \|w - w_\theta^*\|^2. \tag{29}$$

Moreover, let x_i be the i-the column of X, we have

$$\|g_\theta(w)\|^2 = \sum_{i=1}^K \left(x_i^\top D_{\tilde{\pi}_\theta} \left(\Pi_{D_{\tilde{\pi}_\theta}} \mathcal{T}_{\tilde{\pi}_\theta} X w - X w \right) \right)^2$$

$$\leq \sum_{i=1}^K \|x_i\|_{D_{\tilde{\pi}_\theta}}^2 \left\| \Pi_{D_{\tilde{\pi}_\theta}} \mathcal{T}_{\tilde{\pi}_\theta} X w - X w \right\|_{D_{\tilde{\pi}_\theta}}^2$$

(Cauchy-Schwarz inequality)

$$\leq \sum_{i=1}^K \|x_i\|_{D_{\tilde{\pi}_\theta}}^2 \left(\left\| \Pi_{D_{\tilde{\pi}_\theta}} \mathcal{T}_{\tilde{\pi}_\theta} X w - X w_\theta^* \right\|_{D_{\tilde{\pi}_\theta}} + \|X w_\theta^* - X w\|_{D_{\tilde{\pi}_\theta}} \right)^2$$

$$\leq (1 + \gamma)^2 \sum_{i=1}^K \|x_i\|_{D_{\tilde{\pi}_\theta}}^2 \|X w_\theta^* - X w\|_{D_{\tilde{\pi}_\theta}}^2$$

$$= (1 + \gamma)^2 \left(\sum_{i=1}^K \|x_i\|_{D_{\tilde{\pi}_\theta}}^2 \right) \left\| X^\top D_{\tilde{\pi}_\theta} X \right\| \|w - w_\theta^*\|^2.$$

According to the extreme value theorem,

$$C_2 \doteq \sup_\theta \left(\sum_{i=1}^K \|x_i\|_{D_{\pi_\theta}}^2 \right) \left\| X^\top D_{\pi_\theta} X \right\| < \infty.$$

Consequently, we have

$$\|g_\theta(w)\|^2 \leq (1 + \gamma)^2 C_2 \|w - w_\theta^*\|^2. \tag{30}$$

Combining (29) and (30) yields

$$\begin{aligned}
\|f_\theta^\alpha(w) - w_\theta^*\|^2 &= \|w + \alpha g_\theta(w) - w_\theta^*\|^2 \\
&= \|w - w_\theta^*\|^2 + 2\alpha(w - w_\theta^*)^\top g_\theta(w) + \alpha^2 \|g_\theta(w)\|^2 \\
&\leq \left(1 - 2\alpha(1 - \gamma)C_1 + (1 + \gamma)^2 \alpha^2 C_2 \right) \|w - w_\theta^*\|^2.
\end{aligned}$$

Consequently, if

$$\alpha < \bar{\alpha} \doteq \frac{(1 - \gamma)C_1}{(1 + \gamma)^2 C_2},$$

we have

$$1 - 2\alpha(1 - \gamma)C_1 + (1 + \gamma)^2 \alpha^2 C_2 \leq 1 - (1 - \gamma)C_1 \alpha.$$

Defining

$$\kappa_\alpha \doteq \sqrt{1 - (1 - \gamma)C_1 \alpha}$$

then completes the proof. Importantly, both C_1 and C_2 here are independent of C_Γ. □

27

D. Proof of Auxiliary Lemmas

D.1. Proof of Lemma A.1

Lemma D.1. *(Bound of T_1)*

$$T_1 \le L_w L_\theta \alpha_t \big\| \Gamma(u_t) - w_{\theta_t}^* \big\|.$$

Proof.

$$\begin{aligned}
T_1 =& \Big\langle \Gamma(u_t) - w_{\theta_t}^*, w_{\theta_t}^* - w_{\theta_{t+1}}^* \Big\rangle \\
\le& \big\| \Gamma(u_t) - w_{\theta_t}^* \big\| \big\| w_{\theta_t}^* - w_{\theta_{t+1}}^* \big\| \\
\le& \big\| \Gamma(u_t) - w_{\theta_t}^* \big\| L_w L_\theta \alpha_t \quad \text{(Lemma C.5)}
\end{aligned}$$

$\square$

D.2. Proof of Lemma A.2

Lemma D.2. *(Bound of T_2)*

$$T_2 \le -(1 - \kappa_{\alpha_t}) \big\| \Gamma(u_t) - w_{\theta_t}^* \big\|^2.$$

Proof.

