Proceedings of Machine Learning Research vol 336:1–47, 2026 39th Annual Conference on Learning Theory

A Single Stepsize Suffices for Unprojected Linear TD(0): Simultaneous Robust and Fast Rates via Polyak–Ruppert Averaging

Wei-Cheng Lee WEICHENG.FRANK.LEE@GMAIL.COM
Francesco Orabona FRANCESCO@ORABONA.COM
King Abdullah University of Science and Technology (KAUST)
Thuwal, 23955-6900, Kingdom of Saudi Arabia

Editors: Steve Hanneke and Tor Lattimore

Abstract

We study linear TD(0) under Markovian sampling, where data are generated along a single trajectory. We provide high-probability guarantees for a plain *unprojected* TD(0) algorithm with Polyak–Ruppert (PR) averaging, using a *single* stepsize schedule $\eta_t \propto 1/(\tau_{\mathrm{mix}} \log(t) \sqrt{t})$ that depends on mixing time but requires *no prior knowledge of the curvature parameter* ω. Our first result shows that such a choice of the stepsize guarantees that the TD(0) iterates are automatically and uniformly bounded *with high probability*, without projections and without any stability argument based on ω. Building on this result, we establish a simultaneous high-probability convergence guarantee for the PR average: the same stepsize yields both a robust curvature-free $\widetilde{\mathcal{O}}(\tau_{\mathrm{mix}}/\sqrt{T})$ rate and a fast curvature-dependent $\widetilde{\mathcal{O}}(\tau_{\mathrm{mix}}^2/(\omega T))$ rate, with the bound taking the minimum of the two. The core technical ingredient is a Poisson-equation toolkit for geometrically mixing Markov chains, which decomposes Markov noise into a martingale term plus a controlled remainder and enables a new self-bounding inductive argument for pathwise stability.

Keywords: Reinforcement Learning, Temporal Difference Learning, Finite-Time Analysis, Markovian Noise, Stochastic Approximation

1. Introduction

Temporal-difference (TD) learning is one of the central algorithmic primitives in reinforcement learning, used for policy evaluation and as a building block for control methods such as actor–critic. When the state space is large, TD is typically combined with function approximation; in the linear case, the resulting TD(0) recursion admits a clean stochastic-approximation interpretation and converges asymptotically under standard conditions (Sutton, 1988; Tsitsiklis and Van Roy, 1996; Kushner, 2010).

Despite this classical theory, obtaining *finite-time*, *high-probability* guarantees for TD(0) remains challenging, especially in the practically relevant *Markovian sampling* regime where data are generated along a single trajectory.

A major obstacle is that TD(0) is not a standard stochastic gradient method: Even in the linear setting, the update is generally biased and non-symmetric, and the noise is temporally correlated. Moreover, the magnitude of the TD(0) updates depends on the magnitude of the iterates. Therefore, existing non-asymptotic analyses rely on additional structure to control the iterates. One common approach is to enforce boundedness via *projections* onto a known ball (e.g., Bhandari et al., 2018), which simplifies concentration arguments but changes the algorithm and requires *a priori* knowledge of a suitable radius (often tied implicitly to the problem curvature). A second approach is to exploit a *contractive/curvature* structure of the mean dynamics to argue stability without projections;

however, this typically yields stepsize conditions or stability bounds that depend on an unknown curvature parameter, which becomes arbitrarily slow when the curvature is arbitrarily small.

Ideally, one would like to use a *single* stepsize rule that *simultaneously* allows us to obtain rates that are (i) *robust* (curvature-free) and (ii) *fast* when curvature is favorable. Such a stepsize would adapt to the better of the robust and fast high-probability upper bounds, without prior knowledge of the curvature.

Robust and fast rates. To explain the distinction between robust and fast rates, we first briefly describe the potential function we study (Ollivier, 2018; Liu and Olshevsky, 2021) (formally introduced later in (2)), which naturally captures the TD fixed point error under Markovian sampling. This potential always has positive curvature $\omega > 0$, so one can hope for a "fast" rate of order[1] $\widetilde{\mathcal{O}}(\frac{1}{\omega T})$ (up to mixing and logarithmic factors). On the other hand, even without curvature assumptions, one can always achieve a "robust" rate of order $\widetilde{\mathcal{O}}(\frac{1}{\sqrt{T}})$, again up to mixing and logs. It is obvious that, in the finite-time regime, one rate can be better than the other depending on ω, so both rates should be pursued in different situations.

Now, prior work typically achieves these two regimes with *different* stepsize choices and additional algorithmic modifications (e.g., Bhandari et al., 2018). Moreover, the projection-free high-probability analyses under Markovian noise that we are aware of rely on stability arguments whose constants deteriorate with ω (e.g., Samsonov et al., 2024; Durmus et al., 2025).

As far as we know, *obtaining both rates simultaneously with high probability with a single stepsize choice that is independent of* ω *and without projections has not previously been established for TD(0).* In particular, the key technical difficulty in achieving such a result is to control the random iterates, $\boldsymbol{\theta}_t$, *pathwise*. Without projections, the update magnitude scales with $\|\boldsymbol{\theta}_t\|$, and a naive concentration analysis becomes circular: to bound the error one needs bounded updates, but bounded updates require bounded iterates. Moreover, the need to obtain robust rates (i.e., independent of ω) rules out any strategy that shows that the iterates are bounded through a contractive argument based on ω.

Our approach: a self-bounding argument via Poisson equations. In this paper, we show that a simple stepsize, proportional to $\frac{1}{\tau_{\text{mix}} \log t\sqrt{t}}$ and independent of ω, guarantees that TD(0) achieves both fast and robust rates with high probability.

