
Bridging the Gap Between Average and Discounted TD Learning

Anonymous Authors¹

Abstract

The analysis of Temporal Difference (TD) learning in the average-reward setting faces notable theoretical difficulties because the Bellman operator is not contractive with respect to any norm. This complicates standard analyses of stochastic updates that are effective in discounted settings. Although a considerable body of literature addresses these challenges, existing theoretical approaches come with limitations. We introduce a novel algorithm designed explicitly for policy evaluation in the average-reward setting, utilizing sampling from two Markovian trajectories. Our proposed method overcomes previous limitations by guaranteeing convergence to the unique solution of a properly defined projected Bellman equation. Notably, and in contrast to earlier work, our convergence analysis is uniformly applicable to both linear function approximation and tabular settings and does not involve explicit dimension-dependent terms in its convergence bounds. These results align with what is known to hold in the discounted setting. Furthermore, our algorithm achieves improved dependence on the problem’s condition number, reducing the sample complexity from quartic, as in prior literature, to quadratic scaling, and thus matching the efficiency seen in the discounted setting.

1. Introduction

Reinforcement learning with an average-reward objective is well-suited for applications that focus on the long-term performance over an infinite horizon. This framework has proved valuable in a variety of domains, including control systems (Prieto-Rumeau & Hernández-Lerma, 2008; Krishnamurthy, 2016), telecommunications (Bertsekas & Gallager, 2021; Altman, 2021), and production environ-

ments (Feinberg & Shwartz, 2012). While the discounted setting—emphasizing shorter-term returns—has received extensive theoretical attention, especially for Temporal-Difference (TD) learning (Bhandari et al., 2018; Chen et al., 2022; Tian et al., 2023; Srikant & Ying, 2019), analogous understanding of TD methods under the average-reward criterion is comparatively less developed (Tsitsiklis & Van Roy, 1999; Zhang et al., 2021b; Blaser & Zhang, 2024).

Technically, compared to the discounted setting the challenge in the average-reward setting is that the Bellman operator for policy evaluation is not a contraction mapping with respect to any norm. Consequently, the solution to the corresponding Bellman equation is not unique. This significantly complicates the theoretical analysis, particularly in the model-free setting where the stochastic dynamics of the Markov Decision Process (MDP) are unknown and the agent can only obtain random samples through interaction with the environment.

We now make the above remarks more concrete by surveying existing results on average-reward TD and comparing them to their counterparts in the discounted setting. The first theoretical analysis of average-reward TD, due to (Tsitsiklis & Van Roy, 1999), addressed the challenges discussed above by introducing an assumption that guarantees uniqueness of the solution to the Bellman equation. In the setting where the true value function $V_\pi(s)$ of a policy π is approximated as $V_\pi(s) \approx \phi(s)^\top \theta$, that is, as a linear combination of computable features stacked into the vector $\phi(s)$, (Tsitsiklis & Van Roy, 1999) assumed that for any scalar $c \in \mathbb{R}$ and vector $\theta \in \mathbb{R}^d$, we have $\Phi\theta \neq ce$, where Φ is the matrix whose rows are the feature vectors $\phi(s)^\top$, and e is the all-ones vector. While this assumption enabled the first convergence guarantee for average-reward TD, it is not satisfied even in the tabular case, where $\{\phi(s)\}$ form the canonical basis and Φ is the identity matrix. Several subsequent works (Yu & Bertsekas, 2009; Zhang et al., 2021a; Li et al., 2024) adopted the same assumption.

A more recent study (Blaser & Zhang, 2024) established that the value function under a particular update rule converges to a sample-path-dependent fixed point without requiring the assumption $\Phi\theta \neq ce$ for all θ , by leveraging the theory of Stochastic Krasnoselskii–Mann (SKM) iterations (Bravo & Cominetti, 2024). However, since the limit is sample-

¹Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

path-dependent, it may differ across independent runs of the algorithm. A similar result was obtained in the linear function approximation setting in (Zhang et al., 2021b), but without a guarantee that the process converges to a point. A more recent paper (Haque & Maguluri, 2024) establishes convergence to a point, but the convergence rate includes explicit dependence on the dimension d of the parameter vector θ , a dependence that does not explicitly appear in the discounted case.

The recent paper (Li et al., 2024) on actor-critic methods includes some results on policy evaluation in the linear approximation setting, using a fairly intricate nested-loop algorithm based on variance reduction. However, when specialized to the policy evaluation problem, that work also makes the $\Phi\theta \neq ce$ assumption, rendering it inapplicable to the tabular case. Moreover, we note that their sample complexity is stated in terms of the averaged iterate, whereas the other papers discussed above provide guarantees for the last iterate.

To summarize, while a number of papers have developed a convergence theory for average-reward TD, the existing literature consistently includes some mix of caveats compared to the discounted case, ranging from assumptions that exclude the tabular setting, to a lack of point-wise convergence guarantees for the underlying iterates, or explicit dimension dependence in the convergence bounds. Additionally, we note that the state-of-the-art results from (Zhang et al., 2021a; Haque & Maguluri, 2024; Chen et al., 2025) exhibit *quartic* dependence on the condition number, in contrast to the *quadratic* scaling achieved in the discounted setting (Bhandari et al., 2018). Alternatively, the result from (Li et al., 2024) does achieve quadratic scaling, but it relies on the assumption $\Phi\theta \neq ce$ which prevents it from being applied to the tabular case.

In this work, we take a complementary approach: observing that the average-reward value function must satisfy a certain steady-state constraint, we formulate the solution to the Bellman equation as a constrained optimization problem. To solve this problem, we propose a new algorithm that leverages sampling from two independent Markov chains in each iteration. We provide a finite-sample analysis for this method using a relatively new technique known as “gradient splitting” (Liu & Olshevsky, 2021). Finally, building upon ideas from the Gradient TD (GTD) (Sutton et al., 2008), we give a version of our algorithm which only uses a single Markov chain.

Our work improves upon the state of the art by simultaneously having all of the following features.

- **Provably unique, sample-independent fixed point.** Our analysis shows that the iterate sequence $\{\theta_t\}$ converges almost surely to a single, deterministic solution

θ^* that does not depend on the random trajectory or on initialization.

- **Tabular + Linear Function Approximation:** Our analysis is applicable to both the tabular and linear function approximation cases. In particular, we do not assume $e \notin \text{span}(\Phi)$.
- **Dependence on Dimensionality:** Our convergence bound does not have any explicit factors of d , the dimension of θ . While all algorithms have terms like $\|\theta^*\|$ that might implicitly scale with dimension, our algorithm has no terms scaling with d in addition to those.
- **Last Iterate:** our results are based on the last iterate rather than an iterate averaging, matching the corresponding results in the discounted case.
- **Scaling with Condition Number:** In Zhang et al. (2021b) and Haque & Maguluri (2024), the dependence of the convergence time on the condition number is quartic as $O(\eta_2^{-4})$, where η_2 is defined as

$$\eta_2 = \min_{\|x\|=1, x^\top e=0} \|\Phi x\|_{\text{Dir}}^2, \quad (1)$$

where $\|\cdot\|$ is the Euclidean norm and $\|\cdot\|_{\text{Dir}}$ denotes the Dirichlet seminorm, formally defined later in Section 2.3. Similarly, in Kim et al. (2025), the dependence of the convergence time on the condition number is also quartic as $O(\eta_2^{-3}\eta_3^{-1})$, where η_3 is defined as

$$\eta_3 = \left(\min_{\|x\|=1} x^\top \Phi^\top D \Phi x \right) \cdot \left(\min_{\langle y, e \rangle_D = 0, \|y\|_D = 1} y^\top D(I - P)y \right), \quad (2)$$

where D is the diagonal matrix with the stationary distribution μ of the policy on the diagonal. These contrast unfavorably with the corresponding results in the discounted case, which scale with the square rather than fourth power of the condition number (Bhandari et al., 2018). That being said, the definition of condition number is different within the discounted case, and it may be that a slightly different notion of condition number is needed for the average-reward setting.

Indeed, that is just what we show: for our algorithm convergence time scales quadratically as $O(1/\eta_1^2)$, where η_1 is defined as

$$\eta_1 = \min_{x: \|x\|=1} \|\Phi x\|_{\text{Dir}}^2 + (\mu^\top \Phi x)^2, \quad (3)$$

and μ is the stationary distribution of the policy.

Table 1. Comparison with previous work. We highlight several key distinctions between our work and prior research:

- (1) In the table below, **Linear Only** refers to the assumption that $e \notin \text{span}(\Phi)$, which rules out the tabular case.
- (2) A complicating factor in comparing results is that different papers tend to have slightly different notions of condition numbers; this is why the table contains η_1, η_2, η_3 (see definitions in Eq. (3) to Eq. (2)). A detailed discussion and precise definitions can be found in Appendix B. We remark that our condition number η_1 is at least as good as the widely used η_3 , i.e., $\eta_1 \geq \Omega(\eta_3)$, which means bounds based on η_1^{-1} are at least as good as bounds based on η_3^{-1} ; and if we assume the stationary probability μ is not too far from uniform, we also have $\eta_1 \geq \Omega(\eta_2)$. See Appendix B for details.
- (3) We consider a convergence time to be independent of the dimension d if its dependence on d appears only through $\|\theta_0\|$ or $\|\theta^*\|$.
- (4) We make the standard assumption that $\|\phi(s)\| \leq 1$ for all states s , which can be achieved by rescaling. If we instead make the assumption that every entry of $\phi(s)$ is $O(1)$ (so that $\|\phi(s)\| = O(\sqrt{d})$), then our scaling with dimension would be $O(d)$. This modified assumption matches more closely the assumption made in (Haque & Maguluri, 2024), which, unlike this work, analyzes the infinite-dimensional case. Our results are an improvement compared to the larger $\Omega(d^2)$ scaling in that work.
- (5) (Li et al., 2024) includes a scaling with approximation error ($\mathbb{E}\|W^* - \Phi\theta^*\|_D^2$) that no other paper has. In addition, we also note that they use a variance reduction method, not an analogue of plain TD; we do not implement variance reduction and our scaling with condition number is similar to their method.
- (6) In (Li et al., 2024), the condition number is actually defined as $\min_{y \neq e, \|y\|_D=1} y^\top D(I - P)y$. However, our analysis in Appendix B suggests that η_3 is actually greater than their condition number, offering an optimistic approximation of their sample complexity.
- (7) In (Kim et al., 2025), the sample complexity is actually $\tilde{O}(\epsilon^{-1} R_{\text{proj}}^2 \eta_2^{-3})$, where R_{proj} is the projection radius. However, the authors did not provide the choice of this radius. Our analysis in Appendix B shows that R_{proj}^2 is actually $O(1/\eta_3)$.
- (8) In the discounted case, $\eta_{\text{discounted}}$ can be defined as $(1 - \gamma)\sigma_{\min}(\Phi^T D \Phi)$ where γ is the discount factor (Bhandari et al., 2018)

REFERENCE	SETTING	CONVERGES TO A SAMPLE INDEPENDENT POINT	SAMPLE COMPLEXITY	SCALING WITH d	ITERATE	METHOD
(TSITSIKLIS & VAN ROY, 1999)	LINEAR ONLY	YES	$\times$	$\times$	LAST	COUPLED SA
(BLASER & ZHANG, 2024)	TABULAR ONLY	NO	$\times$	$\times$	LAST	COUPLED SA
(ZHANG ET AL., 2021B)	TABULAR & LINEAR	NO	$\tilde{O}(\epsilon^{-1} \eta_2^{-4})$	NO	LAST	COUPLED SA
(HAQUE & MAGULURI, 2024)	TABULAR & LINEAR	YES	$\tilde{O}(\epsilon^{-1} \eta_2^{-4})$	YES	LAST	COUPLED SA
(KIM ET AL., 2025)	TABULAR & LINEAR	NO	$\tilde{O}(\epsilon^{-1} \eta_2^{-3} \eta_3^{-1})$	NO	LAST	COUPLED SA
(CHEN ET AL., 2025)	TABULAR & LINEAR	NO	$\tilde{O}(\epsilon^{-1} \eta_2^{-4})$	NO	LAST	COUPLED SA
(LI ET AL., 2024)	LINEAR ONLY	YES	$\tilde{O}(\epsilon^{-1} \eta_3^{-2})$	NO	AVERAGE	VARIANCE REDUCTION
OUR TWO-CHAIN ALGORITHM	TABULAR & LINEAR	YES	$\tilde{O}(\epsilon^{-1} \eta_1^{-2})$	NO	LAST	COUPLED SA
DISCOUNTED CASE (E.G., (BHANDARI ET AL., 2018))	TABULAR & LINEAR	YES	$\tilde{O}(\epsilon^{-1} \eta_{\text{discounted}}^{-2})$	NO	LAST	SA

- **No variance reduction techniques needed.** While variance-reduction schemes are well-known to improve sample complexity, they typically require multi-level structure (e.g., nested loops, periodic full-batch or long-trajectory reference estimates, and additional bookkeeping) and therefore differ substantially from the “plain TD” template. In contrast, our quadratic condition-number scaling is achieved with a single-timescale, simple stochastic-approximation update without using any variance reduction techniques, matching the algorithmic simplicity of discounted TD.

2. Preliminaries

This section introduces the necessary background on average-reward reinforcement learning to support the algorithm design and convergence analysis of TD learning presented in the subsequent sections.

2.1. Markov Decision Processes (MDP)

We consider an MDP defined by the tuple $(S, A, P_{\text{env}}, r)$, where (i) S is the finite state space, (ii) A is the finite action space, (iii) $P_{\text{env}} = (P_{\text{env}}(s' | s, a))_{s, s' \in S, a \in A}$ is the transition probability kernel, and (iv) $r : S \times A \rightarrow \mathbb{R}$ is the reward function. Let $r_{\max} := \max_{s \in S, a \in A} |r(s, a)|$, which is finite since the state-action space is finite.

A policy $\pi : S \times A \rightarrow \mathbb{R}$ is a function where $\pi(a | s)$ represents the probability of the agent taking action a in state s . Throughout this paper, we focus exclusively on the policy evaluation problem and therefore assume the policy π to be fixed and known. Under this fixed policy, we define the induced transition matrix P as $P = (P(s' | s))_{s, s' \in S}$, where $P(s' | s) = \sum_{a \in A} P_{\text{env}}(s' | s, a) \pi(a | s)$.

We make the following assumption regarding the policy π , which is standard in TD learning (Liu & Olshevsky, 2021; Bhandari et al., 2018).

Assumption 2.1. The Markov chain with transition matrix P is irreducible and aperiodic.

Under the above assumption, the Markov chain with transition matrix P has a unique stationary distribution, denoted by μ , which satisfies $\mu_{\min} := \min_s \mu(s) > 0$ (Levin & Peres, 2017). Moreover, according to Theorem 4.9 in (Levin & Peres, 2017), there exist constants $C > 1$ and $\beta \in [0, 1)$ such that

$$\|p_\tau(\cdot|s) - \mu\|_1 \leq C\beta^\tau, \quad \forall \tau \geq 0, s \in S, \quad (4)$$

where p_τ is the probability distribution of the state of this Markov chain after τ steps starting at s .

2.2. The Long-Term Average Reward

We now discuss value functions within the average-reward framework, highlighting their role in policy evaluation. Let $r_s = \sum_a \pi(a|s) r(s, a)$ denote the expected reward in state s under policy π . We also define $p_{s_0 s}^k$ to be the probability that the agent is at state s after k steps starting from s_0 . Then the value function $v_{s_0}(t)$ reflects the expected cumulative reward starting from state s_0 after t transitions:

$$v_{s_0}(t) = r_{s_0} + \sum_s p_{s_0 s} r_s + \cdots + \sum_s p_{s_0 s}^{t-1} r_s.$$

Defining $V_t = [v_s(t)]_{s \in S}$ as the vector stacking up the value function and $R = [r_s]_{s \in S}$ as the vector stacking up the expected rewards, the above equation can be compactly written as $V_t = \sum_{k=0}^{t-1} P^k R$. Under Assumption 2.1, we have $\lim_{n \rightarrow \infty} P^n = e\mu^\top$ (Levin & Peres, 2017). Let $g = \mu^\top R$ be the steady-state reward per unit of time. Then, the relative value function, denoted by W^* , is defined as

$$W^* = \lim_{t \rightarrow \infty} V_t - tge = \lim_{t \rightarrow \infty} \sum_{k=0}^{t-1} (P^k - e\mu^\top) R. \quad (5)$$

Intuitively, the relative value function W^* quantifies the long-term expected cumulative reward differences across states relative to the steady-state reward. It is a central quantity for policy evaluation in the average-reward setting, as it effectively centers the rewards to focus purely on differences due to transient dynamics.

2.3. Useful Norms

In this subsection, we introduce several useful norms which will play an important role in our analysis. The so-called D -norm and Dirichlet semi-norm have been previously shown to be very useful in TD-like analysis (Ollivier, 2018; Liu & Olshevsky, 2021). Given a vector f with the same number of entries as the number of states in the MDP, its D -norm is defined as

$$\|f\|_D^2 = \langle f, f \rangle_D = \sum_s \mu(s) f(s)^2 \quad (6)$$

and its Dirichlet semi-norm is defined as

$$\|f\|_{\text{Dir}}^2 = \frac{1}{2} \sum_{s, s'} \mu(s) P(s'|s) (f(s) - f(s'))^2. \quad (7)$$

Intuitively, the Dirichlet seminorm measures the difference between f and the all-ones vector e , but in a way that is adapted to the Markov chain with transition matrix P . Finally, throughout the paper, we will use $\|\cdot\|$ to denote the standard Euclidean norm.

2.4. Markov Noise

Let s_t denote the state at time step t . Following standard practice, we consider two distinct sampling scenarios in this paper: (1) *i.i.d. sampling*, where each state s_t is independently drawn from the stationary distribution μ ; and (2) *Markov sampling*, where the Markov chain starts at s_0 and evolves according to the policy. Under Markov sampling, the states s_t remain marginally distributed according to μ , but exhibit temporal correlation across time steps.

2.5. Linear Function Approximation

In practical reinforcement learning applications, the state space S is often extremely large, making it impractical to maintain a vector whose dimension scales with the number of states. To address this challenge, it is common to incorporate function approximation, in particular, a linear function approximator of the form $W = \Phi\theta$, where $\Phi \in \mathbb{R}^{n \times d}$ is the feature matrix and $\theta \in \mathbb{R}^d$ is the parameter vector. Additionally, denote the s -th row of the feature matrix by $\phi(s)^\top$. We assume, without loss of generality, that (i) the features are normalized so that $\max_s \|\phi(s)\| \leq 1$, and (ii) the columns of Φ are linearly independent.

We further define

$$\eta = \min_{x: \|x\|=1} \|\Phi x\|_{\text{Dir}}^2 + (\mu^\top \Phi x)^2, \quad \text{where } x \in \mathbb{R}^d. \quad (8)$$

Intuitively, η measures how close to zero Φx can get: the first term measures the distance between Φx and the all-ones vector, whereas the second term measures the (squared) distance between the weighted average $\mu^\top \Phi x$ and zero. In particular, under the assumption that the columns of Φ are linearly independent, we immediately have $\eta > 0$. The quantity η will act as a condition number in our algorithms for average reward TD.

3. Algorithms

We now introduce the two algorithms studied in this paper. The first, called the **double-chain algorithm**, uses two independent Markov chains. The second, the **single-chain algorithm**, uses only one.

3.1. Double-Chain Algorithm: Motivation and Derivation

Recall that in the discounted setting, TD-learning is designed to solve the projected Bellman equation (Tsitsiklis &

Van Roy, 1996). To motivate our analysis, we next introduce a natural analogue of the projected Bellman equation in the average-reward setting.

Our goal is to compute W^* as defined in Eq.(5). It is well known (Gallager, 1997) that the relative value function W^* satisfies two properties:

$$W^* + ge = PW^* + R \quad \text{and} \quad \mu^\top W^* = 0.$$

Define $\Pi = I - e\mu^\top$ as a projection onto the subspace $\{x \mid \langle x, e \rangle_D = 0\}$ in the inner product $\langle \cdot, \cdot \rangle_D$ and the Bellman operator T_π such that $T_\pi W = R + PW$.

We can write the two properties into an equivalent way:

$$W^* = \Pi T_\pi W^*. \quad (9)$$

It is known that W^* is the unique solution to this equation (Gallager, 1997). Since the matrix $D = \text{diag}(\mu)$ is invertible, the previous equation is equivalent to

$$D(\Pi(R + PW^*) - W^*) = 0,$$

which can be further written as

$$D(R + PW^* - W^*) - \mu\mu^\top(R + W^*) = 0 \quad (10)$$

using the explicit definition of Π .

A natural approach to solve W^* from Eq.(10) is to recursively perform the update

$$W_{t+1} = W_t + \alpha_t \left(D(R + PW_t - W_t) - \mu\mu^\top(R + W_t) \right). \quad (11)$$

Although the above iterative algorithm seems promising, it cannot be implemented directly, since the transition matrix P and the reward function R are unknown. In the remainder of this section, we develop a data-driven stochastic version of the algorithm presented in Eq. (11). Before delving into the details, we first introduce some notation.

We use $1(s = s_0) \in \mathbb{R}^{|S|}$ to denote the vector whose entries are 0 except a 1 at position $s = s_0$. We also use $\mathbb{E}_\mu$ as expectation assuming that the state s_t is drawn from stationary distribution μ while s_{t+1} is still drawn according to the MDP with action taken according to policy π .

With this notation in place, we now describe the intuition behind the equations that we will write down. Keeping in mind that our goal is to provide a stochastic version of Eq. (11), the straightforward approach is to replace the unknown transition matrix P with something depending on samples that has expectation P . This works, but it is the second term in Eq. (11) that causes some trouble: it is surprisingly not straightforward to find a quantity such that its expectation is $\mu\mu^\top(R + W_t)$.

Indeed, to form a stochastic estimator for $\mu\mu^\top(R + W_t)$, observe that while $\mathbb{E}_\mu[1(s = s_t)] = \mu$ and $\mathbb{E}_\mu[r_{s_t} + W_t(s_t)] = \mu^\top(R + W_t)$, we cannot multiply these two estimators to obtain the result we want because $\mathbb{E}[XY] \neq \mathbb{E}[X]\mathbb{E}[Y]$ for random variables X and Y if they are not independent. This is known as the double sampling issue (Sutton et al., 2008). A natural and simple way to solve this issue is to sample two independent Markov Chains and base the two estimates on independent samples.

Denoting the state of these two chains by $\{s_t\}$ and $\{\hat{s}_t\}$, respectively, we therefore consider the following update:

$$W_{t+1} = W_t + \alpha_t (f(s_t, \hat{s}_t, W_t) + g(s_t, s'_t, W_t)), \quad (12)$$

where

$$\begin{aligned} f(s_t, \hat{s}_t, W_t)[s] &= -1(s = \hat{s}_t)(r_{s_t} + W_t(s_t)), \\ g(s_t, s'_t, W_t)[s] &= 1(s = s_t)(r_{s_t} + W_t(s'_t) - W_t(s_t)), \end{aligned}$$

It is then indeed immediate that

$$\begin{aligned} \mathbb{E}_\mu f(s_t, \hat{s}_t, W_t) &= -\mu\mu^\top(R + W_t) \\ \mathbb{E}_\mu g(s_t, s'_t, W_t) &= D(R + PW_t - W_t), \end{aligned}$$

and therefore Eq. (12) is a stochastic version of Eq. (11).

With linear function approximation, the natural generalization of Eq. (9) becomes

$$\Phi\theta^* = \Pi_D \Pi T_\pi \Phi\theta^* \quad (13)$$

where $\Pi_D = \Phi(\Phi^\top D\Phi)^{-1}\Phi^\top D$ is the projection onto the subspace spanned by the columns of Φ in D -inner product. Compared to Eq. (9), this equation uses $\Phi\theta$ as an approximation and adds a projection to the column space of Φ . We rewrite the above equation in the following form, which will be easier to implement:

$$\Phi^\top D(I - \Pi P)\Phi\theta^* = \Phi^\top D\Pi R. \quad (14)$$

The following lemma establishes the existence and uniqueness of θ^* whose proof can be found in Appendix A.

Lemma 3.1. *The solution to the linear system $\Phi^\top D(I - \Pi P)\Phi\theta = \Phi^\top D\Pi R$ exists and is unique.*

Based on this equation, a natural generalization from the tabular to the linear approximation case is therefore

$$\theta_{t+1} = \theta_t + \alpha_t (f(s_t, \hat{s}_t, \theta_t) + g(s_t, s'_t, \theta_t)), \quad (15)$$

where

$$\begin{aligned} f(s_t, \hat{s}_t, \theta_t) &= -(r_{s_t} + \phi(s_t)^\top \theta_t) \phi(\hat{s}_t) \\ g(s_t, s'_t, \theta_t) &= (r_{s_t} + \phi(s'_t)^\top \theta_t - \phi(s_t)^\top \theta_t) \phi(s_t). \end{aligned} \quad (16)$$

3.2. Single-Chain Algorithm

It is natural to wonder whether we can perform the update using only a single Markov chain. Inspired by the GTD method (Sutton et al., 2008), we propose a solution that does so.

Our single-chain algorithm is based on the following observation. Our two-chain algorithm uses the term $\phi(\hat{s}_t)$ – but what if, instead, we replace that by an estimate of the expectation $E[\phi(\hat{s}_t)]$? Because s_t is sampled from μ , this expectation equals $\Phi^\top \mu$. We will therefore introduce a new variable w_t which will converge to $\Phi^\top \mu$ and use it in place of $\phi(\hat{s}_t)$.

Our algorithm is thus as follows:

$$\begin{aligned} w_{t+1} &= \text{Proj}_{R_w} \{w_t + \beta_t f(s_t, w_t)\} \\ \theta_{t+1} &= \text{Proj}_{R_\theta} \{\theta_t + \alpha_t g(s_t, s'_t, w_t, \theta_t)\}, \end{aligned} \quad (17)$$

where

$$\begin{aligned} f(s_t, w_t) &= \phi(s_t) - w_t, \\ g(s_t, s'_t, w_t, \theta_t) &= (r_{s_t} + \phi(s'_t)^\top \theta_t - \phi(s_t)^\top \theta_t) \phi(s_t) - (r_{s_t} \\ &\quad + \phi(s_t)^\top \theta_t) w_t. \end{aligned} \quad (18)$$

Note that the first line clearly drives w_t to $\Phi^\top \mu$ while the second line is identical to the double-chain method except $\phi(\hat{s}_t)$ has been replaced by w_t .