$$\begin{aligned}
T_2 =& \big\langle \Gamma(u_t) - w_{\theta_t}^*, f_{\theta_t}^{\alpha_t}(\Gamma(u_t)) - \Gamma(u_t) \big\rangle \\
=& \big\langle \Gamma(u_t) - w_{\theta_t}^*, f_{\theta_t}^{\alpha_t}(\Gamma(u_t)) - w_{\theta_t}^* \big\rangle - \big\langle \Gamma(u_t) - w_{\theta_t}^*, \Gamma(u_t) - w_{\theta_t}^* \big\rangle \\
\le& \big\| \Gamma(u_t) - w_{\theta_t}^* \big\| \big\| f_{\theta_t}^{\alpha_t}(\Gamma(u_t)) - w_{\theta_t}^* \big\| - \big\| \Gamma(u_t) - w_{\theta_t}^* \big\|^2 \\
\le& \big\| \Gamma(u_t) - w_{\theta_t}^* \big\| \kappa_{\alpha_t} \big\| \Gamma(u_t) - w_{\theta_t}^* \big\| - \big\| \Gamma(u_t) - w_{\theta_t}^* \big\|^2 \quad \text{(Assumption 3.4)} \\
=& - (1 - \kappa_{\alpha_t}) \big\| \Gamma(u_t) - w_{\theta_t}^* \big\|^2.
\end{aligned}$$

$\square$

D.3. Proof of Lemma A.3

Lemma D.3. *(Bound of T_{31})*

$$T_{31} \le 8(L_w L_\theta + 1)\alpha_{t-\tau_{\alpha_t},t-1} \left(A^2 \big\| \Gamma(u_t) - w_{\theta_t}^* \big\|^2 + C^2 \right).$$

Proof.

$$\begin{aligned}
T_{31} =& \Big\langle \Gamma(u_t) - w_{\theta_t}^* - \Big(\Gamma(u_{t-\tau_{\alpha_t}}) - w_{\theta_{t-\tau_{\alpha_t}}}^* \Big), F_{\theta_t}(\Gamma(u_t), Y_t) - \bar{F}_{\theta_t}(\Gamma(u_t)) \Big\rangle \\
\le& \Big\| \Gamma(u_t) - w_{\theta_t}^* - \Big(\Gamma(u_{t-\tau_{\alpha_t}}) - w_{\theta_{t-\tau_{\alpha_t}}}^* \Big) \Big\| \big\| F_{\theta_t}(\Gamma(u_t), Y_t) - \bar{F}_{\theta_t}(\Gamma(u_t)) \big\|.
\end{aligned}$$

For the first term, we have

$$\begin{aligned}
& \Big\| \Gamma(u_t) - w_{\theta_t}^* - \Big(\Gamma(u_{t-\tau_{\alpha_t}}) - w_{\theta_{t-\tau_{\alpha_t}}}^* \Big) \Big\| \\
\le& \big\| \Gamma(u_t) - \Gamma(u_{t-\tau_{\alpha_t}}) \big\| + \Big\| w_{\theta_t}^* - w_{\theta_{t-\tau_{\alpha_t}}}^* \Big\| \\
\le& \big\| \Gamma(u_t) - \Gamma(u_{t-\tau_{\alpha_t}}) \big\| + L_w L_\theta \alpha_{t-\tau_{\alpha_t},t-1} \quad \text{(Lemma C.5)} \\
\le& 4\alpha_{t-\tau_{\alpha_t},t-1}(A\|\Gamma(u_t)\| + B) + L_w L_\theta \alpha_{t-\tau_{\alpha_t},t-1} \quad \text{(Lemma C.2)} \\
\le& 4\alpha_{t-\tau_{\alpha_t},t-1}(A\big\| \Gamma(u_t) - w_{\theta_t}^* \big\| + A\big\| w_{\theta_t}^* \big\| + B) + L_w L_\theta \alpha_{t-\tau_{\alpha_t},t-1} \\
\le& 4\alpha_{t-\tau_{\alpha_t},t-1}(L_w L_\theta + 1)(A\big\| \Gamma(u_t) - w_{\theta_t}^* \big\| + A\big\| w_{\theta_t}^* \big\| + B + 1).
\end{aligned}$$