Our main technical novelty to circumvent the above issues is a new *self-bounding* argument that closes this loop *with high probability without* appealing to curvature ω and *without* artificial projections. Concretely, we develop a Poisson-equation toolkit for geometrically mixing Markov chains that decomposes additive Markov noise into a martingale term plus a controlled remainder. This decomposition allows us to control the Markovian bias terms that appear in the basic potential expansion of TD, and to run an induction in which boundedness at times $\leq t-1$ implies boundedness at time t. The resulting uniform bound on $\sup_t \|\boldsymbol{\theta}_t\|$ holds with high probability and is independent of the curvature ω.

Once boundedness is established, we can analyze the convergence guarantee. The bounded-iterates event provides the missing ingredient needed to control the quadratic variation of the martingale terms (again through the Poisson decomposition), yielding high-probability bounds for the averaged iterate $\bar{\boldsymbol{\theta}}_T$. Importantly, the *same* stepsize delivers:

---

[1] The notation $\widetilde{\mathcal{O}}$ suppresses logarithmic factors..

- a curvature-free robust rate of $\widetilde{\mathcal{O}}(\frac{1}{\sqrt{T}})$, through a classic averaged-stochastic-gradient-descent-like analysis, as one would expect from the choice of the stepsize;
- a curvature-dependent fast rate $\widetilde{\mathcal{O}}(1/(\omega T))$, without requiring ω to set the stepsize, thanks to a Polyak–Ruppert averaging analysis.

2. Related Work

In this section, we briefly survey the main results on the convergence guarantees for TD learning with linear function approximation, as well as the main tools we use in our proofs. Later, in Section 5, we will give a detailed comparison in terms of convergence rates.

Early convergence results for TD learning with linear function approximation trace back to Tsitsiklis and Van Roy (1996), who interpreted TD updates through the lens of stochastic approximation (Kushner, 2010). While foundational, that theory is largely asymptotic and does not yield explicit non-asymptotic rates. Finite-time guarantees were later developed in a sequence of works (Korda and La, 2015; Lakshminarayanan and Szepesvári, 2018; Dalal et al., 2018), but these analyses typically assume i.i.d. samples drawn from the stationary distribution. This assumption sidesteps the temporal dependence present in most reinforcement-learning pipelines, where data are generated sequentially along a single Markov-chain trajectory. Accounting for such Markovian correlations substantially complicates the analysis, even for TD(0).

The seminal work of Bhandari et al. (2018) provided the first finite-time treatment under Markovian sampling. They obtained both fast and robust rates, using two different stepsizes. However, their arguments (and even subsequent refinements for the robust rates such as Liu and Olshevsky (2021)) rely on an explicit projection step to keep the iterates controlled.

The need for a projection is removed in subsequent work in the fast regime by exploiting the contractive nature of the update due to the presence of the curvature. In particular, using a control-theoretic framework, Srikant and Ying (2019) derived finite-time error bounds for linear TD with Markovian data without projections by establishing a contraction-like behavior via Lyapunov theory. However, their stepsize depends on the curvature of the potential function, which is generally unknown in practice. Building on this direction, Patil et al. (2023) eliminated the need to know the curvature parameter when choosing stepsizes, at the cost of introducing a data-dropping modification of TD. Related progress has also focused on simplifying proofs and relaxing algorithmic modifications: Mitra (2025) proposed a streamlined inductive two-step analysis, Li et al. (2026) used exponentially decaying stepsizes to avoid data dropping, and Sun et al. (2022) extended fast-rate analyses to neural networks in the NTK regime. More recently, Samsonov et al. (2024) strengthened the analysis of Patil et al. (2023) and obtained high-probability bounds without using projections. Closest to the high-probability projection-free line, Chandak and Borkar (2025) obtained all-time high-probability control for unprojected TD(0) without data dropping. Their analysis, however, remains contractive in nature, and the resulting constants and burn-in depend on the curvature.

The only result that removes the need for projections with stepsizes independent of the curvature of the potential function, and without using curvature in the analysis, is Lee and Orabona (2025), which proves that the iterates of TD(0) are bounded in expectation. Our approach is inspired by their method, but differs substantially because we need high-probability bounds.

Several closely related works study cheap linear-update methods that are informative but not direct head-to-head baselines for plain TD(0). Raj et al. (2022) analyze linear composition optimization and gradient TD methods for off-policy MSPBE minimization, using primal–dual updates rather than the semi-gradient TD(0) recursion studied here, and obtain adaptive finite-time guarantees but require bounded iterates or prior knowledge of the curvature. In the tabular setting, Li et al. (2020) show that asynchronous Q-learning, and hence tabular TD, can have a leading statistical term matching the i.i.d. sampling complexity, with mixing entering only as an additive transient cost; they also provide data-driven stepsizes that avoid prior knowledge of τ_{mix}. These results address the same broad practical question of obtaining fast, stable, cheap linear updates under limited tuning information, but they do not directly apply to unprojected linear TD(0).

As far as we know, there is no prior work that analyzes the possibility of getting both fast and robust rates with a stepsize schedule independent of the curvature and without using projections for linear TD learning.

The use of Polyak–Ruppert averaging in the analysis of TD learning can be traced back at least to Konda (2002). To the best of our knowledge, Korda and La (2015) were the first to leverage this technique to establish finite-time guarantees. In particular, Polyak–Ruppert (tail) averaging has been employed to enable stepsize choices that avoid explicit dependence on the curvature parameter ω (see, e.g., Lakshminarayanan and Szepesvári, 2018; Patil et al., 2023; Samsonov et al., 2024).

Finally, we make use of the Poisson equation to obtain high-probability bounds. This is a standard tool, used widely in the literature on TD learning and stochastic approximation with Markovian noise (see, e.g., Chandak and Borkar, 2025; Blaser and Zhang, 2024).