To see where $\{w_t\}$ and $\{\theta_t\}$ will converge to in this case, we notice that, under i.i.d. sampling,

$$\begin{aligned} \mathbb{E}_{s_t \sim \mu} f(s_t, w_t) &= \Phi^\top \mu - w_t \\ \mathbb{E}_{s_t \sim \mu} g(s_t, s'_t, w_t, \theta_t) &= \Phi^\top D(R + P\Phi\theta_t - \Phi\theta_t) \\ &\quad - \mu^\top (R + \Phi\theta_t) w_t. \end{aligned}$$

Therefore, we can define w^* and θ^* such that

$$\begin{aligned} \Phi^\top \mu - w^* &= 0 \\ \Phi^\top D(R + P\Phi\theta^* - \Phi\theta^*) - \mu^\top (R + \Phi\theta_t) w^* &= 0. \end{aligned}$$

The intuition behind this method is that while w_t is derived from the same trajectory as s_t , it is an average over past features. We can therefore expect that it will be essentially de-correlated from the instantaneous value s_t , so we will be able to estimate $E[XY] \approx E[X]E[Y]$ up to some error for the product of $X = w_t$ and $Y = r_{s_t} + \phi(s_t)^\top \theta_t$. Naturally, this will come at the cost of an increased convergence time because of the additional error incurred.

4. Main Results

We now present our main results. We will consider both i.i.d. and Markov sampling, as well as both our double-chain and single-chain algorithms.

4.1. Convergence Results for the Double-Chain Method

Our first result assumes that the two Markov chains $\{s_t\}$ and $\{\hat{s}_t\}$ are sampled i.i.d. from the stationary distribution μ . We first consider the case where the stepsizes are constant $\alpha_t = \alpha$. We define τ_{mix} as the mixing time $\tau_{\text{mix}} = \tau_{\text{mix}}(\alpha)$. We also denote $\delta_t = \theta_t - \theta^*$. Our first theorem considers the double-chain method.

Theorem 4.1 (Double-chain, i.i.d. sampling). *Suppose Assumption 2.1 holds. Consider the double-chain algorithm (15) with i.i.d. sampling and constant stepsize $\alpha \leq \eta/18$. Then*

$$\mathbb{E} \|\theta_T - \theta^*\|^2 \leq e^{-\alpha \eta T} \|\theta_0 - \theta^*\|^2 + O\left(\frac{\alpha(r_{\max} + \|\theta^*\|)^2}{\eta}\right).$$

In particular, choosing $\alpha = \tilde{\Theta}(1/(\eta T))$ yields a convergence rate of $\tilde{O}(1/T)$ with sample complexity $\tilde{O}(\epsilon^{-1} \eta^{-2})$.

Our next theorem generalizes this result to Markov sampling, i.e., when $s_t, \hat{s}_t$ are sampled from two independent Markov chains. This requires further information on the mixing time of the Markov Chain as mentioned in Section 2.4. We still choose a constant stepsize. We will require the constants

$$\begin{aligned} B &= 2\|\theta_0 - \theta^*\| + r_{\max} + \|\theta^*\|, \\ G &= 42B^2 + 30(r_{\max} + \|\theta^*\|)^2. \end{aligned}$$

Theorem 4.2 (Double-chain, Markov sampling, constant stepsize). *Suppose Assumption 2.1 holds. Consider the double-chain algorithm (15) with Markov sampling and constant stepsize*

$$\alpha \leq \frac{\eta B^2}{(3\tau_{\text{mix}} + 1)G}.$$

Then for all $T \geq \tau_{\text{mix}}$,

$$\mathbb{E} \|\theta_T - \theta^*\|^2 \leq e^{-2\alpha \eta (T - \tau_{\text{mix}})} B^2 + \frac{\alpha G (3\tau_{\text{mix}} + 1)}{2\eta}.$$

Choosing $\alpha = \tilde{\Theta}(1/(\eta T))$ yields a convergence rate of $\tilde{O}(1/T)$ with sample complexity $\tilde{O}(\epsilon^{-1} \eta^{-2})$.

The rates of both of these theorems match the state-of-the-art for TD learning in the discounted case (Bhandari et al., 2018) in terms of the scaling with the various parameters.

A similar result can be obtained with a decaying stepsize $\alpha_t = a/(t + c_0)^\xi$. In this case, the mixing time is defined as $\tau_{\text{mix}} = \tau_{\text{mix}}(\alpha_T)$. The next theorem formally handles this case.

Theorem 4.3 (Double-chain, Markov sampling, decaying stepsize). *Suppose Assumption 2.1 holds. Consider the double-chain algorithm (15) with Markov sampling and stepsize $\alpha_t = a/(t + c_0)^\xi$ for $\xi \in (0, 1]$, $a > 0$, and c_0 sufficiently large. Then for all $T \geq \tau_{\text{mix}}$:*

1. If $\xi = 1$, then

$$\mathbb{E}\|\theta_T - \theta^*\|^2 \leq O(1) \cdot \left(\frac{\tau_{\text{mix}} + c_0}{T + c_0} \right)^{a\eta} + \tilde{O}\left(\frac{a^2}{(T + c_0)^{\min\{1, a\eta\}}} \right).$$

2. If $\xi \in (0, 1)$, let $\Delta_T = (T + c_0)^{1-\xi} - (\tau_{\text{mix}} + c_0)^{1-\xi}$ then

$$\mathbb{E}\|\theta_T - \theta^*\|^2 \leq O\left(\exp\left(-\frac{\eta a}{1-\xi} \Delta_T \right) \right) + O\left(\frac{a}{\eta(T + c_0)^\xi} \right).$$

Choosing $a = \Theta(1/\eta)$ yields sample complexity $\tilde{O}(\epsilon^{-1/\xi} \eta^{-2/\xi})$ which reduces to $\tilde{O}(\epsilon^{-1} \eta^{-2})$ when $\xi = 1$. The $O(1)$ terms depend polynomially on $\|\theta_0 - \theta^*\|$, r_{max} , and $\|\theta^*\|$.

This is once again consistent with the state-of-the-art results in the discounted setting (Bhandari et al., 2018), where an additional $O(1/\eta)$ factor is introduced when the distance is measured in the parameter space.

4.2. Convergence Results for the Single-Chain Method

For the single-chain algorithm given by Eq. (17), we can also establish similar results, albeit with worse scaling with respect to the condition number. We denote $\delta_t^\theta = \theta_t - \theta^*$ and $\delta_t^w = w_t - w^*$. We define τ_{mix} as the mixing time $\tau_{\text{mix}} = \tau_{\text{mix}}(\min\{\alpha, \beta\})$.

Theorem 4.4 (Single-chain, constant stepsize, Markov sampling). *Consider the single-chain algorithm in Eq. (17) with Markov sampling. Assume constant stepsizes $\alpha_t \equiv \alpha > 0$, $\beta_t \equiv \beta > 0$, and let $\rho_0 := \beta/\alpha \leq 1$. Let $\tau_{\text{mix}} := \tau_{\text{mix}}(\min\{\alpha, \beta\})$ denote the mixing time at accuracy level $\min\{\alpha, \beta\}$. Let $0 < \alpha < \frac{1}{2\zeta}$ and let $\lambda > 0$ satisfy*

$$0 < \lambda^2 < \frac{2\eta}{r_{\text{max}} + 2R_\theta},$$

$$\zeta := \eta - \frac{\lambda^2(r_{\text{max}} + 2R_\theta)}{2} > 0,$$

and define

$$\kappa := 1 - 2\alpha\zeta \in (0, 1), \quad G_1 := e^{-2\rho_0\alpha} \in (0, 1),$$

where $\rho := \max\{\kappa, G_1\} \in (0, 1)$. Then for all $t \geq \tau_{\text{mix}}$,

$$\mathbb{E}\|\theta_t - \theta^*\|^2 \leq \frac{4\alpha(r_{\text{max}} + 2R_\theta)(t - \tau_{\text{mix}})\rho^{t-\tau_{\text{mix}}}}{\lambda^2} + 4R_\theta^2 \kappa^{t-\tau_{\text{mix}}} + \frac{\alpha G_{\text{const}}}{2\zeta},$$

and where

$$G_{\text{const}} := \frac{(16\tau_{\text{mix}} + 6)(r_{\text{max}} + 2R_\theta)}{\lambda^2} + (88\tau_{\text{mix}} + 4)(r_{\text{max}} + 3R_\theta)^2.$$

Sample Complexity. As discussed in Appendix A.3, the projection radii can be chosen so that $R_\theta = O(\eta'^{-1/2})$ and $R_w = O(1)$, where $\eta' = \lambda_{\min}(\Phi^\top D \Phi)$. Moreover, Appendix E shows one may select $\lambda^2 = \Theta(\eta'^{1/2}\eta)$, which implies $\zeta = \Theta(\eta)$ (e.g., by taking λ^2 a fixed fraction of $2\eta/(r_{\text{max}} + 2R_\theta)$). With these choices, the dominant part of G_{const} comes from the $(r_{\text{max}} + 2R_\theta)/\lambda^2$ term and scales as $G_{\text{const}} = \tilde{O}(\tau_{\text{mix}}/(\eta'\eta))$ (up to polynomial factors in r_{max} and R_θ), so the steady-state error term satisfies

$$\frac{\alpha G_{\text{const}}}{\zeta} = \tilde{O}\left(\frac{\alpha \tau_{\text{mix}}}{\eta' \eta^2} \right).$$

Thus, choosing $\alpha = \tilde{\Theta}(1/(\eta T))$ yields a $\tilde{O}(1/T)$ rate (as in the two-chain case), but the overall condition-number dependence is worse than quadratic: plugging $\alpha = \tilde{\Theta}(1/(\eta T))$ into the steady-state term gives $\tilde{O}(\tau_{\text{mix}}/(\eta' \eta^3 T))$, which becomes quartic $\tilde{O}(1/(\eta^4 T))$ in the common regime where $\eta' = \Theta(\eta)$ (the precise relation between η' and η is discussed in Appendix B).

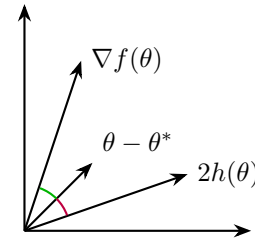


Figure 1. Illustration of key property of the gradient splitting. The splitting vector $h(\theta)$ has the same inner product (up to a factor of 2) as the true gradient with the vector $\theta - \theta^*$.

Finally, we can also give a similar theorem for the single-chain method with decaying step-size

$$\alpha_t = \frac{a}{(t + c_0)^\xi},$$

which parallels Theorem 4.3 for the double-chain algorithm but with the η^{-2} scaling replaced with η^{-4} . For reasons of space, we state this in the appendix as Theorem F.1.

Remark: Some previous works add an additional $(r_t - g)^2$ to the error measure, e.g., (Zhang et al., 2021b; Haque & Maguluri, 2024). We observe that using $r_t = (\sum_{i=0}^{t-1} r_{S_i})/t$ as a stochastic estimate of g allows us to obtain

$$\mathbb{E}[(r_t - g)^2] = O(1/T). \quad (19)$$

Thus, this quantity could easily be estimated separately without affecting the results of this section. A detailed discussion is in Appendix B.

4.3. Proof Idea

Our main observation is that the methods we propose can be analyzed in much the same way as standard *Stochastic Gradient Descent (SGD)*, which is usually more tractable than stochastic approximation-based methods. The critical tool enabling this perspective is the notion of a *gradient splitting*, introduced by (Liu & Olshevsky, 2021).

To illustrate the idea, let us consider a convex quadratic function $f(\theta) = (\theta - \theta^*)^T A (\theta - \theta^*)$, where A is symmetric and positive definite; and let us also consider a linear function $h(\theta) = B (\theta - \theta^*)$. We will say $h(\theta)$ is a gradient splitting of $f(\theta)$ if $B + B^T = 2A$. In other words, each (i, j) -entry of A can be split between the (i, j) and (j, i) -entries of the (generally non-symmetric) matrix B . This decomposition is not unique, since there are many ways to split the entries of A .

An immediate implication of this definition is that for any θ , $(\theta^* - \theta)^T \nabla f(\theta) = 2(\theta^* - \theta)^T h(\theta)$. Hence, updating in the direction $-h(\theta)$ behaves just like an update in the direction of the true gradient $-\nabla f(\theta)$. An illustration is given in Figure 1. One can further conclude

$$\begin{aligned} 2(\theta - \theta^*)^T h(\theta) &= (\theta^* - \theta)^T \nabla f(\theta) \\ &= (\theta^* - \theta)^T (\nabla f(\theta) - \nabla f(\theta^*)) \\ &= 2f(\theta), \end{aligned}$$

where the last equality holds because $f(\theta)$ is a quadratic function.

Concretely, the update rule

$$\theta_{t+1} = \theta_t - \alpha_t [h(\theta_t) + w_t], \quad (20)$$

where w_t is zero-mean i.i.d. noise, exhibits convergence properties analogous to classical SGD. In fact, using a Taylor expansion,

$$\begin{aligned} E[\|\theta_{t+1} - \theta^*\|_2^2 | \theta_t] &= \\ &= \|\theta_t - \theta^*\|_2^2 - \alpha_t \nabla f(\theta_t)^T (\theta_t - \theta^*) + O(\alpha_t^2). \end{aligned}$$

Notice that the middle term is precisely what one would get with a standard gradient step, even though (20) does not explicitly use $\nabla f(\theta_t)$. Thus, standard SGD analysis applies, with adjustments for the different higher-order $O(\alpha_t^2)$ terms.

Building on these ideas, our contribution in this paper is to show that an algorithm for the average reward case can also be written as a gradient splitting. Specifically, consider Eq. (15), where the expected update direction is $E_\mu[f(s_t, \hat{s}_t) + g(s_t, \hat{s}_t)]$, with notation as in Eq. (15). We can rewrite this expected update as $\Phi^T D (R + P\Phi\theta - \Phi\theta) - \Phi^T \mu \mu^T (R + \Phi\theta)$; our key technical observation is to show that it serves as a gradient splitting for the composite function

$$\|\Phi(\theta - \theta^*)\|_{\text{Dir}}^2 + (\mu^T \Phi(\theta - \theta^*))^2.$$

Once this gradient-splitting viewpoint is in place, we can leverage the standard SGD descent recursion, but closing the argument requires new bounds that control the discrepancy between our stochastic update and the true gradient—specifically, we must show the resulting $O(\alpha_t^2)$ and bias terms remain uniformly small/summable under the chosen stepsize.

5. Numerical Results

We compare our algorithms with prior work on fifteen tabular MDPs from OpenAI Gym and MO-Gymnasium. We focus on tabular environments because the exact solution can be computed, allowing us to directly quantify how accurately each method approximates the true solution. This enables a fair comparison between algorithms that converge to a single point and those that converge only to a set, since we evaluate all methods using their approximation error relative to the true solution rather than properties of their iterates.

Across all fifteen environments, we evaluate both our Double-Chain and Single-Chain algorithms, together with the baselines listed in Table 1. Overall, we observe that our methods outperform the prior literature in most settings. Additional experimental details are provided in Appendix G.

6. Conclusion

The theoretical analysis of average reward TD has traditionally lagged behind the discounted setting due to mathematical difficulties, notably the non-contractive nature of the Bellman operator. This disparity manifested in prior works through various limitations: assumptions incompatible with the tabular case, convergence guarantees only to sets or sample-dependent points, explicit dimension scaling, and slower (quartic) convergence rates with respect to condition numbers – which, unlike in the discounted case, could only be removed with more intricate algorithms like variance reduction.

In this work, we employed the gradient splitting technique to provide a finite-sample analysis that closes the gap with discounted TD theory. Our methods guarantee convergence to a unique, well-defined solution for both tabular and linear approximation cases. Furthermore, the convergence bounds are dimension-free (in the standard sense) and exhibit quadratic scaling with the relevant condition number, mirroring the performance characteristics known for discounted TD. This contribution removes the persistent caveats associated with average reward TD analysis.

References

- Altman, E. *Constrained Markov decision processes*. Routledge, 2021.
- Bertsekas, D. and Gallager, R. *Data networks*. Athena Scientific, 2021.
- Bhandari, J., Russo, D., and Singal, R. A finite time analysis of temporal difference learning with linear function approximation. In *Conference on learning theory*, pp. 1691–1692. PMLR, 2018.
- Blaser, E. and Zhang, S. Almost sure convergence of average reward temporal difference learning. *arXiv preprint arXiv:2409.19546*, 2024.
- Bravo, M. and Cominetti, R. Stochastic fixed-point iterations for nonexpansive maps: Convergence and error bounds. *SIAM Journal on Control and Optimization*, 62(1):191–219, 2024.
- Chen, Z., Zhang, S., Doan, T. T., Clarke, J.-P., and Maguluri, S. T. Finite-sample analysis of nonlinear stochastic approximation with applications in reinforcement learning. *Automatica*, 146:110623, 2022.
- Chen, Z., Zhang, S., Zhang, Z., Haque, S. U., and Maguluri, S. T. A non-asymptotic theory of seminorm lyapunov stability: From deterministic to stochastic iterative algorithms. *arXiv preprint arXiv:2502.14208*, 2025.
- Feinberg, E. A. and Shwartz, A. *Handbook of Markov decision processes: methods and applications*, volume 40. Springer Science & Business Media, 2012.
- Felten, F., Alegre, L. N., Nowé, A., Bazzan, A. L. C., Talbi, E. G., Danoy, G., and Silva, B. C. da. A toolkit for reliable benchmarking and research in multi-objective reinforcement learning. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023)*, 2023.
- Gallager, R. G. Discrete stochastic processes. *Journal of the Operational Research Society*, 48(1):103–103, 1997.
- Haque, S. U. and Maguluri, S. T. Stochastic approximation with unbounded Markovian noise: A general-purpose theorem. *arXiv preprint arXiv:2410.21704*, 2024.
- Kim, H., Cho, D. D., and Laber, E. Implicit updates for average-reward temporal difference learning. *arXiv preprint arXiv:2510.06149*, 2025.
- Krishnamurthy, V. *Partially observed Markov decision processes*. Cambridge University Press, 2016.
- Levin, D. A. and Peres, Y. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.
- Li, T., Wu, F., and Lan, G. Stochastic first-order methods for average-reward Markov decision processes. *Mathematics of Operations Research*, 2024.
- Liu, R. and Olshevsky, A. Temporal difference learning as gradient splitting. In *International Conference on Machine Learning*, pp. 6905–6913. PMLR, 2021.
- Meyn, S. P. and Tweedie, R. L. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.
- Ollivier, Y. Approximate temporal difference learning is a gradient descent for reversible policies. *arXiv preprint arXiv:1805.00869*, 2018.
- Prieto-Rumeau, T. and Hernández-Lerma, O. Ergodic control of continuous-time Markov chains with pathwise constraints. *SIAM journal on control and optimization*, 47(4):1888–1908, 2008.
- Srikant, R. and Ying, L. Finite-time error bounds for linear stochastic approximation and TD learning. In *Conference on Learning Theory*, pp. 2803–2830. PMLR, 2019.
- Sutton, R. S., Maei, H., and Szepesvári, C. A convergent $o(n)$ temporal-difference algorithm for off-policy learning with linear function approximation. *Advances in neural information processing systems*, 21, 2008.
- Tian, H., Paschalidis, I. C., and Olshevsky, A. On the performance of temporal difference learning with neural networks. *arXiv preprint arXiv:2312.05397*, 2023.
- Towers, M., Kwiatkowski, A., Terry, J., Balis, J. U., De Cola, G., Deleu, T., Goulão, M., Kallinteris, A., Krimmel, M., KG, A., et al. Gymnasium: A standard interface for reinforcement learning environments. *arXiv preprint arXiv:2407.17032*, 2024.
- Tsitsiklis, J. and Van Roy, B. Analysis of temporal-difference learning with function approximation. *Advances in neural information processing systems*, 9, 1996.
- Tsitsiklis, J. N. and Van Roy, B. Average cost temporal-difference learning. *Automatica*, 35(11):1799–1808, 1999.
- Yu, H. and Bertsekas, D. P. Convergence results for some temporal difference methods based on least squares. *IEEE Transactions on Automatic Control*, 54(7):1515–1531, 2009.
- Zhang, S., Wan, Y., Sutton, R. S., and Whiteson, S. Average-reward off-policy policy evaluation with function approximation. In *International Conference on Machine Learning*, pp. 12578–12588. PMLR, 2021a.

Zhang, S., Zhang, Z., and Maguluri, S. T. Finite sample
analysis of average-reward TD learning and Q -learning.
Advances in Neural Information Processing Systems, 34:
1230–1242, 2021b.

A. Bellman Operator and Projection

In this section, we establish the contractivity of ΠT_π and prove the existence and uniqueness of the solution to Eq.(9). Based on the contraction factor, we subsequently determine the choice of the projection radius used in the single-chain algorithm.

A.1. Contraction of ΠT_π

In this section, we prove that ΠT_π is a contractive operator (recall that $\Pi = I - e\mu^T$).

Lemma A.1. *The operator ΠT_π is a contractive operator satisfying*

$$\|\Pi T_\pi W_1 - \Pi T_\pi W_2\|_D \leq \omega \|W_1 - W_2\|_D,$$

where $\omega = \sqrt{\max_{\langle z, e \rangle_D = 0, \|z\|_D = 1} z^\top P^\top D P z} < 1$.

Proof of Lemma A.1. First, we notice that, since P is irreducible according to Assumption 2.1, the eigenvector whose eigenvalue is 1 is unique (up to a constant factor) and must be the all-one vector.

For any vector $\Delta = W_1 - W_2$, it can be decomposed into $\Delta = \Delta_\parallel + \Delta_\perp$ where $\langle \Delta_\perp, e \rangle_D = 0$ and $\langle \Delta_\parallel, e \rangle_D = \|\Delta_\parallel\|_D \cdot \|e\|_D$. With such decomposition,

$$\|\Pi T_\pi W_1 - \Pi T_\pi W_2\|_D = \|\Pi P \Delta_\perp\|_D.$$

Therefore,

$$\|\Pi T_\pi W_1 - \Pi T_\pi W_2\|_D^2 \leq \max_{\langle \Delta_\perp, e \rangle_D = 0} \Delta_\perp^\top P^\top D P \Delta_\perp \leq \max_{\langle z, e \rangle_D = 0, \|z\|_D = 1} z^\top P^\top D P z \cdot \|\Delta\|_D^2,$$

which indicates the contraction factor is

$$\omega := \sqrt{\max_{\langle z, e \rangle_D = 0, \|z\|_D = 1} z^\top P^\top D P z}.$$

It is easy to see that this factor cannot be larger than 1. If it is exactly 1, then there exists z such that

$$\|z\|_D = 1 = \max_{\langle z, e \rangle_D = 0, \|z\|_D = 1} z^\top P^\top D P z = \|Pz\|_D.$$

This indicates that z must be a multiple of all-one vector, which contradicts with $\langle z, e \rangle_D = 0$. $\square$

A.2. Existence and Uniqueness of θ^*

Recall that in Lemma 3.1, we define θ^* as the solution to the linear system $\Phi^\top D(I - \Pi P)\Phi\theta = \Phi^\top D R$ and we claim that θ^* exists and is unique. In this section, we give the proof of this lemma.

Proof of Lemma 3.1. Recall that in Eq.(8) we already established that

$$\eta = \min_{x: \|x\|=1} \|\Phi x\|_{\text{Dir}}^2 + (\mu^\top \Phi x)^2 > 0.$$

Observe that

$$\|\Phi x\|_{\text{Dir}}^2 + (\mu^\top \Phi x)^2 = x^\top \Phi^\top D(I - P)\Phi x + x^\top \Phi^\top \mu \mu^\top \Phi x = x^\top \Phi^\top D(I - \Pi P)\Phi x.$$

Suppose $\Phi^\top D(I - \Pi P)\Phi$ is not invertible. Then there exists a nonzero vector x_0 such that $x_0^\top \Phi^\top D(I - \Pi P)\Phi x_0 = 0$, which contradicts the fact that $\eta > 0$. Therefore, the matrix $\Phi^\top D(I - \Pi P)\Phi$ must be invertible. It follows that the linear system admits a unique solution θ^* . $\square$

A.3. Choice of Projection Radius

Recall that in our single chain algorithm, we need to project both w and θ onto a ball of radius R_w and R_θ , respectively. In this section, we will discuss on the choice of radius such that w^* and θ^* are in the feasible set.

Bound on $\|w^*\|$: This is easier because we can explicitly write $w^* = \Phi^\top \mu$. Using the fact that $\|\phi(s)\| \leq 1$ we can conclude $\|w^*\| \leq 1$. Therefore, we need $R_w \geq 1$.

Bound on $\|\theta^*\|$: Recall that W^* is defined as

$$W^* = \lim_{t \rightarrow \infty} \sum_{k=0}^{t-1} (P^k - e\mu^\top)R.$$

According to Eq.(4),

$$([P^k]_i - \mu^\top)R \leq \|[P^k]_i - \mu^\top\|_1 \cdot \|R\|_\infty \leq r_{\max} C \beta^k,$$

which indicates $\|W^*\|_D \leq r_{\max} C / (1 - \beta)$.

By the Pythagorean theorem,

$$\begin{aligned} \|\Phi\theta^* - W^*\|_D^2 &= \|W^* - \Pi_D \Pi T_\pi W^*\|_D^2 + \|\Pi_D \Pi T_\pi W^* - \Phi\theta^*\|_D^2 \\ &\leq \|W^* - \Pi_D \Pi T_\pi W^*\|_D^2 + \omega^2 \|\Phi\theta^* - W^*\|_D^2, \end{aligned}$$

where we use the fact $\Pi T_\pi \Phi\theta^* = \Phi\theta^*$ and Lemma A.1. Therefore,

$$\|\Phi\theta^* - W^*\|_D \leq \frac{1}{\sqrt{1 - \omega^2}} \|W^* - \Pi_D \Pi T_\pi W^*\|_D.$$

Further, according to the Pythagorean theorem, $\|W^* - \Pi_D \Pi T_\pi W^*\|_D \leq \|W^*\|_D$. Therefore,

$$\|\Phi\theta^*\|_D \leq \|W^*\|_D + \|\Phi\theta^* - W^*\|_D \leq \frac{2}{\sqrt{1 - \omega^2}} \|W^*\|_D \leq \frac{2r_{\max} C}{(1 - \beta)\sqrt{1 - \omega^2}}.$$

This suggests that

$$\|\theta^*\| \leq \frac{1}{\sqrt{\lambda_{\min}(\Phi^\top D \Phi)}} \|\Phi\theta^*\|_D \leq \frac{2r_{\max} C}{(1 - \beta)\sqrt{(1 - \omega^2)\lambda_{\min}(\Phi^\top D \Phi)}}.$$

Therefore, to ensure θ^* is in the feasible set, we can set R_θ such that

$$R_\theta \geq \frac{2r_{\max} C}{(1 - \beta)\sqrt{(1 - \omega^2)\lambda_{\min}(\Phi^\top D \Phi)}}.$$

B. Comparison with Previous Works

In this section, we will expand our result on reward estimation and compare the difference between condition numbers with previous works.