28

For the second term, we have

$$\left\|F_{\theta_t}(\Gamma(u_t), Y_t) - \bar{F}_{\theta_t}(\Gamma(u_t))\right\|$$
$$\leq\left\|F_{\theta_t}(\Gamma(u_t), Y_t)\right\| + \left\|\bar{F}_{\theta_t}(\Gamma(u_t)) - \bar{F}_{\theta_t}(w_{\theta_t}^*)\right\| + \left\|w_{\theta_t}^*\right\| \quad \text{(Assumption 3.4(i))}$$
$$\leq U_F + L_F\|\Gamma(u_t)\| + \left\|\bar{F}_{\theta_t}(\Gamma(u_t)) - \bar{F}_{\theta_t}(w_{\theta_t}^*)\right\| + \left\|w_{\theta_t}^*\right\| \quad \text{(Lemma D.10)}$$
$$= U_F + L_F\|\Gamma(u_t)\| + \left\|\sum_y d_{\theta_t}(y)\left(F_{\theta_t}(\Gamma(u_t), y) - F_{\theta_t}(w_{\theta_t}^*, y)\right)\right\| + \left\|w_{\theta_t}^*\right\|$$
$$\leq U_F + L_F\|\Gamma(u_t)\| + L_F\left\|\Gamma(u_t) - w_{\theta_t}^*\right\| + \left\|w_{\theta_t}^*\right\|$$
$$\leq U_F + L_F\left\|\Gamma(u_t) - w_{\theta_t}^*\right\| + L_F\left\|w_{\theta_t}^*\right\| + L_F\left\|\Gamma(u_t) - w_{\theta_t}^*\right\| + \left\|w_{\theta_t}^*\right\|$$
$$\leq A\left\|\Gamma(u_t) - w_{\theta_t}^*\right\| + A\left\|w_{\theta_t}^*\right\| + B.$$

Combining the two inequalities together yields

$$\left\langle\Gamma(u_t) - w_{\theta_t}^* - \left(\Gamma(u_{t-\tau_{\alpha_t}}) - w_{\theta_t}^*\right), F_{\theta_t}(\Gamma(u_t), Y_t) - \bar{F}_{\theta_t}(\Gamma(u_t))\right\rangle$$
$$\leq 4(L_w L_\theta + 1)\alpha_{t-\tau_{\alpha_t}, t-1}(A\left\|\Gamma(u_t) - w_{\theta_t}^*\right\| + C)^2$$
$$\leq 8(L_w L_\theta + 1)\alpha_{t-\tau_{\alpha_t}, t-1}(A^2\left\|\Gamma(u_t) - w_{\theta_t}^*\right\|^2 + C^2),$$

which completes the proof. $\square$

D.4. Proof of Lemma A.4

Lemma D.4. *(Bound of T_{32})*

$$T_{32} \leq 16\alpha_{t-\tau_{\alpha_t}, t-1}(1 + L_w L_\theta \alpha_{t-\tau_{\alpha_t}, t-1})\left(A^2\left\|\Gamma(u_t) - w_{\theta_t}^*\right\|^2 + C^2\right).$$

Proof.

$$T_{32} = \left\langle\Gamma(u_{t-\tau_{\alpha_t}}) - w_{\theta_{t-\tau_{\alpha_t}}}^*, F_{\theta_t}(\Gamma(u_t), Y_t) - F_{\theta_t}(\Gamma(u_{t-\tau_{\alpha_t}}), Y_t) + \bar{F}_{\theta_t}(\Gamma(u_{t-\tau_{\alpha_t}})) - \bar{F}_{\theta_t}(\Gamma(u_t))\right\rangle$$
$$\leq\left\|\Gamma(u_{t-\tau_{\alpha_t}}) - w_{\theta_{t-\tau_{\alpha_t}}}^*\right\|\left\|F_{\theta_t}(\Gamma(u_t), Y_t) - F_{\theta_t}(\Gamma(u_{t-\tau_{\alpha_t}}), Y_t) + \bar{F}_{\theta_t}(\Gamma(u_{t-\tau_{\alpha_t}})) - \bar{F}_{\theta_t}(\Gamma(u_t))\right\|.$$

For the first term, we have

$$\left\|\Gamma(u_{t-\tau_{\alpha_t}}) - w_{\theta_{t-\tau_{\alpha_t}}}^*\right\| \tag{31}$$
$$= \left\|\Gamma(u_{t-\tau_{\alpha_t}}) - w_{\theta_{t-\tau_{\alpha_t}}}^* - (w_{\theta_t}^* - w_{\theta_t}^*)\right\|$$
$$\leq\left\|\Gamma(u_{t-\tau_{\alpha_t}}) - w_{\theta_t}^*\right\| + \left\|w_{\theta_t}^* - w_{\theta_{t-\tau_{\alpha_t}}}^*\right\|$$
$$\leq\left\|\Gamma(u_{t-\tau_{\alpha_t}}) - w_{\theta_t}^*\right\| + L_w L_\theta \alpha_{t-\tau_{\alpha_t}, t-1} \quad \text{(Lemma C.5)}$$
$$\leq\left\|\Gamma(u_{t-\tau_{\alpha_t}}) - \Gamma(u_t)\right\| + \left\|\Gamma(u_t) - w_{\theta_t}^*\right\| + L_w L_\theta \alpha_{t-\tau_{\alpha_t}, t-1}$$
$$\leq\|\Gamma(u_t)\| + \frac{B}{A} + \left\|\Gamma(u_t) - w_{\theta_t}^*\right\| + L_w L_\theta \alpha_{t-\tau_{\alpha_t}, t-1} \quad \text{(Lemma C.2)}$$
$$\leq(1 + L_w L_\theta \alpha_{t-\tau_{\alpha_t}, t-1})\left(\left\|w_{\theta_t}^*\right\| + \left\|\Gamma(u_t) - w_{\theta_t}^*\right\| + \frac{B}{A} + \left\|\Gamma(u_t) - w_{\theta_t}^*\right\| + 1\right)$$
$$\leq 2(1 + L_w L_\theta \alpha_{t-\tau_{\alpha_t}, t-1})\left(U_w + \frac{B}{A} + \left\|\Gamma(u_t) - w_{\theta_t}^*\right\| + 1\right)$$
$$\leq\frac{2(1 + L_w L_\theta \alpha_{t-\tau_{\alpha_t}, t-1})}{A}\left(A\left\|\Gamma(u_t) - w_{\theta_t}^*\right\| + C\right).$$