3. Setting and Assumptions

We work directly with the finite discounted Markov reward process (MRP) induced by a fixed policy μ. Thus the action variables have already been averaged under μ. Let $\mathcal{S} := \{1, \dots, n\}$ be a finite state space, let $P^\mu : \mathcal{S} \times \mathcal{S} \to [0,1]$ be the induced transition kernel, and let $r : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ be the one-step reward function. Given an initial state s_0, the state process satisfies $\mathbb{P}(s_{t+1} = j \mid s_t = i) = P^\mu(i,j)$ for $i, j \in \mathcal{S}$. We write $\boldsymbol{P}^\mu \in \mathbb{R}^{n \times n}$ for the corresponding transition matrix, with entries $\boldsymbol{P}^\mu(i,j) = P^\mu(i,j)$, and fix a discount factor $0 < \gamma < 1$.

We are interested in the *policy evaluation* problem in reinforcement learning (Sutton and Barto, 1998; Mannor et al., 2026). For the MRP induced by μ, the value function is

$$V^\mu(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, s_{t+1}) \,\middle|\, s_0 = s \right].$$

The value function $\boldsymbol{V}^\mu$ can be viewed as a vector in $\mathbb{R}^n$ and is the unique fixed point of the Bellman expectation operator $T^\mu : \mathbb{R}^n \to \mathbb{R}^n$ defined by

$$(T^\mu \boldsymbol{V})(s) := r^\mu(s) + \gamma \sum_{s' \in \mathcal{S}} P^\mu(s,s')\, \boldsymbol{V}(s'), \qquad s \in \mathcal{S},$$

where $r^\mu(s) := \sum_{s' \in \mathcal{S}} P^\mu(s,s')\, r(s,s')$.

When n becomes large, solving the Bellman expectation equation via matrix inversion becomes infeasible. Instead, we consider Temporal-Difference (TD) learning with linear function approximation (Sutton, 1988). Let $\boldsymbol{\phi} : \mathcal{S} \to \mathbb{R}^d$ be a fixed feature mapping with $d \ll n$, and let $\boldsymbol{\theta} \in \mathbb{R}^d$. We approximate $\boldsymbol{V}^\mu$ by $\boldsymbol{V}_{\boldsymbol{\theta}}(s) := \boldsymbol{\theta}^\top \boldsymbol{\phi}(s)$, or in vector form $\boldsymbol{V}_{\boldsymbol{\theta}} = \boldsymbol{\Phi}\boldsymbol{\theta}$, where the feature matrix $\boldsymbol{\Phi} \in \mathbb{R}^{n \times d}$ has row $\boldsymbol{\phi}(s)^\top$ corresponding to state s.

Structural assumption. Throughout the paper, we impose the following standard condition on the induced MRP and the feature matrix.

Assumption 1 (Ergodic induced MRP and full-rank features) *The transition matrix $\boldsymbol{P}^{\mu}$ is irreducible and aperiodic, and the feature matrix $\boldsymbol{\Phi}$ has full column rank.*

Since $\boldsymbol{P}^{\mu}$ is finite, irreducible, and aperiodic, it admits a unique stationary distribution π. We write $\boldsymbol{D} := \operatorname{diag}(\pi)$ and use the value-space norm $\|\boldsymbol{v}\|_{\boldsymbol{D}} := \sqrt{\boldsymbol{v}^{\top}\boldsymbol{D}\boldsymbol{v}}$ for $\boldsymbol{v} \in \mathbb{R}^{n}$. We also define $\Sigma := \boldsymbol{\Phi}^{\top}\boldsymbol{D}\boldsymbol{\Phi}$. Under Assumption 1, $\pi(s) > 0$ for every $s \in \mathcal{S}$, and hence $\Sigma \succ 0$.

For any $\boldsymbol{\theta}_0 \in \mathbb{R}^{d}$, the TD(0) algorithm is defined for $t \geq 0$ by

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \eta_t \boldsymbol{g}(\boldsymbol{\theta}_t, Z_t), \qquad Z_t := (s_t, s_{t+1}),$$

where $\eta_t > 0$ is the stepsize and

$$\boldsymbol{g}(\boldsymbol{\theta}, Z_t) := \big(r(s_t, s_{t+1}) + \gamma \boldsymbol{\phi}(s_{t+1})^{\top}\boldsymbol{\theta} - \boldsymbol{\phi}(s_t)^{\top}\boldsymbol{\theta}\big)\boldsymbol{\phi}(s_t).$$

When the dependence on randomness and iterates is clear, we write $\boldsymbol{g}_t := \boldsymbol{g}(\boldsymbol{\theta}_t, Z_t)$.

The TD(0) update can equivalently be written as $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \eta_t(\boldsymbol{b}_{Z_t} - \boldsymbol{A}_{Z_t}\boldsymbol{\theta}_t)$, where

$$\boldsymbol{A}_{Z_t} := \boldsymbol{\phi}(s_t)\big(\boldsymbol{\phi}(s_t) - \gamma\boldsymbol{\phi}(s_{t+1})\big)^{\top} \in \mathbb{R}^{d\times d}, \qquad \boldsymbol{b}_{Z_t} := r(s_t, s_{t+1})\boldsymbol{\phi}(s_t) \in \mathbb{R}^{d}.$$

Let $\mathcal{Z} := \{(i,j) \in \mathcal{S}\times\mathcal{S} : P^{\mu}(i,j) > 0\}$ be the state space of the transition chain $Z_t = (s_t, s_{t+1})$. Its Markov kernel is

$$P_Z\big((i,j),(j',k)\big) = \begin{cases} P^{\mu}(j,k), & j' = j, \\ 0, & j' \neq j, \end{cases} \qquad (i,j),(j',k) \in \mathcal{Z},$$

and its stationary distribution is $\pi_Z(i,j) := \pi(i)P^{\mu}(i,j)$ for $(i,j) \in \mathcal{Z}$. Define the population TD matrix and vector by $\boldsymbol{A} := \mathbb{E}_{Z\sim\pi_Z}[\boldsymbol{A}_Z]$ and $\boldsymbol{b} := \mathbb{E}_{Z\sim\pi_Z}[\boldsymbol{b}_Z]$.