B.1. Reward Estimation

Many previous works (Zhang et al., 2021b; Haque & Maguluri, 2024) also include convergence of the averaged reward function. Although not stated in our theorems, we can also achieve such convergence as mentioned in Eq.(19). In this section, we give a proof of Eq.(19) based on a Central Limit Theorem for Markov chains.

Proof. For all $t = \tau_{\text{mix}}, \dots, T$, we can decompose $\mathbb{E}[(r_t - g)^2]$ as

$$\mathbb{E}[(r_t - g)^2] \leq 2\mathbb{E}[(r_t - \mathbb{E}[r_t])^2] + 2\mathbb{E}[(\mathbb{E}[r_t] - g)^2] = 2\text{Var}(r_t) + 2(\mathbb{E}[r_t] - g)^2.$$

First, we bound the variance of r_t , $\text{Var}(r_t) = \mathbb{E}[(r_t - g)^2]$. Let $\gamma_k = \text{Cov}(r_{S_0}, r_{S_k})$ be the covariance function. It is well known that

$$\text{Var}(r_t) = \frac{1}{t^2} \sum_{i=0}^{t-1} \sum_{j=0}^{t-1} \gamma_{|i-j|}.$$

By changing the order of summation,

$$\sum_{i=0}^{t-1} \sum_{j=0}^{t-1} \gamma_{|i-j|} = \sum_{i=-(t-1)}^{t-1} (t - |i|) \gamma_{|i|} \leq n \sum_{-\infty}^{\infty} \gamma_i = t \left(\gamma_0 + 2 \sum_{i=1}^{\infty} \gamma_i \right).$$

Since all Markov chains considered in this paper is V -uniformly ergodic with $V = 1$ and the reward function is also bounded since the Markov chain is finite, the conditions of Theorem 17.0.1 in (Meyn & Tweedie, 2012) hold (up to a constant factor which will not change the final result) and therefore, the term $\gamma^2 := (\gamma_0 + 2 \sum_{i=1}^{\infty} \gamma_i)$ must be finite. We further conclude

$$\text{Var}(r_t) \leq \frac{1}{t} \gamma^2 = O\left(\frac{1}{t}\right).$$

Next, we bound the other term on the right hand side, $2(\mathbb{E}[r_t] - g)^2$, using Eq.(4). Since each reward is at most r_{max} , the averaged reward g can also be at most r_{max} . Therefore,

$$|\mathbb{E}[r_{\tau_{\text{mix}}}] - g| = \frac{1}{\tau_{\text{mix}}} \sum_{i=0}^{\tau_{\text{mix}}-1} (r_{S_i} - g) \leq 2r_{\text{max}}.$$

Since we also have $r_t = \tau_{\text{mix}} r_{\tau_{\text{mix}}} / t + \sum_{i=\tau_{\text{mix}}}^{t-1} r_t / t$,

$$\begin{aligned} |\mathbb{E}[r_t] - g| &\leq \frac{\tau_{\text{mix}}}{t} |\mathbb{E}[r_{\tau_{\text{mix}}}] - g| + \frac{1}{t} \sum_{i=\tau_{\text{mix}}}^{t-1} |\mathbb{E}[r_i] - g| \\ &\leq \frac{2\tau_{\text{mix}}}{t} r_{\text{max}} + \frac{r_{\text{max}}}{t} \sum_{i=\tau_{\text{mix}}}^{t-1} \|P_i(\cdot | s_0) - \mu\|_1 \\ &\leq \tilde{O}\left(\frac{1}{t}\right). \end{aligned}$$

Therefore, combining both bounds and we conclude that

$$\mathbb{E}[(r_T - g)^2] = O(1/T).$$

□

B.2. Condition Numbers

We notice that, in (Zhang et al., 2021b) and (Li et al., 2024), the definitions on condition number are both different from ours. Our condition number is defined by

$$\eta_1 = \min_{\|x\|=1} \{ \|\Phi x\|_{\text{Dir}}^2 + (\mu^\top \Phi x)^2 \}.$$

Meanwhile, in (Zhang et al., 2021b),

$$\eta_2 = \min_{\|x\|=1, x^\top e=0} \|\Phi x\|_{\text{Dir}}^2$$

and in (Li et al., 2024; Kim et al., 2025),

$$\eta_3 = \left(\min_{\|x\|=1} x^\top \Phi^\top D \Phi x \right) \cdot \left(\min_{\langle y, e \rangle_D = 0, \|y\|_D = 1} y^\top D(I - P)y \right).$$

Therefore, in this section we will describe this difference.

Remark: In (Li et al., 2024), the condition number is actually defined as $\min_{y \neq e, \|y\|_D = 1} y^\top D(I - P)y$. However, we can always decompose it as $y = y_\perp + y_\parallel$ where $\langle y_\perp, e \rangle_D = 0$ and $\langle y_\parallel, e \rangle_D = \|y_\parallel\|_D \cdot \|e\|_D$. We can easily check $y^\top D(I - P)y = y_\perp^\top D(I - P)y_\perp$. Therefore, η_3 is actually greater than the condition number defined in (Li et al., 2024), offering an optimistic approximation of their sample complexity.

B.2.1. DIFFERENCE BETWEEN η_1 AND η_2

Lemma B.1. *We have*

$$\eta_1 \geq \frac{\mu_{\min}}{n\mu_{\max}\|\mu\|^2} \eta_2.$$

Notice that η_1 defined in the above lemma corresponds to the condition number in this paper and η_2 corresponds to the condition number in (Zhang et al., 2021b). We can conclude that when μ is a multiple of the all-one vector, then η_1 is no less than η_2 , which suggests our result is better. In other cases it is difficult to compare them.

Now we provide the proof to Lemma B.1.

Proof of Lemma B.1. It is useful to define the μ -weighted inner product and norm by

$$\langle x, y \rangle_\mu = \sum_{s \in \mathcal{S}} \mu(s) x(s) y(s), \quad \|x\|_\mu = \sqrt{\langle x, x \rangle_\mu}.$$

We first define η'_1, η'_2 to be

$$\eta'_1 = \min_{\|x\|_\mu=1} \{ \|x\|_{\text{Dir}}^2 + (\mu^\top x)^2 \},$$

and

$$\eta'_2 = \min_{\|x\|_\mu=1, \mu^\top x=0} \|x\|_{\text{Dir}}^2.$$

We can show a simple fact that $\eta'_1 \geq \eta'_2$ when $\eta'_2 \leq 1$. We notice that the Dirichlet semi-norm is invariant under addition of a constant. In other words, for any $x \in \mathbb{R}^n$ and $c \in \mathbb{R}$,

$$\|x + ce\|_{\text{Dir}} = \|x\|_{\text{Dir}}.$$

Therefore, we can decompose

$$x = ce + z, \quad \text{where } c \in \mathbb{R} \text{ and } \mu^\top z = 0.$$

On one hand, we obtain $\|x\|_{\text{Dir}} = \|z\|_{\text{Dir}}$. On the other hand,

$$\|x\|_\mu^2 = \|ce\|_\mu^2 + \|z\|_\mu^2 = c^2 \|e\|_\mu^2 + \|z\|_\mu^2 = c^2 + \|z\|_\mu^2 = 1.$$

Therefore,

$$\|x\|_{\text{Dir}}^2 + (\mu^\top x)^2 = \|z\|_{\text{Dir}}^2 + (\mu^\top x)^2 \geq \eta'_2 (1 - c^2) + c^2 = \eta_2 + c^2(1 - \eta_2).$$

where the second equation uses the fact that $\|z\|_{\text{Dir}}^2 \geq \eta'_2 \|z\|_\mu^2$. We conclude that $\eta'_1 \geq \eta'_2$ as long as $\eta'_2 \leq 1$.

Now we return to the original problem. To distinguish, we denote η_1 to be what is defined in our paper. Namely,

$$\eta_1 = \min_{\|x\|=1} \{ \|x\|_{\text{Dir}}^2 + (\mu^\top x)^2 \}.$$

According to (Zhang et al., 2021b), we define η_2 to be

$$\eta_2 = \min_{\|x\|=1, x^\top e=0} \|x\|_{\text{Dir}}^2.$$

We first notice the simple fact

$$\mu_{\max}^{-1} \min_{\|x\|=1} x^\top Ax = \frac{x^\top Ax}{\mu_{\max} \|x\|_2^2} \leq \min_{x: \|x\|_\mu=1} x^\top Ax \leq \frac{x^\top Ax}{\mu_{\min} \|x\|_2^2} = \mu_{\min}^{-1} \min_{\|x\|=1} x^\top Ax.$$

We introduce the following lemma:

Lemma B.2. *Let $A \in \mathbb{R}^{n \times n}$ be a symmetric positive semi-definite matrix satisfying $Ae = 0$. Let $\mu \in \mathbb{R}^n$ be a vector with strictly positive entries satisfying*

$$0 < \mu_{\min} \leq \mu_i \leq \mu_{\max} \quad \text{for } i = 1, \dots, n, \quad \text{and} \quad \sum_{i=1}^n \mu_i = 1.$$

Define

$$\lambda := \min_{\|x\|_2=1, x^\top e=0} x^\top Ax, \quad \lambda^{(\mu)} := \min_{\|x\|_2=1, x^\top \mu=0} x^\top Ax$$

and

$$\lambda' := \min_{\|x\|_2=1} x^\top Ax, \quad \lambda^{(\mu)'} := \min_{\|x\|_\mu=1} x^\top Ax.$$

Then we conclude

$$\frac{\lambda}{n \|\mu\|_2^2} \leq \lambda^{(\mu)}, \quad \mu_{\max}^{-1} \lambda' \leq \lambda^{(\mu)'} \leq \mu_{\min}^{-1} \lambda'.$$

With the above lemma, one can show

$$\eta'_2 \geq \frac{1}{n \mu_{\max} \|\mu\|^2} \eta_2$$

and

$$\eta'_1 \leq \mu_{\min}^{-1} \eta_1.$$

Therefore,

$$\eta_1 \geq \frac{\mu_{\min}}{n \mu_{\max} \|\mu\|^2} \eta_2.$$

Now, we only need to prove the above lemma. We split the proof into two parts.

First, we show that $\mu_{\max}^{-1} \lambda' \leq \lambda^{(\mu)'} \leq \mu_{\min}^{-1} \lambda'$. We can always rewrite λ' and $\lambda^{(\mu)'}$ as

$$\lambda' = \min_{x^\top \mu=0} \frac{x^\top Ax}{\|x\|^2}, \quad \lambda^{(\mu)'} = \min_{x^\top \mu=0} \frac{x^\top Ax}{\|x\|_\mu^2}.$$

The result followed by applying the fact $\mu_{\min} \|x\|^2 \leq \|x\|_\mu^2 \leq \mu_{\max} \|x\|^2$.

Next, we show $\frac{\lambda}{n \|\mu\|_2^2} \leq \lambda^{(\mu)}$. Let $x \in \mathbb{R}^n$ be a vector satisfying

$$x^\top \mu = 0, \quad \|x\| = 1.$$

Define

$$y = x - de, \quad \text{where } d = \frac{x^\top e}{n},$$

we can easily check $y^\top e = 0$ and $y^\top Ay = x^\top Ax$. Consider the norm of y ,

$$\|y\|_2^2 = \|x - de\|_2^2 = \|x\|_2^2 - 2d x^\top e + d^2 \|e\|_2^2 = 1 - \frac{(x^\top e)^2}{n}.$$

To bound $|x^\top e|$, decompose e by

$$e = \frac{1}{\|\mu\|^2} \mu + w, \quad \text{with } w^\top \mu = 0.$$

Then $x^\top e = x^\top w$. By the Cauchy–Schwarz inequality, $|x^\top e| \leq \|w\|_2$. Notice that

$$\|w\|_2^2 = \|e\|_2^2 - \left\| \frac{\mu}{\|\mu\|_2^2} \right\|^2 = n - \frac{1}{\|\mu\|_2^2},$$

Since $e^\top \mu = 1$,

$$\|y\|_2^2 \geq 1 - \frac{n - \frac{1}{\|\mu\|_2^2}}{n} = \frac{1}{n\|\mu\|_2^2}.$$

Now, letting $\hat{y} = y/\|y\|_2$, we have $\hat{y}^\top e = 0$ and thus

$$\hat{y}^\top A \hat{y} \geq \lambda.$$

It follows that

$$x^\top Ax = y^\top Ay = \|y\|_2^2 \hat{y}^\top A \hat{y} \geq \frac{\lambda_2}{n\|\mu\|_2^2}.$$

Take the minimum over all unit vectors x with $x^\top \mu = 0$ yields

$$\lambda^{(\mu)} \geq \frac{\lambda_2}{n\|\mu\|_2^2}.$$

□

B.2.2. DIFFERENCE BETWEEN η_1 AND η_3

Lemma B.3. *We have*

$$\eta_1 \geq \frac{1}{2} \eta_3.$$

Proof. Recall

$$\eta_1 = \min_{\|x\|=1} \left\{ \|\Phi x\|_{\text{Dir}}^2 + (\mu^\top \Phi x)^2 \right\}, \quad \eta_3 = \left(\min_{\|x\|=1} x^\top \Phi^\top D \Phi x \right) \cdot \left(\min_{\langle y, e \rangle_D = 0, \|y\|_D = 1} y^\top D(I - P)y \right),$$

where $D = \text{diag}(\mu)$, $\langle a, b \rangle_D = a^\top D b$, and e is the all-ones vector.

Let

$$\lambda := \min_{\langle y, e \rangle_D = 0, \|y\|_D = 1} y^\top D(I - P)y, \quad \sigma := \min_{\|x\|=1} x^\top \Phi^\top D \Phi x.$$

Then $\eta_3 = \sigma \lambda$.

First, for any $v \in \mathbb{R}^{|S|}$,

$$\begin{aligned} \|v\|_{\text{Dir}}^2 &= \frac{1}{2} \sum_{s, s'} \mu(s) P(s'|s) (v(s) - v(s'))^2 \\ &= \sum_s \mu(s) v(s)^2 - \sum_{s, s'} \mu(s) P(s'|s) v(s) v(s') \\ &= v^\top D(I - P)v, \end{aligned}$$

where we used $\sum_{s'} P(s'|s) = 1$ and stationarity $\sum_s \mu(s) P(s'|s) = \mu(s')$.

We also remark on the standard observation that multiplication by P is a contraction in the D -norm:

$$\|Pv\|_D^2 = \sum_s \mu(s) \left(\sum_{s'} P(s'|s) v(s') \right)^2 \leq \sum_s \mu(s) \sum_{s'} P(s'|s) v(s')^2 = \sum_{s'} \mu(s') v(s')^2 = \|v\|_D^2,$$

where the inequality is Jensen and the equality again uses stationarity. Hence $\|Pv\|_D \leq \|v\|_D$. Therefore for any y with $\|y\|_D = 1$,

$$y^\top D(I - P)y = \|y\|_D^2 - \langle y, Py \rangle_D \leq 1 + |\langle y, Py \rangle_D| \leq 1 + \|y\|_D \|Py\|_D \leq 2.$$

Taking the minimum over the constraint set gives $\lambda \leq 2$.

Having established that, we next fix arbitrary $v \in \mathbb{R}^{|\mathcal{S}|}$. Let

$$c := \langle v, e \rangle_D = v^\top De = \mu^\top v, \quad u := v - ce,$$

so that $\langle u, e \rangle_D = 0$. Since $\|e\|_D^2 = e^\top De = \sum_s \mu(s) = 1$ and $u \perp e$ in $\langle \cdot, \cdot \rangle_D$,

$$\|v\|_D^2 = \|u\|_D^2 + c^2.$$

Also, adding a constant does not change the Dirichlet seminorm, so $\|v\|_{\text{Dir}} = \|u\|_{\text{Dir}}$. Using Step 1 and the definition of λ ,

$$\|v\|_{\text{Dir}}^2 = \|u\|_{\text{Dir}}^2 = u^\top D(I - P)u \geq \lambda \|u\|_D^2.$$

Therefore,

$$\|v\|_{\text{Dir}}^2 + (\mu^\top v)^2 = \|u\|_{\text{Dir}}^2 + c^2 \geq \lambda \|u\|_D^2 + c^2 \geq \frac{\lambda}{2} (\|u\|_D^2 + c^2) = \frac{\lambda}{2} \|v\|_D^2,$$

where we used $\lambda \leq 2$ so that $(\lambda/2)c^2 \leq c^2$.

Step 3: Applying Step 2 to $v = \Phi x$ gives, for every $x \in \mathbb{R}^d$,

$$\|\Phi x\|_{\text{Dir}}^2 + (\mu^\top \Phi x)^2 \geq \frac{\lambda}{2} \|\Phi x\|_D^2 = \frac{\lambda}{2} x^\top \Phi^\top D \Phi x.$$

Taking $\min_{\|x\|=1}$ on both sides yields

$$\eta_1 = \min_{\|x\|=1} \left\{ \|\Phi x\|_{\text{Dir}}^2 + (\mu^\top \Phi x)^2 \right\} \geq \frac{\lambda}{2} \min_{\|x\|=1} x^\top \Phi^\top D \Phi x = \frac{\lambda}{2} \sigma = \frac{1}{2} \eta_3.$$

□

B.2.3. RELATION BETWEEN PROJECTION AND CONDITION NUMBER

In (Kim et al., 2025), their sample complexity depends on the projection radius R_{proj} . Although they did not have a discussion on the projection radius, we will assume $R_{\text{proj}} = R_\theta$ and compare their sample complexity for completeness.

Lemma B.4. Suppose the projection radius is chosen to be

$$R_{\text{proj}} = R_\theta = \frac{2r_{\max}C}{(1 - \beta)\sqrt{(1 - \omega^2)\lambda_{\min}(\Phi^\top D \Phi)}}.$$

We have

$$R_\theta^2 \geq \frac{2r_{\max}^2 C^2}{(1 - \beta)^2 \eta_3}.$$

Proof of Lemma B.4. For any vector y such that $\|y\|_D = 1$ and $\langle y, e \rangle_D = 0$,

$$y^\top D(I - P)y = 1 - \langle y, Py \rangle_D.$$

By Cauchy–Schwarz,

$$\langle y, Py \rangle_D \leq \|y\|_D \|Py\|_D = \|Py\|_D \leq \omega.$$

Therefore, we have

$$\min_{\langle y, e \rangle_D = 0, \|y\|_D = 1} y^\top D(I - P)y \geq 1 - \omega.$$

Since $\omega < 1$ which is already established in Lemma A.1, $1 - \omega \geq (1 - \omega^2)/2$. Therefore,

$$\frac{1}{1 - \omega^2} \geq \frac{1}{2 \min_{\langle y, e \rangle_D = 0, \|y\|_D = 1} y^\top D(I - P)y}$$

Now we obtain

$$R_\theta^2 \geq \frac{(2r_{\max}C)^2}{2(1 - \beta)^2 \min_{\langle y, e \rangle_D = 0, \|y\|_D = 1} y^\top D(I - P)y \cdot \lambda_{\min}(\Phi^\top D\Phi)} = \frac{2r_{\max}^2 C^2}{(1 - \beta)^2 \eta_3}.$$

□

C. Analysis of Double Chain Algorithm with Constant Stepsize

In this section, we will give a detailed proof of Theorem 4.1 and Theorem 4.2. Notice that these proofs can also be applied to the tabular case if one sets $\Phi = I$.

Throughout the analysis, we denote $\mathcal{F}_t$ as the history up to iteration $t - 1$, i.e.,

$$\mathcal{F}_t := \sigma(\{s_0, \dots, s_{t-1}\} \cup \{\hat{s}_0, \dots, \hat{s}_{t-1}\}). \quad (\text{filtration})$$

In particular, θ_t and δ_t is $\mathcal{F}_t$ -measurable. We also denote $B = 2\|\delta_0\| + (r_{\max} + \|\theta^*\|)$.

Before going to the proof, we first record some useful properties about $f(s, \hat{s}, \theta)$ and $g(s, s', \theta)$.

Define the mean field

$$\bar{h}(\theta) := \mathbb{E}_\mu[f(s, \hat{s}, \theta) + g(s, s', \theta)],$$

where under $\mathbb{E}_\mu$ we sample $s \sim \mu$, $\hat{s} \sim \mu$ independently, and then sample $s' \sim P(\cdot | s)$. By Eq. (14), it is straightforward to check that

$$\begin{aligned} \bar{h}(\theta) &= -\Phi^\top \mu \mu^\top (R + \Phi \theta) + \Phi^\top D(R + P\Phi \theta - \Phi \theta) \\ &= -\Phi^\top \mu \mu^\top \Phi \delta + \Phi^\top D(P - I)\Phi \delta, \end{aligned} \quad (21)$$

where $\delta := \theta - \theta^*$.

Beyond the above fact, we have two additional lemmas.

Lemma C.1. *For any $t \geq 0$, we have*

$$\begin{aligned} \|f(s_t, \hat{s}_t, \theta_t)\| &\leq \|\delta_t\| + r_{\max} + \|\theta^*\|, \\ \|g(s_t, s'_t, \theta_t)\| &\leq 2(\|\delta_t\| + r_{\max} + \|\theta^*\|). \end{aligned}$$

Proof of Lemma C.1. According to Eq.(16),

$$\begin{aligned} \|f(s_t, \hat{s}_t, \theta_t)\| &\leq (|r_{s_t}| + |\phi(s_t)^\top \theta_t|) \|\phi(\hat{s}_t)\| \\ &\leq (r_{\max} + \|\theta_t\| \|\phi(s_t)\|) \|\phi(\hat{s}_t)\| && (\text{Cauchy-Schwarz}) \\ &\leq r_{\max} + \|\theta_t\| && (\|\phi(\cdot)\| \leq 1) \\ &\leq r_{\max} + \|\theta_t - \theta^*\| + \|\theta^*\| \\ &= \|\delta_t\| + r_{\max} + \|\theta^*\|. \end{aligned}$$

Similarly,

$$\begin{aligned} \|g(s_t, s'_t, \theta_t)\| &\leq (|r_{s_t}| + |\phi(s'_t)^\top \theta_t| + |\phi(s_t)^\top \theta_t|) \|\phi(s_t)\| \\ &\leq r_{\max} + 2\|\theta_t\| && (\|\phi(\cdot)\| \leq 1) \\ &\leq r_{\max} + 2\|\theta_t - \theta^*\| + 2\|\theta^*\| \\ &\leq 2(\|\delta_t\| + r_{\max} + \|\theta^*\|). \end{aligned}$$

□

Lemma C.2. *The function f is 1-Lipschitz and g is 2-Lipschitz with respect to θ , i.e.,*

$$\|g(s, s', \theta_1) - g(s, s', \theta_2)\| \leq 2\|\theta_1 - \theta_2\|, \quad \|f(s, \hat{s}, \theta_1) - f(s, \hat{s}, \theta_2)\| \leq \|\theta_1 - \theta_2\|.$$

Proof of Lemma C.2. Using Eq.(16) and $\|\phi(\cdot)\| \leq 1$,

$$g(s, s', \theta_1) - g(s, s', \theta_2) = ((\phi(s') - \phi(s))^\top (\theta_1 - \theta_2)) \phi(s),$$

hence

$$\|g(s, s', \theta_1) - g(s, s', \theta_2)\| \leq \|\phi(s') - \phi(s)\| \|\phi(s)\| \|\theta_1 - \theta_2\| \leq 2\|\theta_1 - \theta_2\|.$$

Likewise,

$$f(s, \hat{s}, \theta_1) - f(s, \hat{s}, \theta_2) = -(\phi(s)^\top (\theta_1 - \theta_2)) \phi(\hat{s}),$$

so $\|f(s, \hat{s}, \theta_1) - f(s, \hat{s}, \theta_2)\| \leq \|\theta_1 - \theta_2\|$. □

Now we are ready to provide our proof of Theorem 4.1.

C.1. Proof of Theorem 4.1

Proof of Theorem 4.1. Let $u_t := f(s_t, \hat{s}_t, \theta_t) + g(s_t, s'_t, \theta_t)$. From Eq.(15) with $\alpha_t = \alpha$,

$$\|\delta_{t+1}\|^2 = \|\delta_t\|^2 + 2\alpha \delta_t^\top u_t + \alpha^2 \|u_t\|^2.$$

Taking conditional expectation given $\mathcal{F}_t$, and using that δ_t is $\mathcal{F}_t$ -measurable,

$$\mathbb{E}[\|\delta_{t+1}\|^2 \mid \mathcal{F}_t] = \|\delta_t\|^2 + 2\alpha \delta_t^\top \mathbb{E}[u_t \mid \mathcal{F}_t] + \alpha^2 \mathbb{E}[\|u_t\|^2 \mid \mathcal{F}_t].$$

Step 1: Drift term. We first introduce the following lemma:

Lemma C.3. *The linear function $h(\theta) := \Phi^\top D(I - P)\Phi\theta$ is a gradient splitting of the quadratic function $\|\Phi\theta\|_{\text{Dir}}^2$, i.e.,*

$$\langle \theta, h(\theta) \rangle = \frac{1}{2} \langle \theta, \nabla_\theta \|\Phi\theta\|_{\text{Dir}}^2 \rangle, \quad \forall \theta \in \mathbb{R}^d.$$

As a result, we have $\|\Phi\theta\|_{\text{Dir}}^2 = \theta^\top h(\theta)$ for all $\theta \in \mathbb{R}^d$.

The proof of this lemma can be found in Section C.2.

Under i.i.d. sampling, $(s_t, \hat{s}_t)$ are independent of $\mathcal{F}_t$ with marginals μ , and $s'_t \sim P(\cdot \mid s_t)$. Therefore,

$$\mathbb{E}[u_t \mid \mathcal{F}_t] = \bar{h}(\theta_t),$$

where $\bar{h}(\cdot)$ is defined in Eq.(21).