29

For the second term,

$$\left\| F_{\theta_t}(\Gamma(u_t), Y_t) - F_{\theta_t}(\Gamma(u_{t-\tau_{\alpha_t}}), Y_t) + \bar{F}_{\theta_t}(\Gamma(u_{t-\tau_{\alpha_t}})) - \bar{F}_{\theta_t}(\Gamma(u_t)) \right\|$$

$$\leq \left\| F_{\theta_t}(\Gamma(u_t), Y_t) - F_{\theta_t}(\Gamma(u_{t-\tau_{\alpha_t}}), Y_t) \right\| + \left\| \bar{F}_{\theta_t}(\Gamma(u_{t-\tau_{\alpha_t}})) - \bar{F}_{\theta_t}(\Gamma(u_t)) \right\|$$

$$\leq L_F \left\| \Gamma(u_{t-\tau_{\alpha_t}}) - \Gamma(u_t) \right\| + \left\| \sum_y d_{\theta_t}(y) \left(F_{\theta_t}(\Gamma(u_{t-\tau_{\alpha_t}}), y) - F_{\theta_t}(\Gamma(u_t), y) \right) \right\|$$

$$\leq 2L_F \left\| \Gamma(u_{t-\tau_{\alpha_t}}) - \Gamma(u_t) \right\|$$

$$\leq A \left\| \Gamma(u_{t-\tau_{\alpha_t}}) - \Gamma(u_t) \right\|$$

$$\leq 4A\alpha_{t-\tau_{\alpha_t},t-1} \left(A \|\Gamma(u_t)\| + B \right) \quad \text{(Lemma C.2)}$$

$$\leq 4A\alpha_{t-\tau_{\alpha_t},t-1} (A \|\Gamma(u_t) - w_{\theta_t}^*\| + A \|w_{\theta_t}^*\| + B).$$

Combining the two inequalities together yields

$$T_{32} \leq 8\alpha_{t-\tau_{\alpha_t},t-1}(1 + L_w L_\theta \alpha_{t-\tau_{\alpha_t},t-1})(A \|u_t - w_{\theta_t}^*\| + C)^2$$

$$\leq 16\alpha_{t-\tau_{\alpha_t},t-1}(1 + L_w L_\theta \alpha_{t-\tau_{\alpha_t},t-1}) \left(A^2 \|u_t - w_{\theta_t}^*\|^2 + C^2 \right),$$

which completes the proof. □

D.5. Proof of Lemma A.5

Lemma D.5. *(Bound of T_{331})*

$$\mathbb{E}[T_{331}] \leq \frac{8\alpha_t(1 + L_w L_\theta \alpha_{t-\tau_{\alpha_t},t-1})}{A} \left(A^2 \mathbb{E}\left[\|\Gamma(u_t) - w_{\theta_t}^*\|^2 \right] + C^2 \right).$$

Proof.

$$\mathbb{E}[T_{331}] \tag{32}$$

$$= \mathbb{E}\left[\left\langle \Gamma(u_{t-\tau_{\alpha_t}}) - w_{\theta_{t-\tau_{\alpha_t}}}^*, F_{\theta_{t-\tau_{\alpha_t}}}(\Gamma(u_{t-\tau_{\alpha_t}}), \tilde{Y}_t) - \bar{F}_{\theta_{t-\tau_{\alpha_t}}}(\Gamma(u_{t-\tau_{\alpha_t}})) \right\rangle \right]$$

$$= \mathbb{E}\left[\mathbb{E}\left[\left\langle \Gamma(u_{t-\tau_{\alpha_t}}) - w_{\theta_{t-\tau_{\alpha_t}}}^*, F_{\theta_{t-\tau_{\alpha_t}}}(\Gamma(u_{t-\tau_{\alpha_t}}), \tilde{Y}_t) - \bar{F}_{\theta_{t-\tau_{\alpha_t}}}(\Gamma(u_{t-\tau_{\alpha_t}})) \right\rangle \Big|_{\substack{\theta_{t-\tau_{\alpha_t}} \\ u_{t-\tau_{\alpha_t}} \\ Y_{t-\tau_{\alpha_t}}}} \right] \right]$$