For $z \in \mathcal{Z}$, define

$$\boldsymbol{\xi}(z) := \boldsymbol{b}_z - \boldsymbol{A}_z\boldsymbol{\theta}^{*}, \qquad \boldsymbol{\delta}(z) := \boldsymbol{A} - \boldsymbol{A}_z.$$

When $z = Z_t$, we write

$$\boldsymbol{\xi}_t := \boldsymbol{\xi}(Z_t), \qquad \boldsymbol{\delta}_t := \boldsymbol{\delta}(Z_t).$$

As shown in Lemma 4 below, Assumption 1 implies that $\boldsymbol{A}$ is nonsingular. Hence the expected TD linear system $\boldsymbol{A}\boldsymbol{\theta}^{*} = \boldsymbol{b}$ has a unique solution $\boldsymbol{\theta}^{*} \in \mathbb{R}^{d}$. It is known that the corresponding value approximation $\boldsymbol{V}_{\boldsymbol{\theta}^{*}} = \boldsymbol{\Phi}\boldsymbol{\theta}^{*}$ satisfies the projected Bellman equation (Tsitsiklis and Van Roy, 1996)

$$\boldsymbol{\Phi}\boldsymbol{\theta}^{*} = \Pi_{\boldsymbol{D}} T^{\mu}(\boldsymbol{\Phi}\boldsymbol{\theta}^{*}),$$

where $\Pi_{\boldsymbol{D}}$ is the orthogonal projection operator onto the subspace $\{\boldsymbol{\Phi}\boldsymbol{x} : \boldsymbol{x} \in \mathbb{R}^{d}\}$ with respect to $\|\cdot\|_{\boldsymbol{D}}$. Moreover,

$$\|\boldsymbol{V}^{\mu} - \boldsymbol{V}_{\boldsymbol{\theta}^{*}}\|_{\boldsymbol{D}} \leq \frac{1}{1-\gamma}\|\boldsymbol{V}^{\mu} - \Pi_{\boldsymbol{D}}\boldsymbol{V}^{\mu}\|_{\boldsymbol{D}}.$$

The *mean-path TD update* is defined as

$$\bar{\boldsymbol{g}}(\boldsymbol{\theta}) := \mathbb{E}_{Z\sim\pi_Z}\big[\big(r(s,s') + \gamma\boldsymbol{\phi}(s')^{\top}\boldsymbol{\theta} - \boldsymbol{\phi}(s)^{\top}\boldsymbol{\theta}\big)\boldsymbol{\phi}(s)\big].$$

Using $\boldsymbol{A}\boldsymbol{\theta}^* = \boldsymbol{b}$, the mean-path TD update is linear in $\boldsymbol{e} := \boldsymbol{\theta} - \boldsymbol{\theta}^*$ for all $\boldsymbol{\theta} \in \mathbb{R}^d$:

$$\bar{\boldsymbol{g}}(\boldsymbol{\theta}) = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{\theta} = -\boldsymbol{A}(\boldsymbol{\theta} - \boldsymbol{\theta}^*), \qquad \bar{\boldsymbol{g}}(\boldsymbol{\theta}^*) = \boldsymbol{0}. \tag{1}$$

To characterize the convergence of TD iterates $\boldsymbol{\theta}_t$ to the TD fixed point $\boldsymbol{\theta}^*$, we introduce the Dirichlet semi-norm $\|\boldsymbol{x}\|_{\text{Dir}}^2 := \boldsymbol{x}^\top \boldsymbol{L}_{\text{Dir}} \boldsymbol{x}$, where

$$\boldsymbol{L}_{\text{Dir}} := \boldsymbol{D} - \tfrac{1}{2}\big(\boldsymbol{D}\boldsymbol{P}^{\mu} + (\boldsymbol{P}^{\mu})^\top \boldsymbol{D}\big).$$

By Lemma 23, $\|\cdot\|_{\text{Dir}}$ is indeed a semi-norm since $\boldsymbol{L}_{\text{Dir}}$ is positive semidefinite. The potential function f that we study (Ollivier, 2018; Liu and Olshevsky, 2021) is defined as

$$f(\boldsymbol{\theta}) := (1-\gamma)\,\|\boldsymbol{V}_{\boldsymbol{\theta}} - \boldsymbol{V}_{\boldsymbol{\theta}^*}\|_{\boldsymbol{D}}^2 + \gamma\,\|\boldsymbol{V}_{\boldsymbol{\theta}} - \boldsymbol{V}_{\boldsymbol{\theta}^*}\|_{\text{Dir}}^2. \tag{2}$$

In particular, for $\boldsymbol{e} = \boldsymbol{\theta} - \boldsymbol{\theta}^*$, since $\boldsymbol{e}^\top \boldsymbol{A}\boldsymbol{e} = \boldsymbol{e}^\top \boldsymbol{A}^\top \boldsymbol{e}$, a direct calculation yields

$$f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}^*) = \tfrac{1}{2}\boldsymbol{e}^\top(\boldsymbol{A} + \boldsymbol{A}^\top)\boldsymbol{e} = \boldsymbol{e}^\top \boldsymbol{A}\boldsymbol{e}. \tag{3}$$

Moreover, using (1), we have

$$\langle \bar{\boldsymbol{g}}(\boldsymbol{\theta}), \boldsymbol{\theta}^* - \boldsymbol{\theta}\rangle = \langle -\boldsymbol{A}\boldsymbol{e}, -\boldsymbol{e}\rangle = \boldsymbol{e}^\top \boldsymbol{A}\boldsymbol{e} = f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}^*). \tag{4}$$

Since $f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}^*) \geq 0$ for $\boldsymbol{\theta} \neq \boldsymbol{\theta}^*$, the mean-path update direction $\bar{\boldsymbol{g}}(\boldsymbol{\theta})$ is aligned with the direction $\boldsymbol{\theta}^* - \boldsymbol{\theta}$. In particular, TD(0) under i.i.d. sampling from π_Z acts as a descent method for f. Additionally, any convergence guarantee on f can be translated to a guarantee in the projected value norm:

$$\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}^2 = \|\boldsymbol{V}_{\boldsymbol{\theta}} - \boldsymbol{V}_{\boldsymbol{\theta}^*}\|_{\boldsymbol{D}}^2 \leq \frac{f(\boldsymbol{\theta})}{1-\gamma}. \tag{5}$$

Constants and structural consequences. We next collect the constants and matrix properties used in the high-probability analysis. The boundedness constants are fixed by the finite state space, while the mixing and curvature constants follow from Assumption 1.