Using Eq.(21) with $\delta = \delta_t$,

$$\begin{aligned} \delta_t^\top \bar{h}(\theta_t) &= \delta_t^\top \Phi^\top D(P - I)\Phi\delta_t - \delta_t^\top \Phi^\top \mu \mu^\top \Phi\delta_t \\ &= -\delta_t^\top \Phi^\top D(I - P)\Phi\delta_t - \|\mu^\top \Phi\delta_t\|^2 \\ &= -\|\Phi\delta_t\|_{\text{Dir}}^2 - \|\mu^\top \Phi\delta_t\|^2 \\ &\leq -\eta \|\delta_t\|^2. \end{aligned} \tag{Lemma C.3} \tag{Eq. (8)}$$

Step 2: Second-moment term. By Lemma C.1,

$$\|u_t\| \leq \|f(s_t, \hat{s}_t, \theta_t)\| + \|g(s_t, s'_t, \theta_t)\| \leq 3(\|\delta_t\| + r_{\max} + \|\theta^*\|). \tag{Lemma C.1}$$

Hence,

$$\begin{aligned} \mathbb{E}[\|u_t\|^2 \mid \mathcal{F}_t] &\leq 9(\|\delta_t\| + r_{\max} + \|\theta^*\|)^2 \\ &\leq 18\|\delta_t\|^2 + 18(r_{\max} + \|\theta^*\|)^2. \end{aligned} \tag{((a+b)^2 \leq 2a^2 + 2b^2)}$$

Step 3: Combine. Plugging the two bounds back,

$$\mathbb{E}[\|\delta_{t+1}\|^2 \mid \mathcal{F}_t] \leq (1 - 2\alpha\eta + 18\alpha^2)\|\delta_t\|^2 + 18\alpha^2(r_{\max} + \|\theta^*\|)^2.$$

Since $\alpha \leq \eta/18$, we have $18\alpha^2 \leq \alpha\eta$, hence

$$1 - 2\alpha\eta + 18\alpha^2 \leq 1 - \alpha\eta,$$

and therefore

$$\mathbb{E}[\|\delta_{t+1}\|^2 \mid \mathcal{F}_t] \leq (1 - \alpha\eta)\|\delta_t\|^2 + 18\alpha^2(r_{\max} + \|\theta^*\|)^2.$$

Taking expectation and iterating the recursion yields

$$\mathbb{E}\|\delta_T\|^2 \leq (1 - \alpha\eta)^T \mathbb{E}\|\delta_0\|^2 + 18\alpha^2(r_{\max} + \|\theta^*\|)^2 \sum_{k=0}^{T-1} (1 - \alpha\eta)^k.$$

Using $\sum_{k=0}^{T-1} (1 - \alpha\eta)^k \leq \frac{1}{\alpha\eta}$ and $1 - x \leq e^{-x}$,

$$\mathbb{E}\|\delta_T\|^2 \leq e^{-\alpha\eta T} \mathbb{E}\|\delta_0\|^2 + \frac{18\alpha(r_{\max} + \|\theta^*\|)^2}{\eta}.$$

□

C.2. Proof of Lemma C.3

Proof of Lemma C.3. According to Eq. (7) and the stationarity $\mu^\top = \mu^\top P$,

$$\begin{aligned}\|f\|_{\text{Dir}}^2 &= \frac{1}{2} \sum_{s,s'} \mu(s) P(s'|s) (f(s)^2 + f(s')^2 - 2f(s)f(s')) \\ &= \sum_s \mu(s) f(s)^2 - \sum_{s,s'} \mu(s) P(s'|s) f(s) f(s') \\ &= f^\top D(I - P)f.\end{aligned}$$

Taking $f = \Phi\theta$ gives

$$\|\Phi\theta\|_{\text{Dir}}^2 = \theta^\top \Phi^\top D(I - P)\Phi\theta = \theta^\top h(\theta).$$

Since $\|\Phi\theta\|_{\text{Dir}}^2$ is a quadratic form, its gradient is

$$\nabla_\theta \|\Phi\theta\|_{\text{Dir}}^2 = (\Phi^\top D(I - P)\Phi + \Phi^\top (I - P)^\top D\Phi)\theta,$$

hence

$$\frac{1}{2} \langle \theta, \nabla_\theta \|\Phi\theta\|_{\text{Dir}}^2 \rangle = \theta^\top \Phi^\top D(I - P)\Phi\theta = \langle \theta, h(\theta) \rangle.$$

□

C.3. Proof of Theorem 4.2

Proof of Theorem 4.2. For convenience, define

$$\begin{aligned}\bar{g}(\theta_t) &:= \sum_{s,s'} \mu(s) P(s'|s) g(s, s', \theta_t) = \Phi^\top D(R + P\Phi\theta_t - \Phi\theta_t), \\ \bar{f}(\theta_t) &:= \sum_{s,\hat{s}} \mu(s) \mu(\hat{s}) f(s, \hat{s}, \theta_t) = -\Phi^\top \mu \mu^\top (R + \Phi\theta_t).\end{aligned}\tag{22}$$

Let $u_t := f(s_t, \hat{s}_t, \theta_t) + g(s_t, s'_t, \theta_t)$. Using that δ_t and $\bar{f}(\theta_t) + \bar{g}(\theta_t)$ are $\mathcal{F}_t$ -measurable, we have

$$\begin{aligned}&\mathbb{E}[\delta_t^\top (\Phi^\top D(P - I)\Phi\delta_t - \Phi^\top \mu \mu^\top \Phi\delta_t - \mathbb{E}[u_t | \mathcal{F}_t])] \\ &= \mathbb{E}[\delta_t^\top (\bar{g}(\theta_t) + \bar{f}(\theta_t) - \mathbb{E}[u_t | \mathcal{F}_t])] \\ &= \mathbb{E}[\delta_t^\top (\bar{g}(\theta_t) + \bar{f}(\theta_t) - u_t)] \\ &= -\mathbb{E}[\delta_t^\top ((g(s_t, s'_t, \theta_t) - \bar{g}(\theta_t)) + (f(s_t, \hat{s}_t, \theta_t) - \bar{f}(\theta_t)))] .\end{aligned}\tag{Eq. (21)}$$

Define B and G as in the theorem statement. We introduce the following lemma:

Lemma C.4. *Suppose*

$$\alpha \leq \frac{\eta B^2}{(3\tau_{\text{mix}} + 1)G},$$

then $\mathbb{E}[\|\delta_t\|^2] \leq B^2$ for all $t \geq 0$.

The proof of this lemma can be found in Section C.4.

Since $\mathbb{E}[\|\delta_t\|^2] \leq B^2$ for all $t \geq 0$, we know that the following lemma holds.

Lemma C.5. *Suppose $t \geq \tau_{\text{mix}}$. With $\bar{f}, \bar{g}$ defined in Eq.(22), we have*

$$\begin{aligned}\mathbb{E}[\delta_t^\top (g(s_t, s'_t, \theta_t) - \bar{g}(\theta_t))] &\leq \alpha (3B^2 + (r_{\text{max}} + \|\theta^*\|)^2) + (42B^2 + 30(r_{\text{max}} + \|\theta^*\|)^2) \tau_{\text{mix}} \alpha, \\ \mathbb{E}[\delta_t^\top (f(s_t, \hat{s}_t, \theta_t) - \bar{f}(\theta_t))] &\leq \alpha (3B^2 + (r_{\text{max}} + \|\theta^*\|)^2) + (21B^2 + 15(r_{\text{max}} + \|\theta^*\|)^2) \tau_{\text{mix}} \alpha.\end{aligned}$$

The proof of this lemma can be found in Section C.5.

Summing the two inequalities and multiplying by the factor 2α , we obtain for all $t \geq \tau_{\text{mix}}$,

$$\begin{aligned} & -2\alpha \mathbb{E}[\delta_t^\top (\Phi^\top D(P - I)\Phi \delta_t - \Phi^\top \mu \mu^\top \Phi \delta_t - \mathbb{E}[u_t | \mathcal{F}_t])] \\ & \leq 4\alpha^2 (3B^2 + (r_{\max} + \|\theta^*\|)^2) + \alpha^2 \tau_{\text{mix}} (126B^2 + 90(r_{\max} + \|\theta^*\|)^2). \end{aligned}$$

Using the same I_1 and I_2 bounds as in the i.i.d. case (so that the drift contributes $-(2\alpha\eta)\|\delta_t\|^2$ and the squared-norm term contributes $18\alpha^2(\|\delta_t\|^2 + (r_{\max} + \|\theta^*\|)^2)$), we obtain for any $t \geq \tau_{\text{mix}}$,

$$\begin{aligned} \mathbb{E}\|\delta_{t+1}\|^2 & \leq (1 - 2\alpha\eta)\mathbb{E}\|\delta_t\|^2 + 18\alpha^2 \mathbb{E}\|\delta_t\|^2 + 18\alpha^2 (r_{\max} + \|\theta^*\|)^2 \\ & \quad + 4\alpha^2 (3B^2 + (r_{\max} + \|\theta^*\|)^2) + \alpha^2 \tau_{\text{mix}} (126B^2 + 90(r_{\max} + \|\theta^*\|)^2) \\ & \leq (1 - 2\alpha\eta)\mathbb{E}\|\delta_t\|^2 + 18\alpha^2 B^2 + 22\alpha^2 (r_{\max} + \|\theta^*\|)^2 + 12\alpha^2 B^2 \\ & \quad + \alpha^2 \tau_{\text{mix}} (126B^2 + 90(r_{\max} + \|\theta^*\|)^2) \quad (\text{Lemma C.4}) \\ & \leq (1 - 2\alpha\eta)\mathbb{E}\|\delta_t\|^2 + \alpha^2 (42B^2 + 30(r_{\max} + \|\theta^*\|)^2) (3\tau_{\text{mix}} + 1) \\ & = (1 - 2\alpha\eta)\mathbb{E}\|\delta_t\|^2 + \alpha^2 G(3\tau_{\text{mix}} + 1). \end{aligned}$$

Iterating the above inequality from $t = \tau_{\text{mix}}$ to $T - 1$ gives

$$\mathbb{E}\|\delta_T\|^2 \leq (1 - 2\alpha\eta)^{T - \tau_{\text{mix}}} \mathbb{E}\|\delta_{\tau_{\text{mix}}}\|^2 + \alpha^2 G(3\tau_{\text{mix}} + 1) \sum_{k=0}^{T - \tau_{\text{mix}} - 1} (1 - 2\alpha\eta)^k.$$

Using Lemma C.4, $\sum_{k \geq 0} (1 - 2\alpha\eta)^k \leq \frac{1}{2\alpha\eta}$, and $1 - x \leq e^{-x}$, we obtain

$$\mathbb{E}\|\delta_T\|^2 \leq e^{-2\alpha\eta(T - \tau_{\text{mix}})} B^2 + \frac{\alpha G(3\tau_{\text{mix}} + 1)}{2\eta}.$$

□

C.4. Proof of Lemma C.4

Proof of Lemma C.4. We first handle $t \leq \tau_{\text{mix}}$ via a pathwise bound. From Eq.(15) and Lemma C.1,

$$\begin{aligned} \|\delta_{t+1}\| & \leq \|\delta_t\| + \alpha(\|f(s_t, \hat{s}_t, \theta_t)\| + \|g(s_t, s'_t, \theta_t)\|) \\ & \leq \|\delta_t\| + 3\alpha(\|\delta_t\| + r_{\max} + \|\theta^*\|) = (1 + 3\alpha)\|\delta_t\| + 3\alpha(r_{\max} + \|\theta^*\|). \end{aligned}$$

Iterating the inequality yields, for $t \leq \tau_{\text{mix}}$,

$$\|\delta_t\| \leq (1 + 3\alpha)^t \|\delta_0\| + 3\alpha(r_{\max} + \|\theta^*\|) \sum_{k=0}^{t-1} (1 + 3\alpha)^k.$$

Using $(1 + 3\alpha)^t \leq e^{3\alpha t}$ and $\sum_{k=0}^{t-1} (1 + 3\alpha)^k \leq t(1 + 3\alpha)^t$, we have

$$\|\delta_t\| \leq e^{3\alpha\tau_{\text{mix}}} \|\delta_0\| + 3\alpha\tau_{\text{mix}} e^{3\alpha\tau_{\text{mix}}} (r_{\max} + \|\theta^*\|).$$

The stepsize condition implies $\alpha\tau_{\text{mix}} \leq 1/6$ (indeed $G \geq 42B^2$ and $\eta \leq 3$ imply $\alpha \leq \eta/(42(3\tau_{\text{mix}} + 1)) \leq 1/(6\tau_{\text{mix}})$), so $e^{3\alpha\tau_{\text{mix}}} \leq e^{1/2} < 2$ and $6\alpha\tau_{\text{mix}} \leq 1$. Therefore,

$$\|\delta_t\| \leq 2\|\delta_0\| + 6\alpha\tau_{\text{mix}}(r_{\max} + \|\theta^*\|) \leq 2\|\delta_0\| + (r_{\max} + \|\theta^*\|) = B, \quad \forall t \leq \tau_{\text{mix}}.$$

Hence $\mathbb{E}\|\delta_t\|^2 \leq B^2$ for all $t \leq \tau_{\text{mix}}$.

Now consider $t > \tau_{\text{mix}}$ and use induction on t . Assume $\mathbb{E}\|\delta_k\|^2 \leq B^2$ for all $k \leq t$. Then we have the following lemma:

Lemma C.6. Suppose $t \geq \tau_{\text{mix}}$ and $\mathbb{E}[\|\delta_k\|^2] \leq B^2$ for all $k \leq t$. Then Lemma C.5 holds at time t .

The proof of this lemma can be found in Section C.6.

Plugging those bounds into the recursion in the proof of Theorem 4.2 gives

$$\mathbb{E}\|\delta_{t+1}\|^2 \leq (1 - 2\alpha\eta)B^2 + \alpha^2 G(3\tau_{\text{mix}} + 1).$$

Using the stepsize condition $\alpha \leq \frac{\eta B^2}{(3\tau_{\text{mix}}+1)G}$, we have $\alpha^2 G(3\tau_{\text{mix}} + 1) \leq \alpha\eta B^2$, hence

$$\mathbb{E}\|\delta_{t+1}\|^2 \leq (1 - 2\alpha\eta)B^2 + \alpha\eta B^2 \leq B^2.$$

This completes the induction and proves the claim for all $t \geq 0$. $\square$

C.5. Proof of Lemma C.5

Proof of Lemma C.5. Fix any $t \geq \tau_{\text{mix}}$. Lemma C.4 gives $\mathbb{E}\|\delta_k\|^2 \leq B^2$ for all $k \leq t$, so the assumptions of Lemma C.6 are satisfied. Lemma C.6 therefore implies the two inequalities stated in Lemma C.5. $\square$

C.6. Proof of Lemma C.6

Proof of Lemma C.6. For simplicity, denote

$$g_t(\theta) := g(s_t, s'_t, \theta), \quad f_t(\theta) := f(s_t, \hat{s}_t, \theta).$$

Step 1: the g -term. Decompose

$$\begin{aligned} & \mathbb{E}[(\theta_t - \theta^*)^\top (g_t(\theta_t) - \bar{g}(\theta_t))] \\ &= \underbrace{\mathbb{E}[(\theta_t - \theta_{t-\tau_{\text{mix}}})^\top (g_t(\theta_t) - \bar{g}(\theta_t))]}_{I_1} + \underbrace{\mathbb{E}[(\theta_{t-\tau_{\text{mix}}} - \theta^*)^\top (g_t(\theta_{t-\tau_{\text{mix}}}) - \bar{g}(\theta_{t-\tau_{\text{mix}}}))]}_{I_2} \\ & \quad + \underbrace{\mathbb{E}[(\theta_{t-\tau_{\text{mix}}} - \theta^*)^\top (g_t(\theta_t) - g_t(\theta_{t-\tau_{\text{mix}}}))]}_{I_3} + \underbrace{\mathbb{E}[(\theta_{t-\tau_{\text{mix}}} - \theta^*)^\top (\bar{g}(\theta_{t-\tau_{\text{mix}}}) - \bar{g}(\theta_t))]}_{I_4}. \end{aligned}$$

Term I_1 . We have

$$\theta_t - \theta_{t-\tau_{\text{mix}}} = \sum_{i=t-\tau_{\text{mix}}+1}^t \alpha(g_i(\theta_i) + f_i(\theta_i)),$$

so by Lemma C.1,

$$\|\theta_t - \theta_{t-\tau_{\text{mix}}}\| \leq 3 \sum_{i=t-\tau_{\text{mix}}+1}^t \alpha(\|\delta_i\| + r_{\text{max}} + \|\theta^*\|).$$

Also $\|g_t(\theta_t) - \bar{g}(\theta_t)\| \leq \|g_t(\theta_t)\| + \|\bar{g}(\theta_t)\| \leq 4(\|\delta_t\| + r_{\text{max}} + \|\theta^*\|)$. Therefore, since $2ab \leq a^2 + b^2$, we have

$$\begin{aligned} I_1 &\leq \mathbb{E}[\|\theta_t - \theta_{t-\tau_{\text{mix}}}\| \|g_t(\theta_t) - \bar{g}(\theta_t)\|] \\ &\leq \mathbb{E}\left[12(\|\delta_t\| + c) \sum_{i=t-\tau_{\text{mix}}+1}^t \alpha(\|\delta_i\| + c)\right] \\ &\leq 12 \sum_{i=t-\tau_{\text{mix}}+1}^t \alpha \mathbb{E}[\|\delta_t\|^2 + \|\delta_i\|^2 + 2c^2] \\ &\leq 24(B^2 + c^2) \sum_{i=t-\tau_{\text{mix}}+1}^t \alpha. \end{aligned}$$

Term I_2 . Condition on $\mathcal{F}_{t-\tau_{\text{mix}}}$. Given $\mathcal{F}_{t-\tau_{\text{mix}}}$, the law of (s_t, s'_t) differs from $\mu(s)P(s'|s)$ by at most $\|P_{\tau_{\text{mix}}}(\cdot|s_{t-\tau_{\text{mix}}}) - \mu\|_1$, and by definition of $\tau_{\text{mix}} = \tau_{\text{mix}}(\alpha)$ we have $\|P_{\tau_{\text{mix}}}(\cdot|s) - \mu\|_1 \leq \alpha$ for all s . Using Lemma C.1, $\sup_{s,s'} \|g(s, s', \theta_{t-\tau_{\text{mix}}})\| \leq 2(\|\delta_{t-\tau_{\text{mix}}}\| + c)$, hence

$$\|\mathbb{E}[g_t(\theta_{t-\tau_{\text{mix}}}) - \bar{g}(\theta_{t-\tau_{\text{mix}}}) \mid \mathcal{F}_{t-\tau_{\text{mix}}}] \| \leq 2\alpha(\|\delta_{t-\tau_{\text{mix}}}\| + c).$$

Therefore,

$$\begin{aligned} I_2 &= \mathbb{E}[\mathbb{E}[(\theta_{t-\tau_{\text{mix}}} - \theta^*)^\top (g_t(\theta_{t-\tau_{\text{mix}}}) - \bar{g}(\theta_{t-\tau_{\text{mix}}})) \mid \mathcal{F}_{t-\tau_{\text{mix}}}]] \\ &\leq \mathbb{E}[\|\delta_{t-\tau_{\text{mix}}}\| \cdot 2\alpha(\|\delta_{t-\tau_{\text{mix}}}\| + c)] \\ &\leq \alpha \mathbb{E}[3\|\delta_{t-\tau_{\text{mix}}}\|^2 + c^2] \leq \alpha(3B^2 + c^2). \end{aligned}$$

Terms I_3 and I_4 . By Lemma C.2,

$$\|g_t(\theta_t) - g_t(\theta_{t-\tau_{\text{mix}}})\| \leq 2\|\theta_t - \theta_{t-\tau_{\text{mix}}}\| \leq 6 \sum_{i=t-\tau_{\text{mix}}+1}^t \alpha(\|\delta_i\| + c).$$

Hence

$$\begin{aligned} I_3 &\leq 6 \sum_{i=t-\tau_{\text{mix}}+1}^t \alpha \mathbb{E}[\|\delta_{t-\tau_{\text{mix}}}\|(\|\delta_i\| + c)] \\ &\leq 6 \sum_{i=t-\tau_{\text{mix}}+1}^t \alpha \mathbb{E}[\|\delta_{t-\tau_{\text{mix}}}\|(\|\delta_i\| + \|\delta_{t-\tau_{\text{mix}}}\|c)] \\ &\leq \sum_{i=t-\tau_{\text{mix}}+1}^t \alpha \mathbb{E}\left[6 \cdot \frac{\|\delta_{t-\tau_{\text{mix}}}\|^2 + \|\delta_i\|^2}{2} + 6 \cdot \frac{\|\delta_{t-\tau_{\text{mix}}}\|^2 + c^2}{2}\right] \\ &\leq (9B^2 + 3c^2) \sum_{i=t-\tau_{\text{mix}}+1}^t \alpha. \end{aligned}$$

The same bound applies to I_4 (since $\bar{g}(\cdot)$ is also 2-Lipschitz in θ by Lemma C.2 and linearity of expectation).

Combining I_1 - I_4 ,

$$\mathbb{E}[(\theta_t - \theta^*)^\top (g_t(\theta_t) - \bar{g}(\theta_t))] \leq \alpha(3B^2 + c^2) + (42B^2 + 30c^2) \sum_{i=t-\tau_{\text{mix}}+1}^t \alpha.$$

With constant stepsize, $\sum_{i=t-\tau_{\text{mix}}+1}^t \alpha = \tau_{\text{mix}}\alpha$, proving the first inequality in Lemma C.5.

Step 2: the f -term. Same as before, we can do the same decomposition so that

$$\mathbb{E}[(\theta_t - \theta^*)^\top (f_t(\theta_t) - \bar{f}(\theta_t))] = I'_1 + I'_2 + I'_3 + I'_4.$$

The only change is in I'_2 : given $\mathcal{F}_{t-\tau_{\text{mix}}}$, the joint law of $(s_t, \hat{s}_t)$ equals $P_{\tau_{\text{mix}}}(\cdot|s_{t-\tau_{\text{mix}}}) \otimes P_{\tau_{\text{mix}}}(\cdot|\hat{s}_{t-\tau_{\text{mix}}})$, whose ℓ_1 distance to $\mu \otimes \mu$ is at most $\|P_{\tau_{\text{mix}}}(\cdot|s_{t-\tau_{\text{mix}}}) - \mu\|_1 + \|P_{\tau_{\text{mix}}}(\cdot|\hat{s}_{t-\tau_{\text{mix}}}) - \mu\|_1 \leq 2\alpha$.

Using the same analysis as function g , we conclude

$$\mathbb{E}[(\theta_t - \theta^*)^\top (f_t(\theta_t) - \bar{f}(\theta_t))] \leq \alpha(3B^2 + c^2) + (21B^2 + 15c^2) \sum_{i=t-\tau_{\text{mix}}+1}^t \alpha,$$

which proves the second inequality in Lemma C.5. $\square$

D. Analysis of Double Chain Algorithm with Decaying Stepsize

In this section, we give a detailed proof of Theorem 4.3. We first state the full version of the theorem with all the constants:

Theorem D.1 (Restatement of Theorem 4.3). *Let $a > 0$, $c_0 > 0$, and $\alpha_t = \frac{a}{(t+c_0)^\xi}$ with $\xi \in (0, 1]$. Denote*

$$\begin{aligned}\beta_1 &:= 2\mathbb{E}\left[\left(\|\theta_0\| + r_{\max} + 2\|\theta^*\|\right)^2\right] + 2\mathbb{E}\left[\|\theta_0 - \theta^*\|^2\right], \\ \beta_2 &:= 946(r_{\max} + 3\|\theta^*\|)^2,\end{aligned}$$

and

$$\begin{aligned}L_1 &:= \max\left\{1, \frac{\log C + 1}{\log(1/\beta)}\right\} \\ \beta(T) &:= 4aL_1\beta_2(\log(T + c_0) - \log a + 1)\end{aligned}$$

Assume $c_0 \geq \max\{c_{0,1}(a, \xi), c_{0,2}(a, \xi)\}$, where $c_{0,1}, c_{0,2}$ are chosen so that Lemma D.2 holds.

1. If $\xi \in (0, 1)$ and $c_0 \geq \left(\frac{2\xi}{a\eta}\right)^{\frac{1}{1-\xi}}$, then

$$\begin{aligned}\mathbb{E}[\|\delta_T\|^2] &\leq \beta_1 \exp\left(-\frac{\eta a}{1-\xi}\left((T + c_0)^{1-\xi} - (\tau_{\text{mix}} + c_0)^{1-\xi}\right)\right) \\ &\quad + \frac{\beta(T)}{\eta(T + c_0)^\xi}.\end{aligned}$$

2. If $\xi = 1$ and $c_0 \geq \eta a$, then

$$\mathbb{E}[\|\delta_T\|^2] \leq \beta_1 \left(\frac{\tau_{\text{mix}} + c_0}{T + c_0}\right)^{\eta a} + \frac{\Gamma(T)}{(T + c_0)^q},$$

where $q = \min\{1, \eta a\}$, $\tilde{\beta}(T) := 2L_1\beta_2(\log(T + c_0) - \log a + 1)$ and

$$\Gamma(T) := \begin{cases} \frac{4a^2}{1-\eta a} \tilde{\beta}(T), & 0 < a < 1/\eta, \\ 4a^2 \log(T + c_0) \tilde{\beta}(T), & a = 1/\eta, \\ \frac{4ea^2}{\eta a - 1} \tilde{\beta}(T), & a > 1/\eta. \end{cases}$$

Throughout, we use the same notations as in Appendix C.

D.1. Proof of Theorem 4.3

Proof of Theorem 4.3. To control Markov noise, we need an upper bound on $\sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i$. This is captured by the following lemma.

Lemma D.2. *There exist constants $c_{0,1} = c_{0,1}(a, \xi) > 0$ and $c_{0,2} = c_{0,2}(a, \xi) > 0$ such that if $c_0 \geq \max\{c_{0,1}, c_{0,2}\}$, then for all $t \in [\tau_{\text{mix}}, T]$,*

1. $\sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i \leq 2L_1\alpha_t(\log(1/\alpha_T) + 1);$
2. $\sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i \leq \min\{1/12, \eta/2694\}.$

The proof of Lemma D.2 can be found in Section D.2.

Denote

$$\begin{aligned}\bar{g}(\theta_t) &:= \sum_{s_t, s'_t} \mu(s_t) P(s'_t | s_t) g(s_t, s'_t, \theta_t) = \Phi^\top D(R + P\Phi\theta_t - \Phi\theta_t), \\ \bar{f}(\theta_t) &:= \sum_{s_t, \hat{s}_t} \mu(s_t) \mu(\hat{s}_t) f(s_t, \hat{s}_t, \theta_t) = -\Phi^\top \mu \mu^\top (R + \Phi\theta_t).\end{aligned}$$

We will also need the following two lemmas.