$$= \mathbb{E}\left[\left\langle \Gamma(u_{t-\tau_{\alpha_t}}) - w_{\theta_{t-\tau_{\alpha_t}}}^*, \mathbb{E}\left[F_{\theta_{t-\tau_{\alpha_t}}}(\Gamma(u_{t-\tau_{\alpha_t}}), \tilde{Y}_t) - \bar{F}_{\theta_{t-\tau_{\alpha_t}}}(\Gamma(u_{t-\tau_{\alpha_t}})) \Big|_{\substack{\theta_{t-\tau_{\alpha_t}} \\ u_{t-\tau_{\alpha_t}} \\ Y_{t-\tau_{\alpha_t}}}} \right] \right\rangle \right]$$

$$\leq \mathbb{E}\left[\left\| \Gamma(u_{t-\tau_{\alpha_t}}) - w_{\theta_{t-\tau_{\alpha_t}}}^* \right\| \left\| \mathbb{E}\left[F_{\theta_{t-\tau_{\alpha_t}}}(\Gamma(u_{t-\tau_{\alpha_t}}), \tilde{Y}_t) - \bar{F}_{\theta_{t-\tau_{\alpha_t}}}(\Gamma(u_{t-\tau_{\alpha_t}})) \Big|_{\substack{\theta_{t-\tau_{\alpha_t}} \\ u_{t-\tau_{\alpha_t}} \\ Y_{t-\tau_{\alpha_t}}}} \right] \right\| \right].$$

We now bound the inner expectation.

$$\left\| \mathbb{E}\left[F_{\theta_{t-\tau_{\alpha_t}}}(\Gamma(u_{t-\tau_{\alpha_t}}), \tilde{Y}_t) - \bar{F}_{\theta_{t-\tau_{\alpha_t}}}(\Gamma(u_{t-\tau_{\alpha_t}})) \mid {}^{\theta_{t-\tau_{\alpha_t}}}_{{}^{u_{t-\tau_{\alpha_t}}}_{Y_{t-\tau_{\alpha_t}}}} \right] \right\|$$

$$= \left\| \sum_y \left(\Pr\left(\tilde{Y}_t = y \mid {}^{\theta_{t-\tau_{\alpha_t}}}_{{}^{u_{t-\tau_{\alpha_t}}}_{Y_{t-\tau_{\alpha_t}}}} \right) - d_{\theta_{t-\tau_{\alpha_t}}}(y) \right) F_{\theta_{t-\tau_{\alpha_t}}}(\Gamma(u_{t-\tau_{\alpha_t}}), y) \right\|$$

$$\leq \max_y \left\| F_{\theta_{t-\tau_{\alpha_t}}}(\Gamma(u_{t-\tau_{\alpha_t}}), y) \right\| \sum_y \left| \Pr\left(\tilde{Y}_t = y \mid {}^{\theta_{t-\tau_{\alpha_t}}}_{{}^{u_{t-\tau_{\alpha_t}}}_{Y_{t-\tau_{\alpha_t}}}} \right) - d_{\theta_{t-\tau_{\alpha_t}}}(y) \right|$$

$$\leq \max_y \left\| F_{\theta_{t-\tau_{\alpha_t}}}(\Gamma(u_{t-\tau_{\alpha_t}}), y) \right\| \alpha_t \quad \text{(Definition of } \tau_{\alpha_t}) \tag{33}$$

$$\leq \alpha_t \left(U_F + L_F \left\| \Gamma(u_{t-\tau_{\alpha_t}}) \right\| \right) \quad \text{(Lemma D.10)}$$

$$\leq \alpha_t \left(U_F + L_F \left\| \Gamma(u_{t-\tau_{\alpha_t}}) - \Gamma(u_t) \right\| + L_F \| \Gamma(u_t) \| \right)$$

$$\leq \alpha_t \left(B + A \left(\| \Gamma(u_t) \| + \frac{B}{A} \right) + A \| \Gamma(u_t) \| \right) \quad \text{(Lemma C.2)}$$

$$\leq \alpha_t \left(2B + (A+1) \| \Gamma(u_t) \| \right)$$

$$\leq 2\alpha_t \left(B + A \| \Gamma(u_t) \| \right)$$

$$\leq 2\alpha_t \left(B + A \| \Gamma(u_t) - w_{\theta_t}^* \| + A \| w_{\theta_t}^* \| \right)$$