Lemma 2 (Feature and reward bounds) *There exist finite constants* $\phi_\infty, r_\infty < \infty$ *such that* $\|\boldsymbol{\phi}(s)\| \leq \phi_\infty$ *for all* $s \in \mathcal{S}$ *and* $|r(s,s')| \leq r_\infty$ *for all* $s, s' \in \mathcal{S}$. *For* $z = (s,s') \in \mathcal{Z}$, *set*

$$\varepsilon_z := r(s,s') + \gamma\boldsymbol{\phi}(s')^\top \boldsymbol{\theta}^* - \boldsymbol{\phi}(s)^\top \boldsymbol{\theta}^*.$$

Then $\boldsymbol{\xi}(z) = \varepsilon_z \boldsymbol{\phi}(s)$ *and*

$$\|\boldsymbol{\xi}(z)\| \leq r_\infty \phi_\infty + 2\phi_\infty^2 \|\boldsymbol{\theta}^*\|.$$

Lemma 3 (Geometric mixing of (Z_t)) *The transition chain* $Z_t = (s_t, s_{t+1})$ *has stationary distribution* π_Z. *Moreover, there exists a finite constant* $\tau_{\text{mix}} \geq 1$ *such that*

$$\sup_{z \in \mathcal{Z}} \|P_Z^k(z,\cdot) - \pi_Z\|_{\text{TV}} \leq 4 \cdot 2^{-k/\tau_{\text{mix}}}, \qquad k \geq 0,$$

where $P_Z^k(z,z') := \mathbb{P}(Z_k = z' \mid Z_0 = z)$.

Lemma 4 (Curvature of the TD matrix) *The symmetric part of $\boldsymbol{A}$ is positive definite. More precisely,*

$$\frac{\boldsymbol{A}+\boldsymbol{A}^\top}{2} = (1-\gamma)\boldsymbol{\Phi}^\top \boldsymbol{D}\boldsymbol{\Phi} + \gamma\boldsymbol{\Phi}^\top \boldsymbol{L}_{\mathrm{Dir}}\boldsymbol{\Phi} \succeq (1-\gamma)\lambda_{\min}(\boldsymbol{\Phi}^\top \boldsymbol{D}\boldsymbol{\Phi})\mathbf{I}_d.$$

Consequently, one may take $\omega = (1-\gamma)\lambda_{\min}(\boldsymbol{\Phi}^\top \boldsymbol{D}\boldsymbol{\Phi}) > 0$, and

$$f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}^*) = \boldsymbol{e}^\top \frac{\boldsymbol{A}+\boldsymbol{A}^\top}{2}\boldsymbol{e} \geq \omega\|\boldsymbol{e}\|^2 = \omega\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|^2.$$

Thus f is ω star-strongly convex around $\boldsymbol{\theta}^*$.

The first lemma is immediate from the finiteness of $\mathcal{S}$ and the triangle inequality. Lemma 3 is the standard geometric mixing bound for finite irreducible and aperiodic Markov chains; see, e.g., Levin and Peres (2017, Theorem 4.9). For Lemma 4, expand $\boldsymbol{A} = \boldsymbol{\Phi}^\top \boldsymbol{D}(\mathbf{I}-\gamma\boldsymbol{P}^\mu)\boldsymbol{\Phi}$ and symmetrize. Then Lemma 23 gives $\boldsymbol{L}_{\mathrm{Dir}} \succeq 0$, while Assumption 1 gives $\boldsymbol{\Phi}^\top \boldsymbol{D}\boldsymbol{\Phi} \succ 0$.

Remark 5 *For Theorem 6, the curvature conclusion in Lemma 4 is not needed. The bounded-iterates argument only requires a sample-path independent reference point $\boldsymbol{\theta}^*$ satisfying $\boldsymbol{A}\boldsymbol{\theta}^* = \boldsymbol{b}$; in more singular settings, this may be replaced by an appropriate reference set. In the discounted full-rank MRP setting above, the reference point exists uniquely. If $\gamma = 1$ or if $\boldsymbol{\Phi}$ is not full column rank, its existence requires additional regularity conditions. Whenever such a reference point exists, the robust $\widetilde{\mathcal{O}}(1/\sqrt{T})$ rate continues to provide control in terms of the Dirichlet semi-norm.*

Poisson Equation. A critical tool for handling Markovian noise is the Poisson equation associated with a bounded measurable function $h : \mathcal{Z} \to \mathbb{R}^d$. Define the stationary mean $\pi_Z(h) := \mathbb{E}_{Z\sim\pi_Z}[h(Z)]$ and let $\tilde{h} := h - \pi_Z(h)$ be the centered function. The Poisson equation associated with P_Z is

$$u - P_Z u = \tilde{h}, \qquad \text{i.e.,} \qquad u(z) - \sum_{z'\in\mathcal{Z}} P_Z(z,z')\,u(z') = \tilde{h}(z), \qquad z \in \mathcal{Z}. \tag{6}$$

In Appendix A, we will prove that this Poisson equation admits a solution $u^* : \mathcal{Z} \to \mathbb{R}^d$ such that $\|u^*\|_\infty \leq 16\tau_{\mathrm{mix}}\|h\|_\infty$, where $\|h\|_\infty := \sup_{z\in\mathcal{Z}}\|h(z)\|$.