Lemma D.3. Suppose $t_1 \leq t_2$ and $\sum_{i=t_1}^{t_2-1} \alpha_i \leq 1/12$. Then

1. $\|\theta_{t_2} - \theta_{t_1}\| \leq 6(\|\theta_{t_1}\| + r_{\max} + 2\|\theta^*\|) \sum_{i=t_1}^{t_2-1} \alpha_i;$
2. $\|\theta_{t_2} - \theta_{t_1}\| \leq 12(\|\theta_{t_2}\| + r_{\max} + 2\|\theta^*\|) \sum_{i=t_1}^{t_2-1} \alpha_i.$

Lemma D.4. Suppose $t \in [\tau_{\text{mix}}, T]$ and $\sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i \leq 1/12$. Then

$$\begin{aligned}\mathbb{E}[(\theta_t - \theta^*)^\top (g(s_t, s'_t, \theta_t) - \bar{g}(\theta_t))] &\leq (880\mathbb{E}\|\delta_t\|^2 + 304(r_{\max} + 3\|\theta^*\|)^2) \sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i, \\ \mathbb{E}[(\theta_t - \theta^*)^\top (f(s_t, \hat{s}_t, \theta_t) - \bar{f}(\theta_t))] &\leq (448\mathbb{E}\|\delta_t\|^2 + 160(r_{\max} + 3\|\theta^*\|)^2) \sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i.\end{aligned}$$

The proof of Lemma D.3 can be found in Section D.3, and the proof of Lemma D.4 can be found in Section D.4.

Using Lemma D.4 and summing the two bounds, we have

$$\begin{aligned}2\alpha_t \mathbb{E}[\delta_t^\top (\bar{g}(\theta_t) + \bar{f}(\theta_t) - (g(s_t, s'_t, \theta_t) + f(s_t, \hat{s}_t, \theta_t)))] \\ \leq 2\alpha_t (1328\mathbb{E}\|\delta_t\|^2 + 464(r_{\max} + 3\|\theta^*\|)^2) \sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i.\end{aligned}$$

Combining the basic expansion of $\|\delta_{t+1}\|^2$ in Appendix C with the bounds derived above, for any $t \in [\tau_{\text{mix}}, T]$,

$$\begin{aligned}\mathbb{E}\|\delta_{t+1}\|^2 &\leq (1 - 2\eta\alpha_t)\mathbb{E}\|\delta_t\|^2 + 18\alpha_t^2 (\mathbb{E}\|\delta_t\|^2 + (r_{\max} + \|\theta^*\|)^2) \\ &\quad + 2\alpha_t (1328\mathbb{E}\|\delta_t\|^2 + 464(r_{\max} + 3\|\theta^*\|)^2) \sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i.\end{aligned}$$

Since α_t is nonincreasing and $\tau_{\text{mix}} \geq 1$,

$$\sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i \geq \alpha_{t-1} \geq \alpha_t \quad \Rightarrow \quad \alpha_t^2 \leq \alpha_t \sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i.$$

Moreover, $(r_{\max} + \|\theta^*\|)^2 \leq (r_{\max} + 3\|\theta^*\|)^2$. Therefore,

$$\mathbb{E}\|\delta_{t+1}\|^2 \leq (1 - 2\eta\alpha_t)\mathbb{E}\|\delta_t\|^2 + \alpha_t (2674\mathbb{E}\|\delta_t\|^2 + 946(r_{\max} + 3\|\theta^*\|)^2) \sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i.$$

By Lemma D.2(2), we have $\sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i \leq \eta/2694$, which implies $2674 \sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i \leq \eta$. Hence,

$$\mathbb{E}\|\delta_{t+1}\|^2 \leq (1 - \eta\alpha_t)\mathbb{E}\|\delta_t\|^2 + \beta_2 \alpha_t \sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i.$$

Define $\hat{\alpha}_t = \alpha_t \sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i$. Then

$$\mathbb{E}\|\delta_{t+1}\|^2 \leq (1 - \eta\alpha_t)\mathbb{E}\|\delta_t\|^2 + \beta_2 \hat{\alpha}_t, \quad \forall t \in [\tau_{\text{mix}}, T].$$

Recursively applying the above inequality from τ_{mix} to T , we obtain

$$\mathbb{E}\|\delta_T\|^2 \leq \mathbb{E}\|\delta_{\tau_{\text{mix}}}\|^2 \prod_{j=\tau_{\text{mix}}}^{T-1} (1 - \eta\alpha_j) + \beta_2 \sum_{k=\tau_{\text{mix}}}^{T-1} \hat{\alpha}_k \prod_{j=k+1}^{T-1} (1 - \eta\alpha_j).$$

Step 1: bound $\mathbb{E}\|\delta_{\tau_{\text{mix}}}\|^2$. By Lemma D.2(2) applied at $t = \tau_{\text{mix}}$, $\sum_{i=0}^{\tau_{\text{mix}}-1} \alpha_i \leq 1/12$. Then Lemma D.3 (with $t_1 = 0, t_2 = \tau_{\text{mix}}$) yields

$$\|\theta_{\tau_{\text{mix}}} - \theta_0\| \leq 12(\|\theta_0\| + r_{\text{max}} + 2\|\theta^*\|) \sum_{i=0}^{\tau_{\text{mix}}-1} \alpha_i \leq \|\theta_0\| + r_{\text{max}} + 2\|\theta^*\|.$$

Therefore,

$$\begin{aligned} \mathbb{E}\|\delta_{\tau_{\text{mix}}}\|^2 &\leq 2\mathbb{E}\|\theta_{\tau_{\text{mix}}} - \theta_0\|^2 + 2\mathbb{E}\|\theta_0 - \theta^*\|^2 \\ &\leq 2\mathbb{E}\left[(\|\theta_0\| + r_{\text{max}} + 2\|\theta^*\|)^2\right] + 2\mathbb{E}[\|\theta_0 - \theta^*\|^2] = \beta_1. \end{aligned}$$

Step 2: bound the product term. Let

$$I_1 := \prod_{j=\tau_{\text{mix}}}^{T-1} (1 - \eta\alpha_j).$$

Using $1 - x \leq e^{-x}$, we have

$$I_1 \leq \exp\left(-\eta \sum_{j=\tau_{\text{mix}}}^{T-1} \alpha_j\right) = \exp\left(-\eta a \sum_{j=\tau_{\text{mix}}}^{T-1} \frac{1}{(j+c_0)^\xi}\right) \leq \exp\left(-\eta a \int_{\tau_{\text{mix}}}^T \frac{dx}{(x+c_0)^\xi}\right).$$

Thus,

$$I_1 \leq \begin{cases} \left(\frac{\tau_{\text{mix}} + c_0}{T + c_0}\right)^{\eta a}, & \xi = 1, \\ \exp\left(-\frac{\eta a}{1-\xi} \left((T+c_0)^{1-\xi} - (\tau_{\text{mix}}+c_0)^{1-\xi}\right)\right), & \xi \in (0, 1). \end{cases}$$

Step 3: bound the sum term. Let

$$I_2 := \sum_{k=\tau_{\text{mix}}}^{T-1} \hat{\alpha}_k \prod_{j=k+1}^{T-1} (1 - \eta\alpha_j).$$

By Lemma D.2(1),

$$\hat{\alpha}_k = \alpha_k \sum_{i=k-\tau_{\text{mix}}}^{k-1} \alpha_i \leq 2L_1 \alpha_k^2 (\log(1/\alpha_T) + 1), \quad \forall k \in [\tau_{\text{mix}}, T].$$

Hence,

$$\beta_2 I_2 \leq 2L_1 \beta_2 (\log(1/\alpha_T) + 1) \sum_{k=\tau_{\text{mix}}}^{T-1} \alpha_k^2 \prod_{j=k+1}^{T-1} (1 - \eta\alpha_j).$$

Since $\log(1/\alpha_T) = \xi \log(T + c_0) - \log a \leq \log(T + c_0) - \log a$, we can define

$$\tilde{\beta}(T) := 2L_1\beta_2(\log(T + c_0) - \log a + 1)$$

so that

$$\beta_2 I_2 \leq \tilde{\beta}(T) \cdot \underbrace{\sum_{k=\tau_{\text{mix}}}^{T-1} \alpha_k^2 \prod_{j=k+1}^{T-1} (1 - \eta\alpha_j)}_{=: I_3}.$$

We next bound I_3 in two cases.

Case 1: $\xi = 1$. In this case $\alpha_k = \frac{a}{k+c_0}$, and we assume $c_0 \geq a\eta$. Using $1 - x \leq e^{-x}$,

$$\prod_{j=k+1}^{T-1} \left(1 - \frac{\eta a}{j + c_0}\right) \leq \exp\left(-\eta a \sum_{j=k+1}^{T-1} \frac{1}{j + c_0}\right) \leq \left(\frac{k+1+c_0}{T+c_0}\right)^{\eta a}.$$

Therefore,

$$\begin{aligned} I_3 &= \sum_{k=\tau_{\text{mix}}}^{T-1} \frac{a^2}{(k+c_0)^2} \prod_{j=k+1}^{T-1} \left(1 - \frac{\eta a}{j + c_0}\right) \\ &\leq \sum_{k=\tau_{\text{mix}}}^{T-1} \frac{a^2}{(k+c_0)^2} \left(\frac{k+1+c_0}{T+c_0}\right)^{\eta a} = \frac{a^2}{(T+c_0)^{\eta a}} \sum_{k=\tau_{\text{mix}}}^{T-1} \left(\frac{k+1+c_0}{k+c_0}\right)^2 (k+1+c_0)^{\eta a-2} \\ &\leq \frac{4a^2}{(T+c_0)^{\eta a}} \sum_{k=\tau_{\text{mix}}}^{T-1} (k+1+c_0)^{\eta a-2}. \end{aligned}$$

Standard summation bounds give

$$I_3 \leq \begin{cases} \frac{4a^2}{1-\eta a} \cdot \frac{1}{(T+c_0)^{\eta a}}, & \eta a < 1, \\ 4a^2 \frac{\log(T+c_0)}{T+c_0}, & \eta a = 1, \\ \frac{4ea^2}{\eta a - 1} \cdot \frac{1}{T+c_0}, & \eta a > 1. \end{cases}$$

Define $q = \min\{1, \eta a\}$ and

$$\Gamma(T) := \begin{cases} \frac{4a^2}{1-\eta a} \tilde{\beta}(T), & \eta a < 1, \\ 4a^2 \log(T+c_0) \tilde{\beta}(T), & \eta a = 1, \\ \frac{4ea^2}{\eta a - 1} \tilde{\beta}(T), & \eta a > 1. \end{cases}$$

Then $\tilde{\beta}(T)I_3 \leq \Gamma(T)/(T+c_0)^q$. Combining with the bound on I_1 , we obtain the $\xi = 1$ claim.

Case 2: $\xi \in (0, 1)$. Consider the sequence $\{u_t\}_{t \geq \tau_{\text{mix}}}$ defined by

$$u_{\tau_{\text{mix}}} = 0, \quad u_{t+1} = \left(1 - \frac{\eta a}{(t+c_0)^\xi}\right) u_t + \frac{a^2}{(t+c_0)^{2\xi}}.$$

One can check that $I_3 = u_T$. To bound u_T , we use the following lemma.

Lemma D.5. Given a sequence $\{x_t\}_{t \geq \tau}$ and positive constants c_0, c_1, c_2, ξ , consider the recursion

$$x_{t+1} = \left(1 - \frac{c_1 c_2}{(t + c_0)^\xi}\right) x_t + \frac{c_2^2}{(t + c_0)^{2\xi}},$$

with initial condition $x_\tau \leq \frac{2c_2}{c_1} \frac{1}{(\tau + c_0)^\xi}$. Then $x_t \leq \frac{2c_2}{c_1} \frac{1}{(t + c_0)^\xi}$ for all $t \geq \tau$ if either:

1. $\xi = 1$ and $c_1 c_2 \geq 2$;
2. $\xi \in (0, 1)$ and $\tau \geq (2\xi/(c_1 c_2))^{1/(1-\xi)} - c_0$.

The proof of Lemma D.5 can be found in Section D.5.

Applying Lemma D.5 with $c_1 = \eta$, $c_2 = a$, $\tau = \tau_{\text{mix}}$, and noting that $u_{\tau_{\text{mix}}} = 0 \leq \frac{2a}{\eta} \frac{1}{(\tau_{\text{mix}} + c_0)^\xi}$, we get (under the condition $c_0 \geq (\frac{2\xi}{\eta a})^{1/(1-\xi)}$)

$$I_3 = u_T \leq \frac{2a}{\eta(T + c_0)^\xi}.$$

Therefore,

$$\tilde{\beta}(T) I_3 \leq \frac{2a \tilde{\beta}(T)}{\eta(T + c_0)^\xi} = \frac{\beta(T)}{\eta(T + c_0)^\xi},$$

where $\beta(T) = 4aL_1\beta_2(\log(T + c_0) - \log a + 1)$. Combining with the bound on I_1 proves the $\xi \in (0, 1)$ claim. $\square$

D.2. Proof of Lemma D.2

Proof of Lemma D.2. By definition of $\tau_{\text{mix}} = \tau_{\text{mix}}(\alpha_T)$, we have $C\beta^{\tau_{\text{mix}}} \leq \alpha_T$, which implies

$$\tau_{\text{mix}} \leq \frac{\log C + \log(1/\alpha_T)}{\log(1/\beta)} \leq L_1(\log(1/\alpha_T) + 1), \quad L_1 = \max \left\{ 1, \frac{\log C + 1}{\log(1/\beta)} \right\}.$$

Since α_t is nonincreasing,

$$\sum_{i=\tau_{\text{mix}}}^{t-1} \alpha_i \leq \tau_{\text{mix}} \alpha_{t-\tau_{\text{mix}}} = \tau_{\text{mix}} \alpha_t \left(\frac{t + c_0}{t - \tau_{\text{mix}} + c_0} \right)^\xi.$$

Note that $\left(\frac{t + c_0}{t - \tau_{\text{mix}} + c_0} \right)^\xi \rightarrow 1$ as $c_0 \rightarrow \infty$ uniformly over $t \in [\tau_{\text{mix}}, T]$. Hence, there exists $c_{0,1}(a, \xi)$ large enough such that for all $t \in [\tau_{\text{mix}}, T]$,

$$\left(\frac{t + c_0}{t - \tau_{\text{mix}} + c_0} \right)^\xi \leq 2, \quad \forall c_0 \geq c_{0,1}(a, \xi).$$

Using $\tau_{\text{mix}} \leq L_1(\log(1/\alpha_T) + 1)$, we obtain

$$\sum_{i=\tau_{\text{mix}}}^{t-1} \alpha_i \leq 2L_1 \alpha_t (\log(1/\alpha_T) + 1), \quad \forall t \in [\tau_{\text{mix}}, T],$$

which proves part (1).

For part (2), since $\alpha_t \rightarrow 0$ as $c_0 \rightarrow \infty$ (for fixed T) and the right-hand side in part (1) is $O(\alpha_t \log(1/\alpha_T))$, there exists $c_{0,2}(a, \xi)$ large enough such that for all $c_0 \geq c_{0,2}(a, \xi)$,

$$2L_1 \alpha_t (\log(1/\alpha_T) + 1) \leq \min\{1/12, \eta/2694\}, \quad \forall t \in [\tau_{\text{mix}}, T].$$

Combining with part (1) proves part (2). $\square$

D.3. Proof of Lemma D.3

Proof of Lemma D.3. By Lemma C.1,

$$\|\theta_{t+1}\| - \|\theta_t\| \leq \|\theta_{t+1} - \theta_t\| \leq \alpha_t \|f(s_t, \hat{s}_t, \theta_t) + g(s_t, s'_t, \theta_t)\| \leq 3\alpha_t (\|\theta_t\| + r_{\max} + 2\|\theta^*\|).$$

Thus,

$$\|\theta_{t+1}\| + r_{\max} + 2\|\theta^*\| \leq (1 + 3\alpha_t)(\|\theta_t\| + r_{\max} + 2\|\theta^*\|).$$

For any $t \in [t_1, t_2]$,

$$\begin{aligned} \|\theta_t\| + r_{\max} + 2\|\theta^*\| &\leq \prod_{i=t_1}^{t-1} (1 + 3\alpha_i)(\|\theta_{t_1}\| + r_{\max} + 2\|\theta^*\|) \\ &\leq \exp\left(3 \sum_{i=t_1}^{t-1} \alpha_i\right) (\|\theta_{t_1}\| + r_{\max} + 2\|\theta^*\|) \\ &\leq \left(1 + 6 \sum_{i=t_1}^{t-1} \alpha_i\right) (\|\theta_{t_1}\| + r_{\max} + 2\|\theta^*\|), \end{aligned}$$

where the last step uses $e^x \leq 1 + 2x$ for $x \leq 1/2$, and here $x = 3 \sum_{i=t_1}^{t-1} \alpha_i \leq 3 \cdot (1/12) = 1/4$. In particular, since $\sum_{i=t_1}^{t_2-1} \alpha_i \leq 1/12$,

$$\|\theta_t\| + r_{\max} + 2\|\theta^*\| \leq 2(\|\theta_{t_1}\| + r_{\max} + 2\|\theta^*\|).$$

Therefore,

$$\begin{aligned} \|\theta_{t_2} - \theta_{t_1}\| &\leq \sum_{i=t_1}^{t_2-1} \|\theta_{i+1} - \theta_i\| \leq \sum_{i=t_1}^{t_2-1} 3\alpha_i (\|\theta_i\| + r_{\max} + 2\|\theta^*\|) \\ &\leq 6(\|\theta_{t_1}\| + r_{\max} + 2\|\theta^*\|) \sum_{i=t_1}^{t_2-1} \alpha_i, \end{aligned}$$

which proves part (1). For part (2), using $\|\theta_{t_1}\| \leq \|\theta_{t_2}\| + \|\theta_{t_2} - \theta_{t_1}\|$,

$$\begin{aligned} \|\theta_{t_2} - \theta_{t_1}\| &\leq 6(\|\theta_{t_1}\| + r_{\max} + 2\|\theta^*\|) \sum_{i=t_1}^{t_2-1} \alpha_i \\ &\leq 6(\|\theta_{t_2}\| + \|\theta_{t_2} - \theta_{t_1}\| + r_{\max} + 2\|\theta^*\|) \sum_{i=t_1}^{t_2-1} \alpha_i \\ &\leq \frac{1}{2} \|\theta_{t_2} - \theta_{t_1}\| + 6(\|\theta_{t_2}\| + r_{\max} + 2\|\theta^*\|) \sum_{i=t_1}^{t_2-1} \alpha_i, \end{aligned}$$

where we used $\sum_{i=t_1}^{t_2-1} \alpha_i \leq 1/12$. Rearranging gives

$$\|\theta_{t_2} - \theta_{t_1}\| \leq 12(\|\theta_{t_2}\| + r_{\max} + 2\|\theta^*\|) \sum_{i=t_1}^{t_2-1} \alpha_i,$$

which proves part (2). □

D.4. Proof of Lemma D.4

Proof of Lemma D.4. For simplicity, denote

$$g_t(\theta) = g(s_t, s'_t, \theta), \quad f_t(\theta) = f(s_t, \hat{s}_t, \theta).$$

Throughout this proof we fix $t \in [\tau_{\text{mix}}, T]$ and assume $\sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i \leq 1/12$.

By Lemma D.3 (part (2)),

$$\begin{aligned} \|\theta_t - \theta_{t-\tau_{\text{mix}}}\| &\leq 12(\|\theta_t\| + r_{\text{max}} + 2\|\theta^*\|) \sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i \\ &\leq 12(\|\delta_t\| + r_{\text{max}} + 3\|\theta^*\|) \sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i. \end{aligned}$$

Step 1: bound the noise term for g . Decompose

$$\begin{aligned} &\mathbb{E}[(\theta_t - \theta^*)^\top (g_t(\theta_t) - \bar{g}(\theta_t))] \\ = &\underbrace{\mathbb{E}[(\theta_t - \theta_{t-\tau_{\text{mix}}})^\top (g_t(\theta_t) - \bar{g}(\theta_t))]}_{I_1} + \underbrace{\mathbb{E}[(\theta_{t-\tau_{\text{mix}}} - \theta^*)^\top (g_t(\theta_{t-\tau_{\text{mix}}}) - \bar{g}(\theta_{t-\tau_{\text{mix}}}))]}_{I_2} \\ &+ \underbrace{\mathbb{E}[(\theta_{t-\tau_{\text{mix}}} - \theta^*)^\top (g_t(\theta_t) - g_t(\theta_{t-\tau_{\text{mix}}}))]}_{I_3} + \underbrace{\mathbb{E}[(\theta_{t-\tau_{\text{mix}}} - \theta^*)^\top (\bar{g}(\theta_{t-\tau_{\text{mix}}}) - \bar{g}(\theta_t))]}_{I_4}. \end{aligned}$$

Term I_1 . By Lemma C.1, $\|g_t(\theta_t)\| \leq 2(\|\delta_t\| + r_{\text{max}} + \|\theta^*\|)$, hence $\|g_t(\theta_t) - \bar{g}(\theta_t)\| \leq 4(\|\delta_t\| + r_{\text{max}} + \|\theta^*\|)$. Thus,

$$\begin{aligned} I_1 &\leq \mathbb{E}[\|\theta_t - \theta_{t-\tau_{\text{mix}}}\| \cdot \|g_t(\theta_t) - \bar{g}(\theta_t)\|] \\ &\leq \mathbb{E}\left[12 \sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i (\|\delta_t\| + r_{\text{max}} + 3\|\theta^*\|) \cdot 4(\|\delta_t\| + r_{\text{max}} + \|\theta^*\|)\right] \\ &\leq 48 \sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i \mathbb{E}[(\|\delta_t\| + r_{\text{max}} + 3\|\theta^*\|)^2] \\ &\leq 96 \sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i (\mathbb{E}\|\delta_t\|^2 + (r_{\text{max}} + 3\|\theta^*\|)^2). \end{aligned}$$

Term I_2 . First note that

$$\begin{aligned} \|\delta_{t-\tau_{\text{mix}}}\| &\leq \|\delta_t\| + \|\theta_t - \theta_{t-\tau_{\text{mix}}}\| \\ &\leq \|\delta_t\| + 12(\|\delta_t\| + r_{\text{max}} + 3\|\theta^*\|) \sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i \\ &\leq 2\|\delta_t\| + r_{\text{max}} + 3\|\theta^*\|, \end{aligned}$$

where we used $\sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i \leq 1/12$.

Next, by Eq.(4)

$$\sup_s \|P_{\tau_{\text{mix}}}(\cdot | s) - \mu\|_1 \leq C\beta^{\tau_{\text{mix}}} \leq \alpha_T \leq \alpha_t.$$

Therefore,

$$\begin{aligned} \left\| \mathbb{E}[g_t(\theta_{t-\tau_{\text{mix}}}) - \bar{g}(\theta_{t-\tau_{\text{mix}}}) | \mathcal{F}_{t-\tau_{\text{mix}}}] \right\| &\leq \sup_s \|P_{\tau_{\text{mix}}}(\cdot | s) - \mu\|_1 \cdot \sup_{s, s'} \|g(s, s', \theta_{t-\tau_{\text{mix}}})\| \\ &\leq \alpha_t \cdot 2(\|\delta_{t-\tau_{\text{mix}}}\| + r_{\text{max}} + \|\theta^*\|) \\ &\leq 4\alpha_t(\|\delta_t\| + r_{\text{max}} + 2\|\theta^*\|). \end{aligned}$$

Hence,

$$\begin{aligned}
 I_2 &\leq \mathbb{E} \left[\|\delta_{t-\tau_{\text{mix}}}\| \cdot \left\| \mathbb{E} [g_t(\theta_{t-\tau_{\text{mix}}}) - \bar{g}(\theta_{t-\tau_{\text{mix}}}) \mid \mathcal{F}_{t-\tau_{\text{mix}}}] \right\| \right] \\
 &\leq \mathbb{E} \left[(2\|\delta_t\| + r_{\text{max}} + 3\|\theta^*\|) \cdot 4\alpha_t (\|\delta_t\| + r_{\text{max}} + 2\|\theta^*\|) \right] \\
 &\leq 8\alpha_t \mathbb{E} \left[(\|\delta_t\| + r_{\text{max}} + 2\|\theta^*\|)^2 \right] \\
 &\leq 16\alpha_t \left(\mathbb{E} \|\delta_t\|^2 + (r_{\text{max}} + 2\|\theta^*\|)^2 \right).
 \end{aligned}$$

Terms I_3 and I_4 . By Lemma C.2, $g(\cdot)$ is 2-Lipschitz in θ , and $\bar{g}(\cdot)$ is also 2-Lipschitz. Thus,

$$\begin{aligned}
 I_3 &\leq \mathbb{E} \left[\|\delta_{t-\tau_{\text{mix}}}\| \cdot \|g_t(\theta_t) - g_t(\theta_{t-\tau_{\text{mix}}})\| \right] \leq 2\mathbb{E} \left[\|\delta_{t-\tau_{\text{mix}}}\| \cdot \|\theta_t - \theta_{t-\tau_{\text{mix}}}\| \right], \\
 I_4 &\leq \mathbb{E} \left[\|\delta_{t-\tau_{\text{mix}}}\| \cdot \|\bar{g}(\theta_t) - \bar{g}(\theta_{t-\tau_{\text{mix}}})\| \right] \leq 2\mathbb{E} \left[\|\delta_{t-\tau_{\text{mix}}}\| \cdot \|\theta_t - \theta_{t-\tau_{\text{mix}}}\| \right].
 \end{aligned}$$

Using the bounds on $\|\delta_{t-\tau_{\text{mix}}}\|$ and $\|\theta_t - \theta_{t-\tau_{\text{mix}}}\|$,

$$\begin{aligned}
 I_3 &\leq \mathbb{E} \left[(2\|\delta_t\| + r_{\text{max}} + 3\|\theta^*\|) \cdot 24(\|\delta_t\| + r_{\text{max}} + 3\|\theta^*\|) \sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i \right] \\
 &\leq 48 \sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i \mathbb{E} \left[(2\|\delta_t\| + r_{\text{max}} + 3\|\theta^*\|)^2 \right] \\
 &\leq 96 \sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i \left(4\mathbb{E} \|\delta_t\|^2 + (r_{\text{max}} + 3\|\theta^*\|)^2 \right),
 \end{aligned}$$

and the same bound holds for I_4 .