$$\leq 2\alpha_t \left(A \| \Gamma(u_t) - w_{\theta_t}^* \| + C \right)$$

Using the above inequality and (31) yields

$$\mathbb{E}[T_{331}]$$

$$\leq \mathbb{E}\left[\frac{4\alpha_t(1 + L_w L_\theta \alpha_{t-\tau_{\alpha_t},t-1})}{A} \left(A \| \Gamma(u_t) - w_{\theta_t}^* \| + C \right)^2 \right]$$

$$\leq \mathbb{E}\left[\frac{8\alpha_t(1 + L_w L_\theta \alpha_{t-\tau_{\alpha_t},t-1})}{A} \left(A^2 \| \Gamma(u_t) - w_{\theta_t}^* \|^2 + C^2 \right) \right],$$

which completes the proof. $\qquad \square$

D.6. Proof of Lemma A.6

Lemma D.6. *(Bound of T_{332})*

$$\mathbb{E}[T_{332}] \leq \frac{8|\mathcal{Y}| L_P L_\theta \sum_{j=t-\tau_{\alpha_t}}^{t-1} \alpha_{t-\tau_{\alpha_t},j}(1 + L_w L_\theta \alpha_{t-\tau_{\alpha_t},t-1})}{A} \left(A^2 \mathbb{E}\left[\| \Gamma(u_t) - w_{\theta_t}^* \|^2 \right] + C^2 \right).$$

Proof.

$$\mathbb{E}[T_{332}]$$

$$= \mathbb{E}\left[\left\langle \Gamma(u_{t-\tau_{\alpha_t}}) - w_{\theta_{t-\tau_{\alpha_t}}}^*, F_{\theta_{t-\tau_{\alpha_t}}}(\Gamma(u_{t-\tau_{\alpha_t}}), Y_t) - F_{\theta_{t-\tau_{\alpha_t}}}(\Gamma(u_{t-\tau_{\alpha_t}}), \tilde{Y}_t) \right\rangle \right]$$

$$\leq \mathbb{E}\left[\left\| \Gamma(u_{t-\tau_{\alpha_t}}) - w_{\theta_{t-\tau_{\alpha_t}}}^* \right\| \left\| \mathbb{E}\left[F_{\theta_{t-\tau_{\alpha_t}}}(\Gamma(u_{t-\tau_{\alpha_t}}), Y_t) - F_{\theta_{t-\tau_{\alpha_t}}}(\Gamma(u_{t-\tau_{\alpha_t}}), \tilde{Y}_t) \mid {}^{u_{t-\tau_{\alpha_t}}}_{{}^{\theta_{t-\tau_{\alpha_t}}}_{Y_{t-\tau_{\alpha_t}}}} \right] \right\| \right]$$

$$\text{(Similar to (32))}$$

$$\leq \mathbb{E}\left[\frac{2(1 + L_w L_\theta \alpha_{t-\tau_{\alpha_t},t-1})}{A} \left(A \| \Gamma(u_t) - w_{\theta_t}^* \| + C \right) \right.$$

$$\left. \times 2|\mathcal{Y}| L_P L_\theta \sum_{j=t-\tau_{\alpha_t}}^{t-1} \alpha_{t-\tau_{\alpha_t},j} \left(A \| \Gamma(u_t) - w_{\theta_t}^* \| + C \right) \right]$$

<div align="right">(Using (31) and Lemma D.11)</div>

$$\leq \frac{8|\mathcal{Y}|L_P L_\theta \sum_{j=t-\tau_{\alpha_t}}^{t-1} \alpha_{t-\tau_{\alpha_t},j}(1 + L_w L_\theta \alpha_{t-\tau_{\alpha_t},t-1})}{A} \left(A^2 \mathbb{E}\left[\|\Gamma(u_t) - w^*_{\theta_t}\|^2\right] + C^2\right),$$

which completes the proof. □

D.7. Proof of Lemma A.7

Lemma D.7. *(Bound of T_{333})*

$$T_{333} \leq \frac{4L'_F L_\theta \alpha_{t-\tau_{\alpha_t},t-1}(1 + L_w L_\theta \alpha_{t-\tau_{\alpha_t},t-1})}{A^2}\left(A^2\|\Gamma(u_t) - w^*_{\theta_t}\|^2 + C^2\right).$$

Proof.