4. Main Results

In this section, we present our main results. We first establish a high-probability uniform bound on the iterates, which is a key step in our convergence analysis. We then derive a simultaneous high-probability convergence rate that is both curvature-free/robust and curvature-aware/fast convergence rate for the Polyak–Ruppert averaged iterate.

4.1. High-Probability Boundedness of the Iterates of Unprojected TD(0)

Our first result shows that, even without any projection, the iterates of TD(0) remain bounded by a constant multiple of $\max\{\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|, \|\boldsymbol{\theta}^*\|, \frac{r_\infty}{\phi_\infty}\}$ under a simple stepsize schedule that exploits the update structure of TD(0).

Theorem 6 (High-probability bounded iterates) *Under Assumption 1, let ϕ_∞, r_∞ and τ_{mix} be the constants introduced in Lemmas 2 and 3. Consider the following stepsize schedule:*

$$\eta_t = \eta_{\text{base}} a_t, \quad \textit{where} \quad \eta_{\text{base}} := \frac{1}{c\tau_{\text{mix}}\,\phi_\infty^2}, \tag{7}$$

for some numerical constant $c > 0$, and a non-increasing positive sequence (a_t) such that $\sum_{t=0}^{\infty} a_t^2$ is finite and $a_0 \le 1$. Fix any $\delta \in (0,1)$ and let

$$R_{\text{base}} := \max\left\{\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|, \|\boldsymbol{\theta}^*\|, \frac{r_\infty}{\phi_\infty}\right\}. \tag{8}$$

Define[2] $A_1(\delta) := 1536\sqrt{\sum_{t=0}^{\infty} a_t^2}\sqrt{2\log\frac{2}{\delta}} + 2304$*,* $A_2 := 2706\sum_{t=0}^{\infty} a_t^2$*,*

$$c_{\min}(\delta) := \frac{A_1(\delta) + \sqrt{A_1^2(\delta) + 4A_2}}{2}, \qquad \textit{and} \qquad \rho := \frac{2c}{\sqrt{c^2 - A_1(\delta)c - A_2}}. \tag{9}$$

Then, provided that $c > c_{\min}(\delta)$, with probability at least $1-\delta$, we have

$$\sup_{t \ge 0} \|\boldsymbol{\theta}_t\|_2 \le \rho R_{\text{base}}.$$

Note that a simple choice of a_t that satisfies the assumptions of the theorem is $a_t = \frac{1}{\ln(t+3)\sqrt{t+1}}$. We provide the complete proof in Appendix B and a proof sketch below.

Proof Sketch. Fix $\delta > 0$ and suppress the δ-dependence in the concentration bounds for simplicity. Starting from the definition of the TD(0) update, we have

$$\|\boldsymbol{g}_k\| = \left\|\big(r_k + \gamma\langle\boldsymbol{\phi}(s_{k+1}), \boldsymbol{\theta}_k\rangle - \langle\boldsymbol{\phi}(s_k), \boldsymbol{\theta}_k\rangle\big)\boldsymbol{\phi}(s_k)\right\| \le |r_k|\,\|\boldsymbol{\phi}(s_k)\| + |1+\gamma|\,\|\boldsymbol{\phi}(s_k)\|^2\,\|\boldsymbol{\theta}_k\|. \tag{10}$$

Thus, in each iteration, the update magnitude is at most $r_\infty\phi_\infty + 2\phi_\infty^2\|\boldsymbol{\theta}_k\|$, which is of the same order as $\|\boldsymbol{\theta}_k\|$.
Next, expand $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|^2$ using the TD(0) update and sum from $k = 1$ to t, to obtain

$$\begin{aligned}\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|^2 = \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|^2 + \sum_{k=1}^{t}\eta_{k-1}^2\|\boldsymbol{g}_{k-1}\|^2 + 2\sum_{k=1}^{t}\eta_{k-1}\langle\boldsymbol{g}_{k-1} - \bar{\boldsymbol{g}}(\boldsymbol{\theta}_{k-1}), \boldsymbol{\theta}_{k-1} - \boldsymbol{\theta}^*\rangle \\ + 2\sum_{k=1}^{t}\eta_{k-1}\langle\bar{\boldsymbol{g}}(\boldsymbol{\theta}_{k-1}), \boldsymbol{\theta}_{k-1} - \boldsymbol{\theta}^*\rangle.\end{aligned} \tag{11}$$

To build intuition, suppose for the moment that $\boldsymbol{g}_k = \bar{\boldsymbol{g}}(\boldsymbol{\theta}_k)$. Then, equation (11) simplifies to

$$\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|^2 - \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|^2 \le \sum_{k=1}^{t}\eta_{k-1}^2\|\bar{\boldsymbol{g}}(\boldsymbol{\theta}_{k-1})\|^2 = \mathcal{O}\left(\sum_{k=1}^{t}\eta_{k-1}^2\left(\|\boldsymbol{\theta}_{k-1}\| + \|\boldsymbol{\theta}_{k-1}\|^2\right)\right),$$

where we use $\langle\bar{\boldsymbol{g}}(\boldsymbol{\theta}_{k-1}), \boldsymbol{\theta}_{k-1} - \boldsymbol{\theta}^*\rangle \le 0$ from equation (4). Hence, $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|^2$ is governed by a term of the form $\mathcal{O}\Big(\sum_{k=1}^{t}\eta_{k-1}^2\big(\|\boldsymbol{\theta}_{k-1}\| + \|\boldsymbol{\theta}_{k-1}\|^2\big)\Big)$. Since (η_k) is square-summable, if $\|\boldsymbol{\theta}_k\|$ is

---

[2] We keep track of all numerical constants because future work may focus on sharpening the minimal value of c.