Combining I_1 – I_4 , we get

$$\begin{aligned}
 \mathbb{E} \left[(\theta_t - \theta^*)^\top (g_t(\theta_t) - \bar{g}(\theta_t)) \right] &\leq 16\alpha_t \left(\mathbb{E} \|\delta_t\|^2 + (r_{\text{max}} + 2\|\theta^*\|)^2 \right) \\
 &\quad + \left(864\mathbb{E} \|\delta_t\|^2 + 288(r_{\text{max}} + 3\|\theta^*\|)^2 \right) \sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i.
 \end{aligned}$$

Since $\alpha_t \leq \sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i$, the above implies

$$\mathbb{E} \left[(\theta_t - \theta^*)^\top (g_t(\theta_t) - \bar{g}(\theta_t)) \right] \leq \left(880\mathbb{E} \|\delta_t\|^2 + 304(r_{\text{max}} + 3\|\theta^*\|)^2 \right) \sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i.$$

Step 2: bound the noise term for f . Repeat the same decomposition for f :

$$\begin{aligned}
 &\mathbb{E} \left[(\theta_t - \theta^*)^\top (f_t(\theta_t) - \bar{f}(\theta_t)) \right] \\
 &= \underbrace{\mathbb{E} \left[(\theta_t - \theta_{t-\tau_{\text{mix}}})^\top (f_t(\theta_t) - \bar{f}(\theta_t)) \right]}_{I'_1} + \underbrace{\mathbb{E} \left[(\theta_{t-\tau_{\text{mix}}} - \theta^*)^\top (f_t(\theta_{t-\tau_{\text{mix}}}) - \bar{f}(\theta_{t-\tau_{\text{mix}}})) \right]}_{I'_2} \\
 &\quad + \underbrace{\mathbb{E} \left[(\theta_{t-\tau_{\text{mix}}} - \theta^*)^\top (f_t(\theta_t) - f_t(\theta_{t-\tau_{\text{mix}}})) \right]}_{I'_3} + \underbrace{\mathbb{E} \left[(\theta_{t-\tau_{\text{mix}}} - \theta^*)^\top (\bar{f}(\theta_{t-\tau_{\text{mix}}}) - \bar{f}(\theta_t)) \right]}_{I'_4}.
 \end{aligned}$$

Term I'_1 . By Lemma C.1, $\|f_t(\theta_t)\| \leq \|\delta_t\| + r_{\max} + \|\theta^*\|$, hence $\|f_t(\theta_t) - \bar{f}(\theta_t)\| \leq 2(\|\delta_t\| + r_{\max} + \|\theta^*\|)$. Thus,

$$\begin{aligned} I'_1 &\leq \mathbb{E} \left[\|\theta_t - \theta_{t-\tau_{\text{mix}}}\| \cdot \|f_t(\theta_t) - \bar{f}(\theta_t)\| \right] \\ &\leq \mathbb{E} \left[12 \sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i (\|\delta_t\| + r_{\max} + 3\|\theta^*\|) \cdot 2(\|\delta_t\| + r_{\max} + \|\theta^*\|) \right] \\ &\leq 48 \sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i \left(\mathbb{E} \|\delta_t\|^2 + (r_{\max} + \|\theta^*\|)^2 \right). \end{aligned}$$

Term I'_2 . By Eq.(4) (for both chains s_t and $\hat{s}_t$),

$$\sup_s \|P_{\tau_{\text{mix}}}(\cdot | s) - \mu\|_1 \leq \alpha_t, \quad \sup_{\hat{s}} \|P_{\tau_{\text{mix}}}(\cdot | \hat{s}) - \mu\|_1 \leq \alpha_t.$$

Hence,

$$\begin{aligned} \left\| \mathbb{E} [f_t(\theta_{t-\tau_{\text{mix}}}) - \bar{f}(\theta_{t-\tau_{\text{mix}}}) | \mathcal{F}_{t-\tau_{\text{mix}}}] \right\| &\leq \left(\sup_s \|P_{\tau_{\text{mix}}}(\cdot | s) - \mu\|_1 + \sup_{\hat{s}} \|P_{\tau_{\text{mix}}}(\cdot | \hat{s}) - \mu\|_1 \right) \cdot \sup_{s, \hat{s}} \|f(s, \hat{s}, \theta_{t-\tau_{\text{mix}}})\| \\ &\leq 2\alpha_t \cdot (\|\delta_{t-\tau_{\text{mix}}}\| + r_{\max} + \|\theta^*\|) \\ &\leq 4\alpha_t (\|\delta_t\| + r_{\max} + 2\|\theta^*\|). \end{aligned}$$

Therefore,

$$\begin{aligned} I'_2 &\leq \mathbb{E} \left[\|\delta_{t-\tau_{\text{mix}}}\| \cdot \left\| \mathbb{E} [f_t(\theta_{t-\tau_{\text{mix}}}) - \bar{f}(\theta_{t-\tau_{\text{mix}}}) | \mathcal{F}_{t-\tau_{\text{mix}}}] \right\| \right] \\ &\leq 16\alpha_t \left(\mathbb{E} \|\delta_t\|^2 + (r_{\max} + 2\|\theta^*\|)^2 \right), \end{aligned}$$

where we used the same bound $\|\delta_{t-\tau_{\text{mix}}}\| \leq 2\|\delta_t\| + r_{\max} + 3\|\theta^*\|$.

Terms I'_3 and I'_4 . By Lemma C.2, $f(\cdot)$ is 1-Lipschitz in θ , and $\bar{f}(\cdot)$ is also 1-Lipschitz, hence

$$I'_3 \leq \mathbb{E} [\|\delta_{t-\tau_{\text{mix}}}\| \cdot \|\theta_t - \theta_{t-\tau_{\text{mix}}}\|], \quad I'_4 \leq \mathbb{E} [\|\delta_{t-\tau_{\text{mix}}}\| \cdot \|\theta_t - \theta_{t-\tau_{\text{mix}}}\|].$$

Thus,

$$\begin{aligned} I'_3 &\leq \mathbb{E} \left[(2\|\delta_t\| + r_{\max} + 3\|\theta^*\|) \cdot 12(\|\delta_t\| + r_{\max} + 3\|\theta^*\|) \sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i \right] \\ &\leq 24 \sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i \mathbb{E} \left[(2\|\delta_t\| + r_{\max} + 3\|\theta^*\|)^2 \right] \\ &\leq 48 \sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i \left(4\mathbb{E} \|\delta_t\|^2 + (r_{\max} + 3\|\theta^*\|)^2 \right), \end{aligned}$$

and the same bound holds for I'_4 .

Combining I'_1 - I'_4 , we obtain

$$\begin{aligned} \mathbb{E} \left[(\theta_t - \theta^*)^\top (f_t(\theta_t) - \bar{f}(\theta_t)) \right] &\leq 16\alpha_t \left(\mathbb{E} \|\delta_t\|^2 + (r_{\max} + 2\|\theta^*\|)^2 \right) \\ &\quad + \left(432\mathbb{E} \|\delta_t\|^2 + 144(r_{\max} + 3\|\theta^*\|)^2 \right) \sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i. \end{aligned}$$

Using $\alpha_t \leq \sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i$ gives the desired bound for f , and this completes the proof of Lemma D.4. $\square$

D.5. Proof of Lemma D.5

Proof of Lemma D.5. We use induction. The base case is assumed in the statement. Now suppose $x_t \leq \frac{2c_2}{c_1} \frac{1}{(t+c_0)^\xi}$. Then

$$\begin{aligned} \frac{2c_2}{c_1} \frac{1}{(t+1+c_0)^\xi} - x_{t+1} &= \frac{2c_2}{c_1} \frac{1}{(t+1+c_0)^\xi} - \left(1 - \frac{c_1 c_2}{(t+c_0)^\xi}\right) x_t - \frac{c_2^2}{(t+c_0)^{2\xi}} \\ &\geq \frac{2c_2}{c_1} \frac{1}{(t+1+c_0)^\xi} - \left(1 - \frac{c_1 c_2}{(t+c_0)^\xi}\right) \frac{2c_2}{c_1} \frac{1}{(t+c_0)^\xi} - \frac{c_2^2}{(t+c_0)^{2\xi}} \\ &= \frac{2c_2}{c_1} \left[\frac{1}{(t+1+c_0)^\xi} - \frac{1}{(t+c_0)^\xi} + \frac{c_1 c_2}{2} \frac{1}{(t+c_0)^{2\xi}} \right] \\ &= \frac{2c_2}{c_1} \frac{1}{(t+c_0)^{2\xi}} \left[\frac{c_1 c_2}{2} - (t+c_0)^\xi \left(1 - \left(\frac{t+c_0}{t+1+c_0}\right)^\xi\right) \right]. \end{aligned}$$

Observe that

$$\left(\frac{t+c_0}{t+1+c_0}\right)^\xi = \left(1 + \frac{1}{t+c_0}\right)^{-\xi} \geq \exp\left(-\frac{\xi}{t+c_0}\right) \geq 1 - \frac{\xi}{t+c_0},$$

where we used $e^x \geq 1 + x$. Hence,

$$(t+c_0)^\xi \left(1 - \left(\frac{t+c_0}{t+1+c_0}\right)^\xi\right) \leq (t+c_0)^\xi \cdot \frac{\xi}{t+c_0} = \frac{\xi}{(t+c_0)^{1-\xi}}.$$

Therefore,

$$\frac{2c_2}{c_1} \frac{1}{(t+1+c_0)^\xi} - x_{t+1} \geq \frac{2c_2}{c_1} \frac{1}{(t+c_0)^{2\xi}} \left(\frac{c_1 c_2}{2} - \frac{\xi}{(t+c_0)^{1-\xi}} \right).$$

The last term is nonnegative under either condition in the lemma statement, which yields $x_{t+1} \leq \frac{2c_2}{c_1} \frac{1}{(t+1+c_0)^\xi}$. This completes the induction. $\square$

E. Analysis of Single Markov Chain Case with Constant Step-size

Recall the single-chain update components in Eq. (17):

$$f(s, w) := \phi(s) - w, \quad g(s, s', w, \theta) := (r_s + \phi(s')^\top \theta - \phi(s)^\top \theta) \phi(s) - (r_s + \phi(s)^\top \theta) w.$$

Under stationarity $(s, s') \sim (\mu, P)$, define the mean fields

$$\bar{f}(w) := \mathbb{E}_{s \sim \mu} [f(s, w)] = \Phi^\top \mu - w = w^* - w, \quad (23)$$

$$\bar{g}(w, \theta) := \mathbb{E}_{s \sim \mu, s' \sim P(\cdot|s)} [g(s, s', w, \theta)] = \Phi^\top D(R + P\Phi\theta - \Phi\theta) - \mu^\top (R + \Phi\theta) w. \quad (24)$$

The fixed points are $w^* = \Phi^\top \mu$, and θ^* satisfies

$$\bar{g}(w^*, \theta^*) = 0. \quad (25)$$

First, we introduce some useful basic properties.

Lemma E.1. For all $t \geq 0$, $\|w_t\| \leq 1$ and $\|\delta_t^w\| \leq 2$.

Proof of Lemma E.1. The update $w_{t+1} = (1 - \beta)w_t + \beta\phi(s_t)$ implies

$$\|w_{t+1}\| \leq (1 - \beta)\|w_t\| + \beta\|\phi(s_t)\| \leq (1 - \beta)\|w_t\| + \beta.$$

Since $\|w_0\| \leq 1$, induction gives $\|w_t\| \leq 1$ for all t . Also $\|w^*\| = \|\mathbb{E}_{s \sim \mu} [\phi(s)]\| \leq \mathbb{E}\|\phi(s)\| \leq 1$, hence $\|\delta_t^w\| \leq 2$. $\square$

Lemma E.2. For all $t \geq 0$,

$$\|f(s_t, w_t)\| \leq 2, \quad \|g(s_t, s'_t, w_t, \theta_t)\| \leq 2r_{\max} + 6R_\theta = 2(r_{\max} + 3R_\theta).$$

Proof of Lemma E.2. For $f(s, w) = \phi(s) - w$, $\|f(s_t, w_t)\| \leq \|\phi(s_t)\| + \|w_t\| \leq 2$. For g , using $\|\phi(\cdot)\| \leq 1$ and $\|w_t\| \leq 1$,

$$\begin{aligned} \|g(s_t, s'_t, w_t, \theta_t)\| &\leq |r_{s_t} + (\phi(s'_t) - \phi(s_t))^\top \theta_t| \cdot \|\phi(s_t)\| + |r_{s_t} + \phi(s_t)^\top \theta_t| \cdot \|w_t\| \\ &\leq (r_{\max} + 2\|\theta_t\|) + (r_{\max} + \|\theta_t\|) \leq 2r_{\max} + 3\|\theta_t\| \leq 2r_{\max} + 6R_\theta. \end{aligned}$$

$\square$

Lemma E.3. For all s, s' and all $w, w_1, w_2, \theta, \theta_1, \theta_2$,

$$\|f(s, w_1) - f(s, w_2)\| \leq \|w_1 - w_2\|,$$

$$\|g(s, s', w, \theta_1) - g(s, s', w, \theta_2)\| \leq 3\|\theta_1 - \theta_2\|,$$

and

$$\|g(s, s', w_1, \theta) - g(s, s', w_2, \theta)\| \leq (r_{\max} + 2R_\theta) \|w_1 - w_2\|.$$

Proof of Lemma E.3. The f claim is immediate: $\|f(s, w_1) - f(s, w_2)\| = \|w_1 - w_2\|$.

For θ ,

$$g(s, s', w, \theta_1) - g(s, s', w, \theta_2) = ((\phi(s') - \phi(s))^\top (\theta_1 - \theta_2)) \phi(s) - (\phi(s)^\top (\theta_1 - \theta_2)) w,$$

so

$$\|g(\cdot, \theta_1) - g(\cdot, \theta_2)\| \leq \|\phi(s') - \phi(s)\| \cdot \|\theta_1 - \theta_2\| \cdot \|\phi(s)\| + \|\phi(s)\| \cdot \|\theta_1 - \theta_2\| \cdot \|w\| \leq (2 + 1)\|\theta_1 - \theta_2\|.$$

For w ,

$$g(s, s', w_1, \theta) - g(s, s', w_2, \theta) = -(r_s + \phi(s)^\top \theta)(w_1 - w_2),$$

hence $\|g(\cdot, w_1) - g(\cdot, w_2)\| \leq (r_{\max} + \|\theta\|)\|w_1 - w_2\| \leq (r_{\max} + 2R_\theta)\|w_1 - w_2\|$. $\square$

E.1. Proof of Theorem 4.4

We first restate the theorem with all the constants.

Theorem E.4 (Full statement of Theorem 4.4). *Consider the single-chain algorithm in Eq. (17) with Markov sampling. Assume constant stepsizes $\alpha_t \equiv \alpha > 0$, $\beta_t \equiv \beta > 0$ and $\rho_0 := \beta/\alpha \leq 1$. Let $\lambda > 0$ satisfy $0 < \lambda^2 < \frac{2\eta}{r_{\max} + 2R_\theta}$ and define*

$$\begin{aligned}\zeta &:= \eta - \frac{\lambda^2(r_{\max} + 2R_\theta)}{2} > 0, \\ \kappa &:= 1 - 2\alpha\zeta \in (0, 1), \\ G_1 &:= e^{-2\rho_0\alpha}.\end{aligned}$$

Then for any $t \geq \tau_{\text{mix}}$,

$$\begin{aligned}\mathbb{E}\|\delta_t^\theta\|^2 &\leq \frac{4\alpha(r_{\max} + 2R_\theta)(t - \tau_{\text{mix}}) \max\{\kappa, G_1\}^{t - \tau_{\text{mix}}}}{\lambda^2} \\ &\quad + 4R_\theta^2 \kappa^{t - \tau_{\text{mix}}} + \frac{\alpha G_{\text{const}}}{2\zeta},\end{aligned}$$

where

$$\begin{aligned}G_{\text{const}} &:= \frac{(16\tau_{\text{mix}} + 6)(r_{\max} + 2R_\theta)}{\lambda^2} \\ &\quad + (88\tau_{\text{mix}} + 4)(r_{\max} + 3R_\theta)^2.\end{aligned}$$

Proof of Theorem 4.4. By non-expansiveness of projection,

$$\|\delta_{t+1}^\theta\|^2 \leq \|\delta_t^\theta + \alpha g(s_t, s'_t, w_t, \theta_t)\|^2. \quad (26)$$

Taking expectation and expanding,

$$\begin{aligned}\mathbb{E}\|\delta_{t+1}^\theta\|^2 &\leq \mathbb{E}\|\delta_t^\theta\|^2 + 2\alpha \mathbb{E}\left[\delta_t^{\theta^\top} g(s_t, s'_t, w_t, \theta_t)\right] + \alpha^2 \mathbb{E}\|g(s_t, s'_t, w_t, \theta_t)\|^2 \\ &= \mathbb{E}\|\delta_t^\theta\|^2 + 2\alpha \mathbb{E}\left[\delta_t^{\theta^\top} (g(s_t, s'_t, w_t, \theta_t) - \bar{g}(w_t, \theta_t))\right] + 2\alpha \mathbb{E}\left[\delta_t^{\theta^\top} \bar{g}(w_t, \theta_t)\right] + \alpha^2 \mathbb{E}\|g(s_t, s'_t, w_t, \theta_t)\|^2.\end{aligned} \quad (27)$$

Step 1: Markov-noise term.

Lemma E.5. *For any $t \geq \tau_{\text{mix}}$,*

$$\mathbb{E}\left[\delta_t^{\theta^\top} (g(s_t, s'_t, w_t, \theta_t) - \bar{g}(w_t, \theta_t))\right] \leq 44\alpha\tau_{\text{mix}}(r_{\max} + 3R_\theta)^2.$$

The proof of this lemma can be found in Section E.2.

By Lemma E.5, for $t \geq \tau_{\text{mix}}$,

$$2\alpha \mathbb{E}\left[\delta_t^{\theta^\top} (g(s_t, s'_t, w_t, \theta_t) - \bar{g}(w_t, \theta_t))\right] \leq 88\alpha^2\tau_{\text{mix}}(r_{\max} + 3R_\theta)^2. \quad (28)$$

Step 2: Mean-drift term. Using $\bar{g}(w^*, \theta^*) = 0$,

$$\delta_t^{\theta^\top} \bar{g}(w_t, \theta_t) = -\|\Phi\delta_t^\theta\|_{\text{Dir}}^2 - (\mu^\top \Phi\delta_t^\theta)^2 - (\mu^\top (R + \Phi\theta_t))(\delta_t^{\theta^\top} \delta_t^w).$$

By Eq.(8) and Young's inequality, for any $\lambda > 0$,

$$|\mu^\top (R + \Phi\theta_t)| |\delta_t^{\theta^\top} \delta_t^w| \leq (r_{\max} + 2R_\theta) \cdot \frac{1}{2} \left(\lambda^2 \|\delta_t^\theta\|^2 + \|\delta_t^w\|^2 / \lambda^2 \right).$$

Therefore, with $\zeta = \eta - \lambda^2(r_{\max} + 2R_\theta)/2 > 0$,

$$2\alpha \mathbb{E}[\delta_t^{\theta^\top} \bar{g}(w_t, \theta_t)] \leq -2\alpha\zeta \mathbb{E}\|\delta_t^\theta\|^2 + \alpha \frac{r_{\max} + 2R_\theta}{\lambda^2} \mathbb{E}\|\delta_t^w\|^2. \quad (29)$$

Step 3: Second-moment term. By Lemma E.2,

$$\alpha^2 \mathbb{E}\|g\|^2 \leq 4\alpha^2 (r_{\max} + 3R_\theta)^2. \quad (30)$$

Combining (27)–(30), for $t \geq \tau_{\text{mix}}$,

$$\mathbb{E}\|\delta_{t+1}^\theta\|^2 \leq (1 - 2\alpha\zeta) \mathbb{E}\|\delta_t^\theta\|^2 + \alpha \frac{r_{\max} + 2R_\theta}{\lambda^2} \mathbb{E}\|\delta_t^w\|^2 + \alpha^2(88\tau_{\text{mix}} + 4)(r_{\max} + 3R_\theta)^2. \quad (31)$$

To deal with $\mathbb{E}\|\delta_t^w\|^2$, we introduce the following lemma:

Lemma E.6. For any $t \geq \tau_{\text{mix}}$,

$$\mathbb{E}\|\delta_t^w\|^2 \leq 4e^{-2\beta(t-\tau_{\text{mix}})} + (16\tau_{\text{mix}} + 6)\beta = 4G_1^{t-\tau_{\text{mix}}} + (16\tau_{\text{mix}} + 6)\beta.$$

The proof of this lemma can be found in Section E.3.

Now apply Lemma E.6 and use $\beta = \rho_0\alpha \leq \alpha$:

$$\mathbb{E}\|\delta_{t+1}^\theta\|^2 \leq \kappa \mathbb{E}\|\delta_t^\theta\|^2 + \frac{4\alpha(r_{\max} + 2R_\theta)}{\lambda^2} G_1^{t-\tau_{\text{mix}}} + \rho_0\alpha^2 \frac{(16\tau_{\text{mix}} + 6)(r_{\max} + 2R_\theta)}{\lambda^2} + \alpha^2(88\tau_{\text{mix}} + 4)(r_{\max} + 3R_\theta)^2. \quad (32)$$

Iterate (32) from τ_{mix} to t and use $\|\delta_{\tau_{\text{mix}}}^\theta\| \leq 2R_\theta$:

$$\mathbb{E}\|\delta_t^\theta\|^2 \leq \kappa^{t-\tau_{\text{mix}}} \mathbb{E}\|\delta_{\tau_{\text{mix}}}^\theta\|^2 + \frac{4\alpha(r_{\max} + 2R_\theta)}{\lambda^2} \sum_{j=\tau_{\text{mix}}}^{t-1} \kappa^{t-1-j} G_1^{j-\tau_{\text{mix}}} + \left(\alpha^2 G_{\text{const}}\right) \sum_{j=\tau_{\text{mix}}}^{t-1} \kappa^{t-1-j},$$

where

$$G_{\text{const}} := \frac{\rho_0(16\tau_{\text{mix}} + 6)(r_{\max} + 2R_\theta)}{\lambda^2} + (88\tau_{\text{mix}} + 4)(r_{\max} + 3R_\theta)^2.$$

Use

$$\sum_{j=\tau_{\text{mix}}}^{t-1} \kappa^{t-1-j} G_1^{j-\tau_{\text{mix}}} \leq (t - \tau_{\text{mix}}) \max\{\kappa, G_1\}^{t-\tau_{\text{mix}}}, \quad \sum_{j=\tau_{\text{mix}}}^{t-1} \kappa^{t-1-j} \leq \frac{1}{1-\kappa} = \frac{1}{2\alpha\zeta}.$$

Then we complete the proof. $\square$

E.2. Proof of Lemma E.5

Proof of Lemma E.5. For brevity write $g_t := g(s_t, s'_t, w_t, \theta_t)$ and $\bar{g}_t := \bar{g}(w_t, \theta_t)$. Using $\delta_t^\theta = \delta_{t-\tau_{\text{mix}}}^\theta + (\theta_t - \theta_{t-\tau_{\text{mix}}})$,

$$\mathbb{E}[\delta_t^{\theta^\top} (g_t - \bar{g}_t)] = \mathbb{E}[\delta_{t-\tau_{\text{mix}}}^{\theta^\top} (g_t - \bar{g}_t)] + \mathbb{E}[(\theta_t - \theta_{t-\tau_{\text{mix}}})^\top (g_t - \bar{g}_t)].$$

We first note the crude drift bounds over the last τ_{mix} steps:

$$\|\theta_t - \theta_{t-\tau_{\text{mix}}}\| \leq \sum_{i=t-\tau_{\text{mix}}}^{t-1} \|\theta_{i+1} - \theta_i\| \leq \alpha \sum_{i=t-\tau_{\text{mix}}}^{t-1} \|g(s_i, s'_i, w_i, \theta_i)\| \leq 2\alpha\tau_{\text{mix}}(r_{\max} + 3R_\theta), \quad (33)$$

$$\|w_t - w_{t-\tau_{\text{mix}}}\| \leq \sum_{i=t-\tau_{\text{mix}}}^{t-1} \|w_{i+1} - w_i\| = \beta \sum_{i=t-\tau_{\text{mix}}}^{t-1} \|f(s_i, w_i)\| \leq 2\beta\tau_{\text{mix}}. \quad (34)$$

Now decompose

$$\begin{aligned}\delta_{t-\tau_{\text{mix}}}^\theta{}^\top (g_t - \bar{g}_t) &= \delta_{t-\tau_{\text{mix}}}^\theta{}^\top (g_t - g(s_t, s'_t, w_{t-\tau_{\text{mix}}}, \theta_{t-\tau_{\text{mix}}})) \\ &\quad + \delta_{t-\tau_{\text{mix}}}^\theta{}^\top (\bar{g}(w_{t-\tau_{\text{mix}}}, \theta_{t-\tau_{\text{mix}}}) - \bar{g}_t) \\ &\quad + \delta_{t-\tau_{\text{mix}}}^\theta{}^\top (g(s_t, s'_t, w_{t-\tau_{\text{mix}}}, \theta_{t-\tau_{\text{mix}}}) - \bar{g}(w_{t-\tau_{\text{mix}}}, \theta_{t-\tau_{\text{mix}}}))\end{aligned}$$

Using Lemma E.3, Lemma E.1, and (33)–(34),

$$\begin{aligned}\mathbb{E}[\delta_{t-\tau_{\text{mix}}}^\theta{}^\top (g_t - g(s_t, s'_t, w_{t-\tau_{\text{mix}}}, \theta_{t-\tau_{\text{mix}}}))] &\leq \mathbb{E}[\|\delta_{t-\tau_{\text{mix}}}^\theta\| (3\|\theta_t - \theta_{t-\tau_{\text{mix}}}\| + (r_{\max} + 2R_\theta)\|w_t - w_{t-\tau_{\text{mix}}}\|)] \\ &\leq 2R_\theta (3 \cdot 2\alpha\tau_{\text{mix}}(r_{\max} + 3R_\theta) + (r_{\max} + 2R_\theta) \cdot 2\beta\tau_{\text{mix}}) \\ &\leq 16\alpha\tau_{\text{mix}}(r_{\max} + 3R_\theta)^2,\end{aligned}$$

and the same bound holds with g replaced by $\bar{g}$.