$$\begin{aligned}
T_{333} &= \left\langle \Gamma(u_{t-\tau_{\alpha_t}}) - w^*_{\theta_{t-\tau_{\alpha_t}}}, F_{\theta_t}(\Gamma(u_{t-\tau_{\alpha_t}}), Y_t) - F_{\theta_{t-\tau_{\alpha_t}}}(\Gamma(u_{t-\tau_{\alpha_t}}), Y_t)\right\rangle \\
&\leq \left\|\Gamma(u_{t-\tau_{\alpha_t}}) - w^*_{\theta_{t-\tau_{\alpha_t}}}\right\|\left\|F_{\theta_t}(\Gamma(u_{t-\tau_{\alpha_t}}), Y_t) - F_{\theta_{t-\tau_{\alpha_t}}}(\Gamma(u_{t-\tau_{\alpha_t}}), Y_t)\right\| \\
&\leq \frac{2(1 + L_w L_\theta \alpha_{t-\tau_{\alpha_t},t-1})}{A}\left(A\|\Gamma(u_t) - w^*_{\theta_t}\| + C\right) \\
&\quad \times L'_F L_\theta \alpha_{t-\tau_{\alpha_t},t-1}\left(\|\Gamma(u_{t-\tau_{\alpha_t}})\| + U'_F\right) \quad \text{(Using (31) and Lemma C.5)} \quad .
\end{aligned}$$

Since

$$\begin{aligned}
&\left\|\Gamma(u_{t-\tau_{\alpha_t}})\right\| \\
&\leq \left\|\Gamma(u_{t-\tau_{\alpha_t}}) - \Gamma(u_t)\right\| + \|\Gamma(u_t)\| \\
&\leq 2\|\Gamma(u_t)\| + \frac{B}{A} \quad \text{(Lemma C.2)} \\
&\leq 2\|\Gamma(u_t) - w^*_{\theta_t}\| + 2\|w^*_{\theta_t}\| + \frac{B}{A},
\end{aligned} \tag{34}$$

we have

$$T_{333} \leq \frac{8L'_F L_\theta \alpha_{t-\tau_{\alpha_t},t-1}(1 + L_w L_\theta \alpha_{t-\tau_{\alpha_t},t-1})}{A^2}\left(A^2\|\Gamma(u_t) - w^*_{\theta_t}\|^2 + C^2\right),$$

which completes the proof. □

D.8. Proof of Lemma A.8

Lemma D.8. *(Bound of T_{334})*

$$T_{334} \leq \frac{4L''_F L_\theta \alpha_{t-\tau_{\alpha_t},t-1}(1 + L_w L_\theta \alpha_{t-\tau_{\alpha_t},t-1})}{A^2}\left(A^2\|\Gamma(u_t) - w^*_{\theta_t}\|^2 + C^2\right).$$

Proof.

$$\begin{aligned}
T_{334} &= \left\langle \Gamma(u_{t-\tau_{\alpha_t}}) - w^*_{\theta_{t-\tau_{\alpha_t}}}, \bar{F}_{\theta_{t-\tau_{\alpha_t}}}(\Gamma(u_{t-\tau_{\alpha_t}})) - \bar{F}_{\theta_t}(\Gamma(u_{t-\tau_{\alpha_t}}))\right\rangle \\
&\leq \left\|\Gamma(u_{t-\tau_{\alpha_t}}) - w^*_{\theta_{t-\tau_{\alpha_t}}}\right\|\left\|\bar{F}_{\theta_t}(\Gamma(u_{t-\tau_{\alpha_t}})) - \bar{F}_{\theta_{t-\tau_{\alpha_t}}}(\Gamma(u_{t-\tau_{\alpha_t}}))\right\| \\
&\leq \frac{2(1 + L_w L_\theta \alpha_{t-\tau_{\alpha_t},t-1})}{A}\left(A\|\Gamma(u_t) - w^*_{\theta_t}\| + C\right) \\
&\quad \times L''_F L_\theta \alpha_{t-\tau_{\alpha_t},t-1}\left(\|\Gamma(u_{t-\tau_{\alpha_t}})\| + U''_F\right) \quad \text{(Using (31) and Lemma C.5)} \quad .
\end{aligned}$$

Using (34) completes the proof. □

D.9. Proof of Lemma A.9

Lemma D.9. *(Bound of T_5)*

$$T_5 \leq 2\left(A^2\left\|\Gamma(u_t) - w_{\theta_t}^*\right\|^2 + C^2\right).$$

Proof.

$$
\begin{aligned}
T_5 &= \left\|F_{\theta_t}(\Gamma(u_t), Y_t) - \Gamma(u_t)\right\|^2 \\
&\leq \left(\left\|F_{\theta_t}(\Gamma(u_t), Y_t)\right\| + \left\|\Gamma(u_t)\right\|\right)^2 \\
&\leq \left(U_F + (L_F + 1)\left\|\Gamma(u_t)\right\|\right)^2 \quad \text{(Lemma D.10)} \\
&\leq \left(B + A\left\|\Gamma(u_t)\right\|\right)^2 \\
&\leq \left(B + A\left\|\Gamma(u_t) - w_{\theta_t}^*\right\| + A\left\|w_{\theta_t}^*\right\|\right)^2 \\
&\leq 2\left(A^2\left\|\Gamma(u_t) - w_{\theta_t}^*\right\|^2 + C^2\right)
\end{aligned}
$$

$\square$

Lemma D.10. *For any θ, w, y,*

$$\left\|F_\theta(w, y)\right\| \leq U_F + L_F\|w\|$$

Proof. Assumption 3.5 implies that

$$
\begin{aligned}
\left\|F_\theta(w, y)\right\| - \left\|F_\theta(0, y)\right\| &\leq \left\|F_\theta(0, y) - F_\theta(w, y)\right\| \\
&\leq L_F\|w - 0\|,
\end{aligned}
$$

which completes the proof. $\square$

Lemma D.11.