bounded by R_{base} for all $k \le t-1$, then $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|$ is also bounded by a constant multiple of R_{base}. Moreover, by the triangle inequality, $\|\boldsymbol{\theta}_t\| \le \|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\| + \|\boldsymbol{\theta}^*\|$. In other words, a bound on $\|\boldsymbol{\theta}_k\|$ for all $k \le t-1$ implies a bound at time t. This motivates the following induction hypothesis for some $\rho > 2$:

$$\max_{1 \le k \le t-1} \|\boldsymbol{\theta}_{k-1}\| \le \rho R_{\text{base}} .$$

The core of the proof then becomes to choose the stepsize parameter c (and hence ρ) so that

$$\|\boldsymbol{\theta}_t\|^2 \le 2\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|^2 + 2\|\boldsymbol{\theta}^*\|^2 \le \rho^2 R_{\text{base}}^2,$$

where we use the induction hypothesis to control $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|$.

We now return to the Markovian bias term $\sum_{k=1}^t \eta_{k-1}\langle \boldsymbol{g}_{k-1} - \bar{\boldsymbol{g}}(\boldsymbol{\theta}_{k-1}), \boldsymbol{\theta}_{k-1} - \boldsymbol{\theta}^*\rangle$ in (11). In general, for $\mathcal{F}_t := \sigma(Z_0, \ldots, Z_t)$, $\mathbb{E}\big[\boldsymbol{g}_{k-1} - \bar{\boldsymbol{g}}(\boldsymbol{\theta}_{k-1}) \bigm| \mathcal{F}_{k-2}\big]$ need not be zero: $\boldsymbol{g}_{k-1}$ depends on the fresh randomness in Z_{k-1}, and the conditional distribution of Z_{k-1} given Z_{k-2} does not necessarily coincide with the stationary distribution π_Z. To control this bias, freeze the past iterate $\boldsymbol{\theta}_{k-1}$ and define the centered scalar forcing term $h_{k-1}(z) := \langle \boldsymbol{g}(\boldsymbol{\theta}_{k-1}, z) - \bar{\boldsymbol{g}}(\boldsymbol{\theta}_{k-1}), \boldsymbol{\theta}_{k-1} - \boldsymbol{\theta}^*\rangle$. For this forcing term, consider the Poisson equation on the transition chain Z_t:

$$u_{k-1} : \mathcal{Z} \to \mathbb{R}, \quad u_{k-1}(z) - \mathbb{E}[u_{k-1}(Z_1) \mid Z_0 = z] = h_{k-1}(z) . \tag{12}$$

By Lemma 8, the equation (12) is solved by $u_{k-1}^*(z) := \sum_{i=0}^{\infty} \mathbb{E}[h_{k-1}(Z_i) \mid Z_0 = z]$, and we also have

$$\left\|u_{k-1}^*\right\|_\infty := \sup_z |u_{k-1}^*(z)| \le 16\tau_{\text{mix}} \left\|h_{k-1}\right\|_\infty = \mathcal{O}\left(\tau_{\text{mix}} \left\|\boldsymbol{\theta}_{k-1}\right\|^2\right) . \tag{13}$$

The solution u_{k-1}^* provides a convenient decomposition of $\sum_{k=1}^t \eta_{k-1} h_{k-1}(Z_{k-1})$ into a martingale term M_t plus a remainder term R_t:

$$\begin{aligned}\sum_{k=1}^t \eta_{k-1} h_{k-1}(Z_{k-1}) &= \sum_{k=1}^t \eta_{k-1}\left(u_{k-1}^*(Z_k) - \mathbb{E}\big[u_{k-1}^*(Z_k) \bigm| Z_{k-1}\big] + u_{k-1}^*(Z_{k-1}) - u_{k-1}^*(Z_k)\right) \\ &= M_t + R_t, \end{aligned} \tag{14}$$

where $\Delta M_k := \eta_{k-1}(u_{k-1}^*(Z_k) - \mathbb{E}[u_{k-1}^*(Z_k)|Z_{k-1}])$ and $M_t := \sum_{k=1}^t \Delta M_k$. Note that $\|\Delta M_k\|_\infty = \mathcal{O}(\eta_{k-1}\tau_{\text{mix}} \left\|\boldsymbol{\theta}_{k-1}\right\|^2) = \mathcal{O}(a_{k-1} \left\|\boldsymbol{\theta}_{k-1}\right\|^2)$ from (13). Consequently, the accumulated bounded difference for M_t is controlled by $\mathcal{O}(\max_{k \le t} \left\|\boldsymbol{\theta}_{k-1}\right\|^2)$. Assuming again that $\max_{k \le t} \left\|\boldsymbol{\theta}_{k-1}\right\| \le \rho R_{\text{base}}$, Pinelis' inequality (Pinelis, 1994) for martingale concentration (Lemma 11) yields $|M_t| = \mathcal{O}(\rho^2 R_{\text{base}}^2)$ with high probability.

To control the remainder term $R_t = \sum_{k=1}^t \eta_{k-1}(u_{k-1}^*(Z_{k-1}) - u_{k-1}^*(Z_k))$, we rewrite it (see Lemma 10) as

$$R_t = \eta_0 u_0^*(Z_0) - \eta_t u_t^*(Z_t) - \sum_{k=1}^t (\eta_{k-1} - \eta_k) u_{k-1}^*(Z_k) - \sum_{k=1}^t \eta_k \left(u_{k-1}^*(Z_k) - u_k^*(Z_k)\right) .$$

Since $\left\|u_k^* - u_{k-1}^*\right\|_\infty$ is controlled by $\mathcal{O}(\eta_{k-1}\tau_{\text{mix}} \left\|\boldsymbol{\theta}_{k-1}\right\|^2)$ (because $\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-1} = \eta_{k-1}\boldsymbol{g}_{k-1}$), the induction hypothesis again implies $|R_t| = \mathcal{O}(\rho^2 R_{\text{base}}^2)$. Combining the bounds for M_t and R_t, we obtain $\sum_{k=1}^t \eta_{k-1} h_{k-1}(Z_{k-1}) = \mathcal{O}(\rho^2 R_{\text{base}}^2)$, which fits perfectly with our inductive proof, since $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|^2$ remains the same order $\mathcal{O}(\rho^2 R_{\text{base}}^2)$ as in the idealized case $\boldsymbol{g}_k = \bar{\boldsymbol{g}}(\boldsymbol{\theta}_k)$.