For the remaining bias term, by the definition of $\tau_{\text{mix}}(\beta)$ we have $\sup_s \|P_{\tau_{\text{mix}}}(\cdot|s) - \mu\|_1 \leq \beta$, hence

$$\left\| \mathbb{E}[g(s_t, s'_t, w_{t-\tau_{\text{mix}}}, \theta_{t-\tau_{\text{mix}}}) - \bar{g}(w_{t-\tau_{\text{mix}}}, \theta_{t-\tau_{\text{mix}}}) \mid \mathcal{F}_{t-\tau_{\text{mix}}}] \right\| \leq \beta \cdot \sup_{s, s'} \|g(s, s', w_{t-\tau_{\text{mix}}}, \theta_{t-\tau_{\text{mix}}})\| \leq 2\beta(r_{\max} + 3R_\theta),$$

so

$$\mathbb{E}[\delta_{t-\tau_{\text{mix}}}^\theta{}^\top (g(s_t, s'_t, w_{t-\tau_{\text{mix}}}, \theta_{t-\tau_{\text{mix}}}) - \bar{g}(w_{t-\tau_{\text{mix}}}, \theta_{t-\tau_{\text{mix}}}))] \leq 2R_\theta \cdot 2\beta(r_{\max} + 3R_\theta) \leq 4\alpha(r_{\max} + 3R_\theta)^2.$$

Finally, for the second main term,

$$\mathbb{E}[(\theta_t - \theta_{t-\tau_{\text{mix}}})^\top (g_t - \bar{g}_t)] \leq \|\theta_t - \theta_{t-\tau_{\text{mix}}}\| \cdot \mathbb{E}[\|g_t\| + \|\bar{g}_t\|] \leq 2\alpha\tau_{\text{mix}}(r_{\max} + 3R_\theta) \cdot 4(r_{\max} + 3R_\theta) = 8\alpha\tau_{\text{mix}}(r_{\max} + 3R_\theta)^2.$$

Summing the pieces,

$$\mathbb{E}[\delta_t^\theta{}^\top (g_t - \bar{g}_t)] \leq (16 + 16 + 4 + 8)\alpha\tau_{\text{mix}}(r_{\max} + 3R_\theta)^2 = 44\alpha\tau_{\text{mix}}(r_{\max} + 3R_\theta)^2.$$

□

E.3. Proof of Lemma E.6

Proof of Lemma E.6. Using non-expansiveness of projection, for $t \geq 0$,

$$\begin{aligned}\mathbb{E}\|\delta_{t+1}^w\|^2 &\leq \mathbb{E}\|\delta_t^w + \beta f(s_t, w_t)\|^2 \\ &= \mathbb{E}\|\delta_t^w\|^2 + 2\beta \mathbb{E}[(\delta_t^w)^\top f(s_t, w_t)] + \beta^2 \mathbb{E}\|f(s_t, w_t)\|^2.\end{aligned}\tag{35}$$

By Lemma E.2, $\beta^2 \mathbb{E}\|f(s_t, w_t)\|^2 \leq 4\beta^2$.

For $t \geq \tau_{\text{mix}}$, write $\bar{f}(w_t) = w^* - w_t = -\delta_t^w$ and decompose

$$\mathbb{E}[(\delta_t^w)^\top f(s_t, w_t)] = \mathbb{E}[(\delta_t^w)^\top (f(s_t, w_t) - \bar{f}(w_t))] + \mathbb{E}[(\delta_t^w)^\top \bar{f}(w_t)] = \mathbb{E}[(\delta_t^w)^\top (f(s_t, w_t) - \bar{f}(w_t))] - \mathbb{E}\|\delta_t^w\|^2.$$

A standard τ_{mix} -step decomposition together with $\|w_t - w_{t-\tau_{\text{mix}}}\| \leq 2\beta\tau_{\text{mix}}$, $\|f(s_t, w_t)\| \leq 2$, $\|\delta_t^w\| \leq 2$, and the fact $\sup_s \|P_{\tau_{\text{mix}}}(\cdot|s) - \mu\|_1 \leq \beta$ yields

$$\mathbb{E}[(\delta_t^w)^\top (f(s_t, w_t) - \bar{f}(w_t))] \leq (16\tau_{\text{mix}} + 4)\beta.$$

Therefore, for $t \geq \tau_{\text{mix}}$,

$$\mathbb{E}[(\delta_t^w)^\top f(s_t, w_t)] \leq -\mathbb{E}\|\delta_t^w\|^2 + (16\tau_{\text{mix}} + 4)\beta.$$

Plugging into (35) gives

$$\mathbb{E}\|\delta_{t+1}^w\|^2 \leq (1 - 2\beta)\mathbb{E}\|\delta_t^w\|^2 + (32\tau_{\text{mix}} + 12)\beta^2.$$

Iterating from $t = \tau_{\text{mix}}$ and using $\mathbb{E}\|\delta_{\tau_{\text{mix}}}^w\|^2 \leq 4$,

$$\mathbb{E}\|\delta_t^w\|^2 \leq (1 - 2\beta)^{t-\tau_{\text{mix}}} \cdot 4 + (32\tau_{\text{mix}} + 12)\beta^2 \sum_{i=0}^{t-\tau_{\text{mix}}-1} (1 - 2\beta)^i \leq 4e^{-2\beta(t-\tau_{\text{mix}})} + (16\tau_{\text{mix}} + 6)\beta.$$

□

F. Analysis of Single Markov Chain Algorithm with decaying step-size

In this section, we provide the proof to Theorem F.1.

We now restate and prove the theorem for the single-chain algorithm with decaying step-size.

Theorem F.1. Suppose Assumption 2.1 holds. Let $a, c_0 > 0$, $\rho_0 \in (0, 1]$, and choose step-sizes

$$\alpha_t = \frac{a}{(t + c_0)^\xi}, \quad \beta_t = \rho_0 \alpha_t = \frac{\rho_0 a}{(t + c_0)^\xi}.$$

Let

$$\begin{aligned} c_{1,\lambda} &:= \eta - \frac{\lambda^2(r_{\max} + 2R_\theta)}{2}, \\ c_{2,\lambda} &:= \frac{r_{\max} + 2R_\theta}{\lambda^2}, \\ \lambda^2 &\leq \frac{2\eta}{r_{\max} + 2R_\theta}. \end{aligned}$$

Define $\Delta_T := (T + c_0)^{1-\xi} - (\tau_{\text{mix}} + c_0)^{1-\xi}$, and $G_1(T)$ as in (37), $G_2(T)$ as in (42), Γ_0 as in (43), Γ_1 as in (44). Then:

1. If $\xi \in (0, 1)$ and

$$c_0 \geq \max \left\{ \left(\frac{\xi}{ac_{1,\lambda}} \right)^{\frac{1}{1-\xi}}, \left(\frac{\xi}{\rho_0 a} \right)^{\frac{1}{1-\xi}} \right\},$$

then

$$\begin{aligned} &\mathbb{E}[\|\delta_T^\theta\|^2] \\ &\leq 4R_\theta^2 \cdot \exp\left(-\frac{ac_{1,\lambda}}{1-\xi}\Delta_T\right) + \frac{a(G_1(T) + \rho_0 c_{2,\lambda} G_2(T))}{c_{1,\lambda}(T + c_0)^\xi} \\ &\quad + \frac{4ac_{2,\lambda}}{1-\xi} \exp\left(\frac{ac_{1,\lambda}}{c_0^\xi}\right) \Delta_T \exp\left(-\frac{a}{1-\xi}\rho_{\min}\Delta_T\right). \end{aligned}$$

where $\rho_{\min} = \min\{c_{1,\lambda}, 2\rho_0\}$

2. If $\xi = 1$, $a\eta < 1$, and $\lambda^2 \geq \frac{2\eta - 4\rho_0}{r_{\max} + 2R_\theta}$, then

$$\mathbb{E}[\|\delta_T^\theta\|^2] \leq 4R_\theta^2 \cdot \left(\frac{\tau_{\text{mix}} + c_0}{T + c_0} \right)^{ac_{1,\lambda}} + \frac{\Gamma_1 + \Gamma_0 G_1(T)}{(T + c_0)^{ac_{1,\lambda}}},$$

where Γ_0 and Γ_1 are defined in Appendix F.

Proof of Theorem F.1. As before, the dynamic of θ_t satisfies

$$\begin{aligned} \mathbb{E}\|\delta_{t+1}^\theta\|^2 &\leq \mathbb{E}\|\delta_t^\theta\|^2 + 2\alpha_t \mathbb{E}\left[\delta_t^{\theta^\top} g(s_t, s'_t, w_t, \theta_t)\right] + \alpha_t^2 \mathbb{E}[\|g(s_t, s'_t, w_t, \theta_t)\|^2] \\ &= \mathbb{E}\|\delta_t^\theta\|^2 + \underbrace{2\alpha_t \mathbb{E}\left[\delta_t^{\theta^\top} (g(s_t, s'_t, w_t, \theta_t) - \bar{g}(w_t, \theta_t))\right]}_{I_1} \\ &\quad + \underbrace{2\alpha_t \mathbb{E}\left[\delta_t^{\theta^\top} \bar{g}(w_t, \theta_t)\right]}_{I_2} + \underbrace{\alpha_t^2 \mathbb{E}[\|g(s_t, s'_t, w_t, \theta_t)\|^2]}_{I_3}, \end{aligned}$$

where

$$\bar{g}(w_t, \theta_t) := \mathbb{E}_{s \sim \mu} g(s, s', w_t, \theta_t) = \Phi^\top D(R + P\Phi\theta_t - \Phi\theta_t) - \mu^\top (R + \Phi\theta_t)w_t.$$

Term I_1 : We use the following Markov-noise bound.

Lemma F.2. Suppose $t \geq \tau_{\text{mix}}$. Then

$$\mathbb{E} \left[\delta_t^\theta{}^\top (g(s_t, s'_t, w_t, \theta_t) - \bar{g}(w_t, \theta_t)) \right] \leq \rho_1 (r_{\max} + 3R_\theta)^2 \sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i,$$

where $\rho_1 := 28 + 8\rho_0$.

The proof is deferred to Section F.1. Using Lemma F.2,

$$I_1 \leq 2\rho_1 \alpha_t (r_{\max} + 3R_\theta)^2 \sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i.$$

Term I_2 : By the same argument as Eq. (29) in Appendix E, we have

$$I_2 \leq -2\alpha_t c_{1,\lambda} \mathbb{E} \|\delta_t^\theta\|^2 + \alpha_t c_{2,\lambda} \mathbb{E} \|\delta_t^w\|^2,$$

where $c_{1,\lambda} = \eta - \lambda^2(r_{\max} + 2R_\theta)/2$ and $c_{2,\lambda} = (r_{\max} + 2R_\theta)/\lambda^2$.

Term I_3 : By Lemma E.2,

$$I_3 \leq 4\alpha_t^2 (r_{\max} + 3R_\theta)^2.$$

Combining the three terms yields, for $t \geq \tau_{\text{mix}}$,

$$\begin{aligned} \mathbb{E} \|\delta_{t+1}^\theta\|^2 &\leq (1 - 2\alpha_t c_{1,\lambda}) \mathbb{E} \|\delta_t^\theta\|^2 + \alpha_t c_{2,\lambda} \mathbb{E} \|\delta_t^w\|^2 + 4\alpha_t^2 (r_{\max} + 3R_\theta)^2 \\ &\quad + 2\rho_1 \alpha_t (r_{\max} + 3R_\theta)^2 \sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i. \end{aligned}$$

Using Lemma D.2, $\sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i \leq 2L_1 \alpha_t (\log(1/\alpha_t) + 1)$, we further obtain

$$\mathbb{E} \|\delta_{t+1}^\theta\|^2 \leq (1 - 2\alpha_t c_{1,\lambda}) \mathbb{E} \|\delta_t^\theta\|^2 + \alpha_t c_{2,\lambda} \mathbb{E} \|\delta_t^w\|^2 + G_1(T) \alpha_t^2, \quad (36)$$

where we used the monotonicity of α_t to bound $\log(1/\alpha_t) \leq \log(1/\alpha_T)$ for all $t \leq T$, and defined

$$G_1(T) := \left(4\rho_1 L_1 (\log(T + c_0) - \log a + 1) + 4 \right) (r_{\max} + 3R_\theta)^2. \quad (37)$$

For convenience in the iteration, we relax the contraction factor as

$$(1 - 2\alpha_t c_{1,\lambda}) \leq (1 - \alpha_t c_{1,\lambda}) \quad (\text{since } \alpha_t c_{1,\lambda} \geq 0),$$

so from (36),

$$\mathbb{E} \|\delta_{t+1}^\theta\|^2 \leq (1 - \alpha_t c_{1,\lambda}) \mathbb{E} \|\delta_t^\theta\|^2 + \alpha_t c_{2,\lambda} \mathbb{E} \|\delta_t^w\|^2 + G_1(T) \alpha_t^2. \quad (38)$$

Iterating (38) from τ_{mix} to T gives

$$\begin{aligned} \mathbb{E} \|\delta_T^\theta\|^2 &\leq \underbrace{\prod_{i=\tau_{\text{mix}}}^{T-1} (1 - c_{1,\lambda} \alpha_i)}_{I'_1} \mathbb{E} \|\delta_{\tau_{\text{mix}}}^\theta\|^2 + \sum_{j=\tau_{\text{mix}}}^{T-1} \left(\prod_{i=j+1}^{T-1} (1 - c_{1,\lambda} \alpha_i) \right) \alpha_j c_{2,\lambda} \mathbb{E} \|\delta_j^w\|^2 \\ &\quad + G_1(T) \sum_{j=\tau_{\text{mix}}}^{T-1} \alpha_j^2 \left(\prod_{i=j+1}^{T-1} (1 - c_{1,\lambda} \alpha_i) \right). \end{aligned} \quad (39)$$

We first bound I'_1 . Using $1 - x \leq e^{-x}$,

$$I'_1 \leq \exp\left(-c_{1,\lambda} \sum_{i=\tau_{\text{mix}}}^{T-1} \alpha_i\right) = \exp\left(-ac_{1,\lambda} \sum_{i=\tau_{\text{mix}}}^{T-1} \frac{1}{(i+c_0)^\xi}\right) \leq \exp\left(-ac_{1,\lambda} \int_{\tau_{\text{mix}}}^T \frac{dx}{(x+c_0)^\xi}\right), \quad (40)$$

hence

$$I'_1 \leq \begin{cases} \left(\frac{\tau_{\text{mix}}+c_0}{T+c_0}\right)^{ac_{1,\lambda}}, & \xi = 1, \\ \exp\left(-\frac{ac_{1,\lambda}}{1-\xi} \left((T+c_0)^{1-\xi} - (\tau_{\text{mix}}+c_0)^{1-\xi}\right)\right), & \xi \in (0, 1). \end{cases} \quad (41)$$

Next we control $\mathbb{E}\|\delta_j^w\|^2$.

Lemma F.3. *Let*

$$G_2(T) := 64L_1(\log(T+c_0) - \log(\rho_0 a) + 1) + 12. \quad (42)$$

Then for all $t \in [\tau_{\text{mix}}, T]$:

1. *If $\xi \in (0, 1)$ and $c_0 \geq \left(\frac{\xi}{\rho_0 a}\right)^{\frac{1}{1-\xi}}$, then*

$$\mathbb{E}\|\delta_t^w\|^2 \leq 4 \exp\left(-\frac{2\rho_0 a}{1-\xi} \left((t+c_0)^{1-\xi} - (\tau_{\text{mix}}+c_0)^{1-\xi}\right)\right) + G_2(T) \cdot \frac{\rho_0 a}{(t+c_0)^\xi}.$$

2. *If $\xi = 1$, then:*

- (a) *If $\rho_0 a \in (0, 1/2)$, then*

$$\mathbb{E}\|\delta_t^w\|^2 \leq 4 \left(\frac{\tau_{\text{mix}}+c_0}{t+c_0}\right)^{2\rho_0 a} + G_2(T) \cdot \frac{4\rho_0^2 a^2}{(1-2\rho_0 a)(t+c_0)^{2\rho_0 a}}.$$

- (b) *If $\rho_0 a = 1/2$, then*

$$\mathbb{E}\|\delta_t^w\|^2 \leq 4 \left(\frac{\tau_{\text{mix}}+c_0}{t+c_0}\right)^{2\rho_0 a} + G_2(T) \cdot \frac{4\rho_0^2 a^2 \log(t+c_0)}{t+c_0}.$$

- (c) *If $\rho_0 a \in (1/2, \infty)$, then*

$$\mathbb{E}\|\delta_t^w\|^2 \leq 4 \left(\frac{\tau_{\text{mix}}+c_0}{t+c_0}\right)^{2\rho_0 a} + G_2(T) \cdot \frac{4e\rho_0^2 a^2}{(2\rho_0 a - 1)(t+c_0)}.$$

The proof is deferred to Section F.2.

We now split into two cases.

Case 1: $\xi = 1$. Using (39) and (41) (with $\xi = 1$),

$$\begin{aligned} \mathbb{E}\|\delta_T^\theta\|^2 &\leq \underbrace{\left(\frac{\tau_{\text{mix}}+c_0}{T+c_0}\right)^{ac_{1,\lambda}} \mathbb{E}\|\delta_{\tau_{\text{mix}}}^\theta\|^2 + \sum_{j=\tau_{\text{mix}}}^{T-1} \left(\frac{j+1+c_0}{T+c_0}\right)^{ac_{1,\lambda}} \frac{ac_{2,\lambda}}{j+c_0} \mathbb{E}\|\delta_j^w\|^2}_{I'_2} \\ &\quad + \underbrace{G_1(T) \sum_{j=\tau_{\text{mix}}}^{T-1} \frac{a^2}{(j+c_0)^2} \left(\frac{j+1+c_0}{T+c_0}\right)^{ac_{1,\lambda}}}_{I'_3}. \end{aligned}$$

Bound on I'_3 . Since $(j+1+c_0)/(j+c_0) \leq 1+1/c_0 \leq 2$ when $c_0 \geq 1$,

$$\begin{aligned} I'_3 &= \frac{a^2}{(T+c_0)^{ac_{1,\lambda}}} \sum_{j=\tau_{\text{mix}}}^{T-1} \frac{(j+1+c_0)^{ac_{1,\lambda}}}{(j+c_0)^2} \leq \frac{4a^2}{(T+c_0)^{ac_{1,\lambda}}} \sum_{j=\tau_{\text{mix}}}^{T-1} (j+1+c_0)^{ac_{1,\lambda}-2} \\ &\leq \frac{4a^2}{(1-ac_{1,\lambda})(T+c_0)^{ac_{1,\lambda}}} \quad (ac_{1,\lambda} \leq a\eta < 1). \end{aligned}$$

Define

$$\Gamma_0 := \frac{4a^2}{1-ac_{1,\lambda}}. \quad (43)$$

Then $I'_3 \leq \Gamma_0/(T+c_0)^{ac_{1,\lambda}}$.

Bound on I'_2 . Apply Lemma F.3 (case $\xi = 1$). If $\rho_0 a \in (0, 1/2)$, then

$$\begin{aligned} I'_2 &\leq \sum_{j=\tau_{\text{mix}}}^{T-1} \left(\frac{j+1+c_0}{T+c_0} \right)^{ac_{1,\lambda}} \frac{ac_{2,\lambda}}{j+c_0} \left[4 \left(\frac{\tau_{\text{mix}}+c_0}{j+c_0} \right)^{2\rho_0 a} + G_2(T) \cdot \frac{4\rho_0^2 a^2}{(1-2\rho_0 a)(j+c_0)^{2\rho_0 a}} \right] \\ &= \frac{ac_{2,\lambda}}{(T+c_0)^{ac_{1,\lambda}}} \left(4(\tau_{\text{mix}}+c_0)^{2\rho_0 a} + \frac{4\rho_0^2 a^2}{1-2\rho_0 a} G_2(T) \right) \sum_{j=\tau_{\text{mix}}}^{T-1} \frac{(j+1+c_0)^{ac_{1,\lambda}}}{(j+c_0)^{1+2\rho_0 a}} \\ &\leq \frac{2^{ac_{1,\lambda}} ac_{2,\lambda}}{(T+c_0)^{ac_{1,\lambda}}} \left(4(\tau_{\text{mix}}+c_0)^{2\rho_0 a} + \frac{4\rho_0^2 a^2}{1-2\rho_0 a} G_2(T) \right) \sum_{j=\tau_{\text{mix}}}^{T-1} (j+c_0)^{ac_{1,\lambda}-1-2\rho_0 a} \\ &\leq \frac{2^{ac_{1,\lambda}} ac_{2,\lambda}}{(T+c_0)^{ac_{1,\lambda}}} \left(4(\tau_{\text{mix}}+c_0)^{2\rho_0 a} + \frac{4\rho_0^2 a^2}{1-2\rho_0 a} G_2(T) \right) \cdot \frac{1}{2\rho_0 a - ac_{1,\lambda}}, \end{aligned}$$

where the last step uses $ac_{1,\lambda} < 2\rho_0 a$, which follows from $\lambda^2 \geq \frac{2\eta-4\rho_0}{r_{\text{max}}+2R_\theta}$ (equivalently, $c_{1,\lambda} \leq 2\rho_0$). The cases $\rho_0 a = 1/2$ and $\rho_0 a > 1/2$ are handled similarly, yielding the same $(T+c_0)^{-ac_{1,\lambda}}$ scaling. Define

$$\Gamma_1 := \begin{cases} \frac{2^{ac_{1,\lambda}} ac_{2,\lambda}}{2\rho_0 a - ac_{1,\lambda}} \left(4(\tau_{\text{mix}}+c_0)^{2\rho_0 a} + \frac{4\rho_0^2 a^2}{1-2\rho_0 a} G_2(T) \right), & \rho_0 a \in (0, 1/2), \\ \frac{4ac_{2,\lambda}}{1-ac_{1,\lambda}} (4(\tau_{\text{mix}}+c_0)^{2\rho_0 a} + 4\rho_0^2 a^2 \log(T+c_0) G_2(T)), & \rho_0 a = 1/2, \\ \frac{4ac_{2,\lambda}}{1-ac_{1,\lambda}} \left(4(\tau_{\text{mix}}+c_0)^{2\rho_0 a} + \frac{4e\rho_0^2 a^2}{2\rho_0 a - 1} G_2(T) \right), & \rho_0 a \in (1/2, \infty). \end{cases} \quad (44)$$

Then $I'_2 \leq \Gamma_1/(T+c_0)^{ac_{1,\lambda}}$.

Putting the bounds together and using $\mathbb{E}\|\delta_{\tau_{\text{mix}}}^\theta\|^2 \leq 4R_\theta^2$, we obtain

$$\mathbb{E}\|\delta_T^\theta\|^2 \leq 4R_\theta^2 \left(\frac{\tau_{\text{mix}}+c_0}{T+c_0} \right)^{ac_{1,\lambda}} + \frac{\Gamma_1 + \Gamma_0 G_1(T)}{(T+c_0)^{ac_{1,\lambda}}}.$$

Case 2: $\xi \in (0, 1)$. Recall from Lemma F.3 (case $\xi \in (0, 1)$) that for all $t \in [\tau_{\text{mix}}, T]$,

$$\mathbb{E}\|\delta_t^w\|^2 \leq 4 \exp\left(-A\left((t+c_0)^{1-\xi} - (\tau_{\text{mix}}+c_0)^{1-\xi}\right)\right) + \rho_0 G_2(T) \alpha_t, \quad A := \frac{2\rho_0 a}{1-\xi}. \quad (45)$$

Define

$$\Delta_T := (T+c_0)^{1-\xi} - (\tau_{\text{mix}}+c_0)^{1-\xi}, \quad S_T := \sum_{j=\tau_{\text{mix}}}^{T-1} \alpha_j^2 \left(\prod_{i=j+1}^{T-1} (1-c_{1,\lambda}\alpha_i) \right).$$

We also define the exponential-part sum

$$S_T^{\text{exp}} := \sum_{j=\tau_{\text{mix}}}^{T-1} \alpha_j \left(\prod_{i=j+1}^{T-1} (1-c_{1,\lambda}\alpha_i) \right) \exp\left(-A\left((j+c_0)^{1-\xi} - (\tau_{\text{mix}}+c_0)^{1-\xi}\right)\right). \quad (46)$$

Lemma F.4. Let $\xi \in (0, 1)$. Then

$$S_T^{\text{exp}} \leq \exp\left(\frac{ac_{1,\lambda}}{c_0^\xi}\right) \cdot \frac{a}{1-\xi} \Delta_T \exp\left(-\frac{a}{1-\xi} \min\{c_{1,\lambda}, 2\rho_0\} \Delta_T\right). \quad (47)$$

The proof is deferred to Section F.3.