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

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

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

$$\Pr\left(\tilde{Y}_t = y'\right)$$
$$= \sum_y \Pr\left(\tilde{Y}_{t-1} = y\right) P_{\theta_{t-\tau_{\alpha_t}}}(y, y')$$
$$= \sum_y \Pr\left(\tilde{Y}_{t-1} = y\right) P_{\theta_{t-\tau_{\alpha_t}}}(y, y') \sum_{z \in \Theta_t} \Pr(\theta_t = z | Y_{t-1} = y)$$
$$= \sum_y \sum_{z \in \Theta_t} \Pr\left(\tilde{Y}_{t-1} = y\right) P_{\theta_{t-\tau_{\alpha_t}}}(y, y') \Pr(\theta_t = z | Y_{t-1} = y)$$

Consequently,

$$\sum_{y'} \left| \Pr(Y_t = y') - \Pr\left(\tilde{Y}_t = y'\right) \right|$$
$$\leq \sum_{y,y'} \sum_{z \in \Theta_t} \left| \Pr(Y_{t-1} = y) P_z(y, y') - \Pr\left(\tilde{Y}_{t-1} = y\right) P_{\theta_{t-\tau_{\alpha_t}}}(y, y') \right| \Pr(\theta_t = z \mid Y_{t-1} = y).$$

Since for any $z \in \Theta_t$,

$$\left| \Pr(Y_{t-1} = y) P_z(y, y') - \Pr\left(\tilde{Y}_{t-1} = y\right) P_{\theta_{t-\tau_{\alpha_t}}}(y, y') \right|$$
$$\leq \left| \Pr(Y_{t-1} = y) P_z(y, y') - \Pr\left(\tilde{Y}_{t-1} = y\right) P_z(y, y') \right|$$
$$+ \left| \Pr\left(\tilde{Y}_{t-1} = y\right) P_z(y, y') - \Pr\left(\tilde{Y}_{t-1} = y\right) P_{\theta_{t-\tau_{\alpha_t}}}(y, y') \right|$$
$$\leq \left| \Pr(Y_{t-1} = y) - \Pr\left(\tilde{Y}_{t-1} = y\right) \right| P_z(y, y') + L_P L_\theta \alpha_{t-\tau_{\alpha_t}, t-1} \Pr\left(\tilde{Y}_{t-1} = y\right) \quad \text{(Lemma C.5)} \quad,$$

we have

$$\sum_{y'} \left| \Pr(Y_t = y') - \Pr\left(\tilde{Y}_t = y'\right) \right|$$
$$\leq \sum_y \left| \Pr(Y_{t-1} = y) - \Pr\left(\tilde{Y}_{t-1} = y\right) \right| + |\mathcal{Y}| L_P L_\theta \alpha_{t-\tau_{\alpha_t}, t-1}.$$

Applying the above inequality recursively yields

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

Consequently,

$$\left\| \mathbb{E}\left[F_{\theta_{t-\tau_{\alpha_t}}}\left(\Gamma(u_{t-\tau_{\alpha_t}}), Y_t\right) - F_{\theta_{t-\tau_{\alpha_t}}}\left(\Gamma(u_{t-\tau_{\alpha_t}}), \tilde{Y}_t\right) \right] \right\|$$
$$= \left\| \sum_y \left(\Pr(Y_t = y) - \Pr\left(\tilde{Y}_t = y\right) \right) F_{\theta_{t-\tau_{\alpha_t}}}\left(\Gamma(u_{t-\tau_{\alpha_t}}), y\right) \right\|$$
$$\leq \max_y \left\| F_{\theta_{t-\tau_{\alpha_t}}}\left(\Gamma(u_{t-\tau_{\alpha_t}}), y\right) \right\| |\mathcal{Y}| L_P L_\theta \sum_{j=t-\tau_{\alpha_t}}^{t-1} \alpha_{t-\tau_{\alpha_t}, j}$$
$$\leq 2|\mathcal{Y}| L_P L_\theta \sum_{j=t-\tau_{\alpha_t}}^{t-1} \alpha_{t-\tau_{\alpha_t}, j} \left(A \|\Gamma(u_t) - w_{\theta_t}^*\| + C \right) \quad \text{(Using (33))},$$

which completes the proof. $\qquad\square$