Now plug (45) into the second term of (39). Using $c_{2,\lambda}\alpha_j \rho_0 G_2(T)\alpha_j = \rho_0 c_{2,\lambda} G_2(T)\alpha_j^2$, we obtain

$$\begin{aligned} & \sum_{j=\tau_{\text{mix}}}^{T-1} \left(\prod_{i=j+1}^{T-1} (1 - c_{1,\lambda}\alpha_i) \right) \alpha_j c_{2,\lambda} \mathbb{E} \|\delta_j^w\|^2 \\ & \leq 4c_{2,\lambda} S_T^{\text{exp}} + \rho_0 c_{2,\lambda} G_2(T) S_T. \end{aligned}$$

Therefore, by (39) and (41) (with $\xi \in (0, 1)$),

$$\mathbb{E} \|\delta_T^\theta\|^2 \leq \exp\left(-\frac{ac_{1,\lambda}}{1-\xi} \Delta_T\right) \mathbb{E} \|\delta_{\tau_{\text{mix}}}^\theta\|^2 + (G_1(T) + \rho_0 c_{2,\lambda} G_2(T)) S_T + 4c_{2,\lambda} S_T^{\text{exp}}.$$

By Lemma D.5 and the condition $c_0 \geq \left(\frac{\xi}{ac_{1,\lambda}}\right)^{\frac{1}{1-\xi}}$, we have

$$S_T \leq \frac{a}{c_{1,\lambda}(T + c_0)^\xi}.$$

Applying Lemma F.4 (i.e., (47)) and using $\mathbb{E} \|\delta_{\tau_{\text{mix}}}^\theta\|^2 \leq 4R_\theta^2$ yields the following bound for Case 2:

$$\begin{aligned} \mathbb{E} \|\delta_T^\theta\|^2 & \leq 4R_\theta^2 \exp\left(-\frac{ac_{1,\lambda}}{1-\xi} \Delta_T\right) + \frac{a(G_1(T) + \rho_0 c_{2,\lambda} G_2(T))}{c_{1,\lambda}(T + c_0)^\xi} \\ & \quad + 4c_{2,\lambda} \exp\left(\frac{ac_{1,\lambda}}{c_0^\xi}\right) \cdot \frac{a}{1-\xi} \Delta_T \exp\left(-\frac{a}{1-\xi} \min\{c_{1,\lambda}, 2\rho_0\} \Delta_T\right). \end{aligned}$$

□

F.1. Proof of Lemma F.2

Proof of Lemma F.2. The iteration of w suggests that

$$\begin{aligned} \|w_t - w_{t-\tau_{\text{mix}}}\| & \leq \sum_{i=t-\tau_{\text{mix}}}^{t-1} \|w_{i+1} - w_i\| \leq \sum_{i=t-\tau_{\text{mix}}}^{t-1} \beta_i \|f(s_i, w_i)\| \\ & \leq 2 \sum_{i=t-\tau_{\text{mix}}}^{t-1} \beta_i. \end{aligned} \quad (48)$$

For simplicity, denote $g_t(w_t, \theta_t) := g(s_t, s'_t, w_t, \theta_t)$. First, by telescoping and Lemma E.2,

$$\begin{aligned} \|\theta_t - \theta_{t-\tau_{\text{mix}}}\| & \leq \sum_{i=t-\tau_{\text{mix}}}^{t-1} \|\theta_{i+1} - \theta_i\| \leq \sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i \|g(s_i, s'_i, w_i, \theta_i)\| \\ & \leq (2r_{\text{max}} + 6R_\theta) \sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i = 2(r_{\text{max}} + 3R_\theta) \sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i. \end{aligned} \quad (49)$$

Decompose

$$\begin{aligned} & \mathbb{E} \left[\delta_t^{\theta^\top} (g_t(w_t, \theta_t) - \bar{g}(w_t, \theta_t)) \right] \\ & = \mathbb{E} \left[\delta_{t-\tau_{\text{mix}}}^{\theta^\top} (g_t(w_t, \theta_t) - g_t(w_{t-\tau_{\text{mix}}}, \theta_{t-\tau_{\text{mix}}})) \right] + \mathbb{E} \left[\delta_{t-\tau_{\text{mix}}}^{\theta^\top} (\bar{g}(w_{t-\tau_{\text{mix}}}, \theta_{t-\tau_{\text{mix}}}) - \bar{g}(w_t, \theta_t)) \right] \\ & \quad + \mathbb{E} \left[\delta_{t-\tau_{\text{mix}}}^{\theta^\top} (g_t(w_{t-\tau_{\text{mix}}}, \theta_{t-\tau_{\text{mix}}}) - \bar{g}(w_{t-\tau_{\text{mix}}}, \theta_{t-\tau_{\text{mix}}})) \right] + \mathbb{E} \left[(\theta_t - \theta_{t-\tau_{\text{mix}}})^\top (g_t(w_t, \theta_t) - \bar{g}(w_t, \theta_t)) \right] \\ & =: I_1 + I_2 + I_3 + I_4. \end{aligned}$$

Terms I_1 and I_2 . By Lemma E.3, $\|g(\cdot, \cdot, w, \theta_1) - g(\cdot, \cdot, w, \theta_2)\| \leq 2\|\theta_1 - \theta_2\|$ and $\|g(\cdot, \cdot, w_1, \theta) - g(\cdot, \cdot, w_2, \theta)\| \leq (r_{\max} + 2R_\theta)\|w_1 - w_2\|$. Also, by (48) and $\beta_i = \rho_0\alpha_i$,

$$\|w_t - w_{t-\tau_{\text{mix}}}\| \leq 2 \sum_{i=t-\tau_{\text{mix}}}^{t-1} \beta_i = 2\rho_0 \sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i.$$

Thus, using $\|\delta_{t-\tau_{\text{mix}}}^\theta\| \leq 2R_\theta$,

$$\begin{aligned} I_1 &\leq \mathbb{E}[\|\delta_{t-\tau_{\text{mix}}}^\theta\| \cdot 2\|\theta_t - \theta_{t-\tau_{\text{mix}}}\|] + \mathbb{E}[\|\delta_{t-\tau_{\text{mix}}}^\theta\| \cdot (r_{\max} + 2R_\theta)\|w_t - w_{t-\tau_{\text{mix}}}\|] \\ &\leq 2R_\theta \cdot 4(r_{\max} + 3R_\theta) \sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i + 2R_\theta \cdot (r_{\max} + 2R_\theta) \cdot 2\rho_0 \sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i \\ &\leq (8 + 4\rho_0)R_\theta(r_{\max} + 3R_\theta) \sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i. \end{aligned}$$

The same bound applies to I_2 since $\bar{g}$ is the expectation of g .

Term I_3 . Conditioning on $\mathcal{F}_{t-\tau_{\text{mix}}}$ and using the mixing bound $\|P_{\tau_{\text{mix}}}(\cdot | s) - \mu\|_1 \leq C\beta^{\tau_{\text{mix}}} \leq \beta_T \leq \alpha_t$ (for $t \leq T$),

$$\begin{aligned} I_3 &\leq \mathbb{E}\left[\|\delta_{t-\tau_{\text{mix}}}^\theta\| \cdot \|P_{\tau_{\text{mix}}}(\cdot | s_{t-\tau_{\text{mix}}}) - \mu\|_1 \cdot \sup_{s, s'} \|g(s, s', w_{t-\tau_{\text{mix}}}, \theta_{t-\tau_{\text{mix}}})\|\right] \\ &\leq 2R_\theta \cdot \alpha_t \cdot (2r_{\max} + 6R_\theta) = 4\alpha_t R_\theta(r_{\max} + 3R_\theta). \end{aligned}$$

Term I_4 . Using (49) and Lemma E.2,

$$\begin{aligned} I_4 &\leq \mathbb{E}[\|\theta_t - \theta_{t-\tau_{\text{mix}}}\| \cdot (\|g_t(w_t, \theta_t)\| + \|\bar{g}(w_t, \theta_t)\|)] \\ &\leq 2(r_{\max} + 3R_\theta) \sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i \cdot 2 \cdot (2r_{\max} + 6R_\theta) \\ &= 8(r_{\max} + 3R_\theta)^2 \sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i. \end{aligned}$$

Summing the four terms and using $R_\theta \leq r_{\max} + 3R_\theta$ and $\alpha_t \leq \sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i$,

$$\begin{aligned} \mathbb{E}\left[\delta_t^{\theta^\top} (g_t(w_t, \theta_t) - \bar{g}(w_t, \theta_t))\right] &\leq (16 + 8\rho_0)R_\theta(r_{\max} + 3R_\theta) \sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i + 4\alpha_t R_\theta(r_{\max} + 3R_\theta) \\ &\quad + 8(r_{\max} + 3R_\theta)^2 \sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i \\ &\leq (28 + 8\rho_0)(r_{\max} + 3R_\theta)^2 \sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha_i, \end{aligned}$$

which is the desired result with $\rho_1 = 28 + 8\rho_0$. □

F.2. Proof of Lemma F.3

Proof of Lemma F.3. The iterates of w suggests that

$$\mathbb{E}\|\delta_{t+1}^w\|^2 \leq \mathbb{E}\|\delta_t^w\|^2 + 2\beta_t \mathbb{E}[(\delta_t^w)^\top f(s_t, w_t)] + \beta_t^2 \mathbb{E}\|f(s_t, w_t)\|^2.$$

By Lemma E.2, $\mathbb{E}\|f(s_t, w_t)\|^2 \leq 4$, hence the third term is $\leq 4\beta_t^2$.

For the second term, decompose as in the constant-step proof (Appendix E):

$$\mathbb{E}[(\delta_t^w)^\top f(s_t, w_t)] = I_1 + I_2 + I_3 + I_4 + I_5,$$

where the five terms are exactly those in Section E.3 with β replaced by β_t (and sums over the last τ_{mix} indices). Using the same arguments but keeping the time-varying step-size, we obtain:

$$I_1 \leq 4 \sum_{i=t-\tau_{\text{mix}}}^{t-1} \beta_i,$$

$$I_2 \leq 4\beta_t,$$

$$I_3 \leq 4 \sum_{i=t-\tau_{\text{mix}}}^{t-1} \beta_i,$$

$$I_4 \leq 8 \sum_{i=t-\tau_{\text{mix}}}^{t-1} \beta_i,$$

$$I_5 = -\mathbb{E}\|\delta_t^w\|^2,$$

and therefore

$$\mathbb{E}[(\delta_t^w)^\top f(s_t, w_t)] \leq -\mathbb{E}\|\delta_t^w\|^2 + 16 \sum_{i=t-\tau_{\text{mix}}}^{t-1} \beta_i + 4\beta_t.$$

Plugging back yields

$$\mathbb{E}\|\delta_{t+1}^w\|^2 \leq (1 - 2\beta_t)\mathbb{E}\|\delta_t^w\|^2 + 32\beta_t\hat{\beta}_t + 12\beta_t^2, \quad \hat{\beta}_t := \sum_{i=t-\tau_{\text{mix}}}^{t-1} \beta_i. \quad (50)$$

As in Lemma D.2, one can show that

$$\hat{\beta}_t \leq 2L_1 \beta_t (\log(1/\beta_t) + 1), \quad t \leq T,$$

where $L_1 = \max\left\{1, \frac{\log C+1}{\log(1/\beta)}\right\}$. Substituting this bound into (50) and using $\log(1/\beta_t) \leq \log(1/\beta_T) = \log(T + c_0) - \log(\rho_0 a)$ for $t \leq T$ gives

$$\mathbb{E}\|\delta_{t+1}^w\|^2 \leq (1 - 2\beta_t)\mathbb{E}\|\delta_t^w\|^2 + G_2(T) \beta_t^2,$$

with $G_2(T)$ as defined in Lemma F.3. Unrolling the recursion yields

$$\mathbb{E}\|\delta_t^w\|^2 \leq \underbrace{\prod_{i=\tau_{\text{mix}}}^{t-1} (1 - 2\beta_i)}_{J_1} \mathbb{E}\|\delta_{\tau_{\text{mix}}}^w\|^2 + G_2(T) \underbrace{\sum_{i=\tau_{\text{mix}}}^{t-1} \beta_i^2 \prod_{j=i+1}^{t-1} (1 - 2\beta_j)}_{J_2}.$$

The term J_1 is bounded by $J_1 \leq \exp(-2\sum_{i=\tau_{\text{mix}}}^{t-1} \beta_i)$, which gives the stated exponential bounds for $\xi \in (0, 1)$ and the stated power bounds for $\xi = 1$. The term J_2 is bounded by the standard summation estimates, yielding the three sub-cases for $\xi = 1$ and, for $\xi \in (0, 1)$ under $c_0 \geq \left(\frac{\xi}{\rho_0 a}\right)^{\frac{1}{1-\xi}}$, the bound $J_2 \leq \frac{\rho_0 a}{(t+c_0)^\xi}$ via Lemma D.5. Finally, Lemma E.1 gives $\mathbb{E}\|\delta_{\tau_{\text{mix}}}^w\|^2 \leq 4$, completing the proof. $\square$

F.3. Proof of Lemma F.4

Proof of Lemma F.4. Let $u(t) := (t + c_0)^{1-\xi}$ and $\Delta_T = u(T) - u(\tau_{\text{mix}})$. For any $j \in [\tau_{\text{mix}}, T - 1]$, using $1 - x \leq e^{-x}$ and $\alpha_i \geq 0$,

$$\prod_{i=j+1}^{T-1} (1 - c_{1,\lambda} \alpha_i) \leq \exp\left(-c_{1,\lambda} \sum_{i=j+1}^{T-1} \alpha_i\right).$$

Since $x \mapsto (x + c_0)^{-\xi}$ is decreasing,

$$\sum_{i=j+1}^{T-1} \alpha_i = a \sum_{i=j+1}^{T-1} \frac{1}{(i + c_0)^\xi} \geq a \int_{j+1}^T \frac{dx}{(x + c_0)^\xi} = \frac{a}{1 - \xi} (u(T) - u(j + 1)).$$

Moreover,

$$u(j + 1) = u(j) + (u(j + 1) - u(j)) \leq u(j) + (1 - \xi)(j + c_0)^{-\xi} = u(j) + \frac{1 - \xi}{a} \alpha_j,$$

hence

$$u(T) - u(j + 1) \geq u(T) - u(j) - \frac{1 - \xi}{a} \alpha_j.$$

Combining the above gives

$$\prod_{i=j+1}^{T-1} (1 - c_{1,\lambda} \alpha_i) \leq \exp\left(-\frac{ac_{1,\lambda}}{1 - \xi} (u(T) - u(j)) + c_{1,\lambda} \alpha_j\right) \leq \exp\left(\frac{ac_{1,\lambda}}{c_0^\xi}\right) \cdot \exp\left(-\frac{ac_{1,\lambda}}{1 - \xi} (u(T) - u(j))\right),$$

where we used $\alpha_j \leq \alpha_0 = a/c_0^\xi$.

Therefore, each summand in (46) satisfies

$$\begin{aligned} & \alpha_j \left(\prod_{i=j+1}^{T-1} (1 - c_{1,\lambda} \alpha_i) \right) \exp(-A(u(j) - u(\tau_{\text{mix}}))) \\ & \leq \exp\left(\frac{ac_{1,\lambda}}{c_0^\xi}\right) \cdot \alpha_j \exp\left(-\frac{ac_{1,\lambda}}{1 - \xi} (u(T) - u(j)) - \frac{2\rho_0 a}{1 - \xi} (u(j) - u(\tau_{\text{mix}}))\right). \end{aligned}$$

Since $u(j) \in [u(\tau_{\text{mix}}), u(T)]$, we have for all such j ,

$$c_{1,\lambda} (u(T) - u(j)) + 2\rho_0 (u(j) - u(\tau_{\text{mix}})) \geq \min\{c_{1,\lambda}, 2\rho_0\} \cdot (u(T) - u(\tau_{\text{mix}})) = \min\{c_{1,\lambda}, 2\rho_0\} \Delta_T,$$

and hence the exponential factor is at most $\exp\left(-\frac{a}{1 - \xi} \min\{c_{1,\lambda}, 2\rho_0\} \Delta_T\right)$. Summing and using

$$\sum_{j=\tau_{\text{mix}}}^{T-1} \alpha_j \leq a \int_{\tau_{\text{mix}}}^T \frac{dx}{(x + c_0)^\xi} = \frac{a}{1 - \xi} \Delta_T$$

yields (47). □

G. Numerical Results

In this section, we provide numerical results on our proposed algorithms. We consider tasks from OpenAI Gym (Towers et al., 2024), MO-Gymnasium (Felten et al., 2023) and Gridworlds. All of the tasks share the following settings:

- Policy: the policy π is learned by using the Q -learning algorithm.
- Continuous task: some tasks are episodic. To make it a continuous task, the agent will proceed to the starting point after reaching any terminal state with a reward of 0.
- Ergodic MDP: to ensure the induced Markov chain under this policy is ergodic, we modify the transition matrix as follows: for each state s , we examine the s -th row of the transition matrix P . If the original transition probabilities under the learned policy contains entries that are 0 (e.g., $P(\cdot|s) = [0, 0, 1, 0, 0]$), we redistribute a small portion of the probability mass to all previously unreachable states. Specifically, we assign a small probability ϵ equally among the zero-probability entries, and reduce the original non-zero entries accordingly to ensure the row still sums to 1. For instance, when $\epsilon = 0.1$, the row above becomes $[0.025, 0.025, 0.9, 0.025, 0.025]$. This adjustment is applied to all rows of P , ensuring that every state has a non-zero probability of transitioning to every other state, thus enforcing ergodicity. The choice of ϵ for different tasks can be found in Table 2.
- Reward function: the reward function R is a vector whose s -th row $R(s)$ is defined by the deterministic one-step reward of performing the policy π in state s .
- Stationary distribution: the stationary distribution μ is obtained by solving $\mu^T = \mu^T P$.
- Averaged reward: the average reward function g is defined as $g = \mu^T R$.
- Feature matrix: the feature matrix Φ is defined to be a $|S| \times d$ matrix. We first generate a matrix $\tilde{\Phi} \in \mathbb{R}^{|S| \times (d-2)}$, where each element is drawn from the Bernoulli distribution with success probability $p = 0.5$. Then, we construct Φ by stacking the all-ones vector e and the true value function W^* as columns into the matrix $\tilde{\Phi}$, i.e., $\Phi = [\tilde{\Phi}, e, W^*]$. The process is repeated until the feature matrix has full column rank. We further normalize the features to ensure $\|\phi(s)\| \leq 1$ for all $s \in S$.

We plot the value function error $\|W_t - W^*\|$ for four algorithms. Two of them are proposed in this paper, namely the Double-Chain and Single-Chain algorithms. As baselines, we include representative prior methods from Table 1. We note that (Tsitsiklis & Van Roy, 1999), (Blaser & Zhang, 2024), (Haque & Maguluri, 2024), (Haque & Maguluri, 2024), and (Chen et al., 2025) use essentially the same update rule, while (Li et al., 2024) incorporates variance reduction and therefore converges much more slowly.

Here, for our algorithms, we compute W^* as the unique solution to

$$W^* + ge = PW^* + R, \quad \mu^T W^* = 0.$$

For the other algorithms, we measure the error in value-function space modulo additive constants, since in the average-reward setting the relative value function is only defined up to a constant shift. Specifically, rather than comparing W_t with a single representative W^* , we measure the Euclidean distance from W_t to the affine space

$$\mathcal{W}^* := \{W^* + ce : c \in \mathbb{R}\},$$

where e denotes the all-one vector. Equivalently, we remove from $W_t - W^*$ its component in the constant direction e , and retain only the orthogonal component. The resulting error metric is

$$\text{dist}(W_t, \mathcal{W}^*) = \left\| \left(I - \frac{ee^T}{e^T e} \right) (W_t - W^*) \right\|_2.$$

Each curve is averaged over three independent runs. The step-size schedule α_t and the total number of iterations T are reported in Table 2.

Table 2. Parameters in different tasks

Task	d	α_t	T	ϵ
Random Walk (50)	5	$150/(t + 1000)$	150000	N/A
Random Walk (100)	20	$150/(t + 1000)$	150000	N/A
Random Walk (1000)	100	$150/(t + 1000)$	150000	N/A
Frozen Lake	10	$150/(t + 1000)$	150000	0.1
Cliff Walking	20	$150/(t + 1000)$	150000	0.1
Taxi	100	$150/(t + 1000)$	150000	0.5
Grid World (5x5)	10	$200/(t + 1000)$	200000	0.2
Grid World (10x10)	40	$300/(t + 1000)$	300000	0.2
Grid World (2x11)	10	$300/(t + 1000)$	300000	0.2
Deep sea	20	$200/(t + 1000)$	200000	0.2
Deep sea (concave)	20	$200/(t + 1000)$	200000	0.2
Resource Gathering	50	$500/(t + 1000)$	500000	0.5
Fruit Tree (depth = 5)	50	$150/(t + 1000)$	150000	0.2
Fruit Tree (depth = 6)	50	$150/(t + 1000)$	150000	0.2
Fruit Tree (depth = 7)	50	$300/(t + 1000)$	300000	0.2

Table 3. Distance in value space in different tasks (mean $\pm$ std)

TASK	DOUBLECHAIN	SINGLECHAIN	(HAQUE & MAGULURI, 2024) & (ZHANG ET AL., 2021B) & (CHEN ET AL., 2025)	(KIM ET AL., 2025)
RANDOM WALK (50)	0.12 ± 0.02	0.04 ± 0.00	0.04 ± 0.00	0.04 ± 0.00
RANDOM WALK (100)	0.19 ± 0.02	0.07 ± 0.01	0.06 ± 0.01	0.06 ± 0.01
RANDOM WALK (1000)	2.94 ± 0.13	2.82 ± 0.03	2.84 ± 0.03	2.92 ± 0.03
FROZEN LAKE	0.44 ± 0.00	0.44 ± 0.01	0.44 ± 0.01	0.44 ± 0.01
CLIFF WALKING	0.62 ± 0.12	0.85 ± 0.02	0.88 ± 0.01	0.88 ± 0.01
TAXI	15.88 ± 2.96	5.90 ± 0.38	9.05 ± 0.45	8.96 ± 0.42
GRID WORLD (5x5)	0.04 ± 0.01	0.01 ± 0.00	0.12 ± 0.00	0.12 ± 0.00
GRID WORLD (10x10)	0.44 ± 0.03	0.49 ± 0.01	0.45 ± 0.01	0.48 ± 0.01
GRID WORLD (2x11)	0.04 ± 0.00	0.01 ± 0.00	0.07 ± 0.00	0.08 ± 0.00
DEEP SEA	0.22 ± 0.06	0.09 ± 0.02	0.28 ± 0.01	0.29 ± 0.01
DEEP SEA (CONCAVE)	0.15 ± 0.03	0.05 ± 0.01	0.15 ± 0.01	0.15 ± 0.01
RESOURCE GATHERING	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
FRUIT TREE (DEPTH = 5)	0.87 ± 0.14	0.49 ± 0.04	0.51 ± 0.02	0.52 ± 0.02
FRUIT TREE (DEPTH = 6)	1.22 ± 0.10	0.51 ± 0.03	0.85 ± 0.01	0.85 ± 0.01
FRUIT TREE (DEPTH = 7)	0.86 ± 0.17	0.38 ± 0.05	0.46 ± 0.06	0.45 ± 0.05

Table 4. Condition numbers in different tasks

Task	η_1	η_2	η_3
Random Walk (50)	3.64×10^{-3}	2.77×10^{-3}	3.04×10^{-3}
Random Walk (100)	3.68×10^{-3}	2.63×10^{-3}	3.19×10^{-3}
Random Walk (1000)	1.91×10^{-4}	1.09×10^{-4}	1.82×10^{-4}
Frozen Lake	4.81×10^{-4}	4.26×10^{-4}	5.20×10^{-5}
Cliff Walking	5.04×10^{-4}	3.27×10^{-4}	7.27×10^{-5}
Taxi	2.29×10^{-4}	1.75×10^{-4}	1.69×10^{-5}
Grid World (5x5)	4.21×10^{-3}	3.94×10^{-3}	1.60×10^{-3}
Grid World (10x10)	4.09×10^{-4}	2.44×10^{-4}	7.83×10^{-5}
Grid World (2x11)	3.77×10^{-3}	3.88×10^{-3}	1.23×10^{-3}
Deep sea	1.73×10^{-3}	1.19×10^{-3}	4.16×10^{-4}
Deep sea (concave)	1.92×10^{-3}	1.28×10^{-3}	3.72×10^{-4}
Resource Gathering	4.35×10^{-3}	4.36×10^{-3}	2.07×10^{-3}
Fruit Tree (depth = 5)	1.28×10^{-5}	7.94×10^{-6}	5.71×10^{-6}
Fruit Tree (depth = 6)	1.07×10^{-4}	6.36×10^{-5}	3.75×10^{-5}
Fruit Tree (depth = 7)	1.92×10^{-4}	1.11×10^{-4}	5.75×10^{-5}

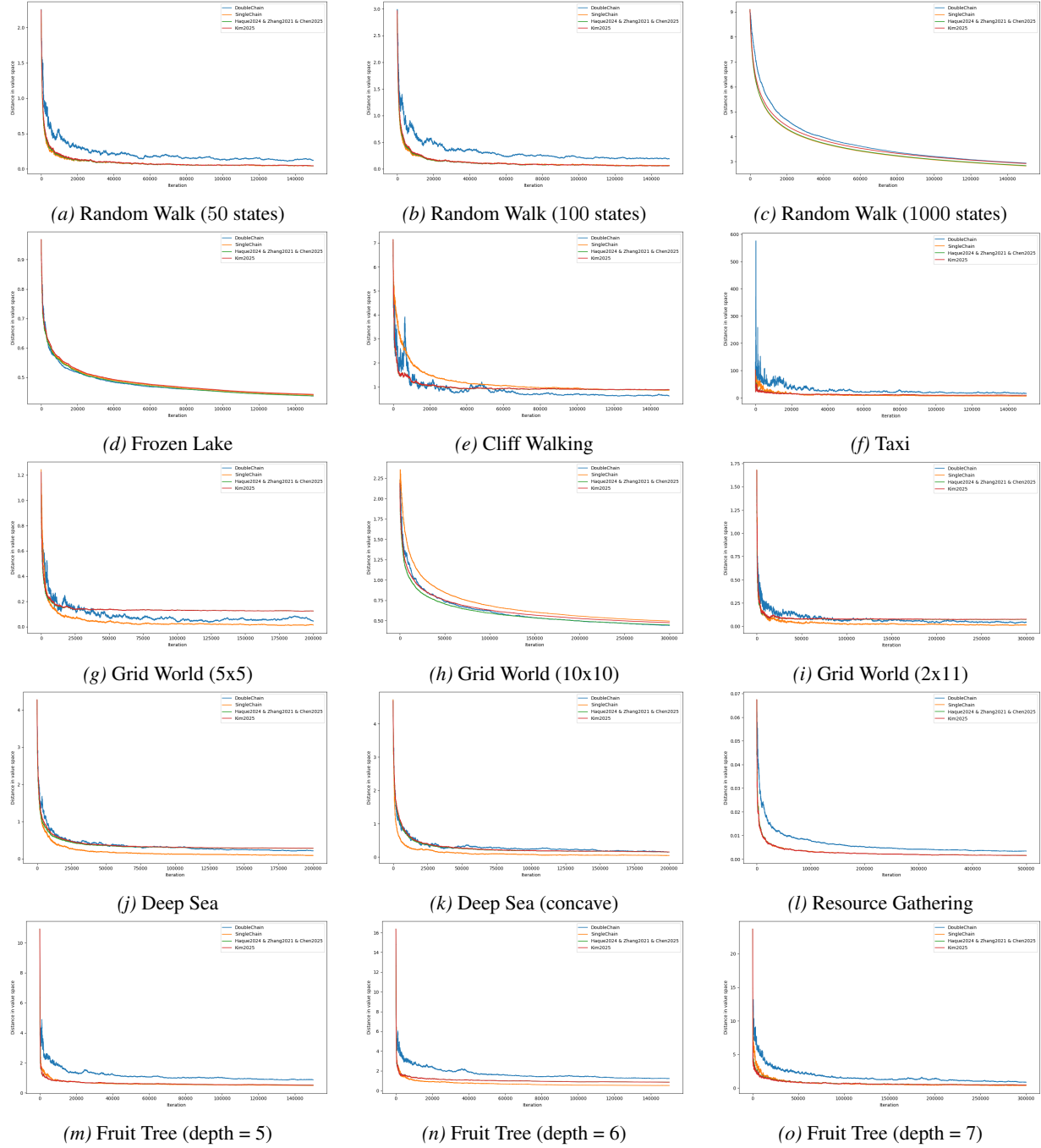


Figure 2. Simulation results