4.2. High-Probability Convergence Rate for Polyak–Ruppert Averaging

Our second result studies the convergence rate of the Polyak–Ruppert (PR) averaged iterate. Let

$$S_T := \sum_{t=1}^{T} \eta_{t-1} \qquad \bar{\boldsymbol{\theta}}_T := \frac{1}{S_T} \sum_{t=1}^{T} \eta_{t-1} \boldsymbol{\theta}_{t-1} \,.$$

Recall the following potential defined in (2):

$$f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}^*) = (1-\gamma) \left\| \boldsymbol{V}_{\boldsymbol{\theta}} - \boldsymbol{V}_{\boldsymbol{\theta}^*} \right\|_{\boldsymbol{D}}^2 + \gamma \left\| \boldsymbol{V}_{\boldsymbol{\theta}} - \boldsymbol{V}_{\boldsymbol{\theta}^*} \right\|_{\mathrm{Dir}}^2 .$$

We have the following high-probability guarantees for the PR average $\bar{\boldsymbol{\theta}}_T$.

Theorem 7 (High-probability rate for PR averaging) *Under Assumption 1, let $\phi_\infty, r_\infty, \tau_{\mathrm{mix}}$ and ω be the constants introduced in Lemmas 2–4. Consider the same stepsize schedule (η_t) as in (7) and the same R_{base} in (8). Also, define $H := \sum_{t=0}^{\infty} \eta_t^2$. Fix any $\delta \in (0,1)$ and let $A_1(\delta) = 1536\sqrt{\sum_{t=0}^{\infty} a_t^2}\sqrt{2\log\frac{8}{\delta}} + 2304$ and $A_2 = 2706\sum_{t=0}^{\infty} a_t^2$. Define*

$$c_{\min}(\delta) := \frac{A_1(\delta) + \sqrt{A_1(\delta)^2 + 4A_2}}{2}, \qquad \rho := \frac{2c}{\sqrt{c^2 - A_1(\delta)c - A_2}} \,.$$

Suppose $c > c_{\min}(\delta)$. Then, with probability at least $1-\delta$, we have for all $T \geq 1$,

$$f(\bar{\boldsymbol{\theta}}_T) - f(\boldsymbol{\theta}^*) \leq \min\left\{ \frac{C_{\mathrm{fast}}^2}{\omega S_T^2}, \frac{C_{\mathrm{robust}}}{S_T} \right\},$$

where

$$\begin{aligned} C_{\mathrm{fast}} &:= \rho R_{\mathrm{base}} \left[2 + 2\tau_{\mathrm{mix}}\phi_\infty^2 \left(264\eta_0 + 176\sqrt{H}\sqrt{2\log\frac{8}{\delta}} \right) + 192\tau_{\mathrm{mix}}\phi_\infty^4 H \right], \\ C_{\mathrm{robust}} &:= \rho^2 R_{\mathrm{base}}^2 \left[0.5 + 2\tau_{\mathrm{mix}}\phi_\infty^2 \left(288\eta_0 + 192\sqrt{H}\sqrt{2\log\frac{8}{\delta}} \right) + (672\tau_{\mathrm{mix}} + 4.5)\,\phi_\infty^4 H \right] . \end{aligned}$$

A choice of the stepsize that satisfies the assumptions of the theorem is (see Appendix D)

$$\eta_t = (c\,\phi_\infty^2\,\tau_{\mathrm{mix}})^{-1} a_t, \qquad \text{where} \qquad a_t = (\log(t+3)\sqrt{t+1})^{-1} \,. \tag{15}$$

The logarithmic correction in a_t is used at the square-summability boundary: it ensures that $\sum_{t=0}^{\infty} \eta_t^2$ is finite while keeping $S_T = \sum_{t=1}^{T} \eta_{t-1}$ within a logarithmic factor of $\sqrt{T}$. Such a choice results in a robust rate of $\widetilde{\mathcal{O}}(\frac{\|\boldsymbol{\theta}^*\|^2 \tau_{\mathrm{mix}} \phi_\infty^2}{\sqrt{T}})$ and a fast rate of $\widetilde{\mathcal{O}}(\frac{\|\boldsymbol{\theta}^*\|^2 \tau_{\mathrm{mix}}^2 \phi_\infty^4}{\omega T})$ with high probability.

In Appendix E, we also show an alternative stepsize choice that does not require knowledge of τ_{mix}, at the cost of a doubly exponential dependence on τ_{mix} in the constants.

Proof Sketch of Theorem 7. By Theorem 6, with high probability we have $\sup_{t\geq 0}\ \|\boldsymbol{\theta}_t\|_2 \leq \rho\, R_{\mathrm{base}}$. It then follows that $\sum_{t=1}^{T}\eta_{t-1}^2\left\|\boldsymbol{g}_{t-1}\right\|^2$ is of order $\mathcal{O}(R_{\mathrm{base}}^2)$, using the gradient bound (10). Moreover, the cumulated Markov bias term $\sum_{t=1}^{T}\eta_{t-1}\langle\boldsymbol{g}_{t-1}-\bar{\boldsymbol{g}}(\boldsymbol{\theta}_{t-1}),\boldsymbol{\theta}_{t-1}-\boldsymbol{\theta}^*\rangle$ is also of order $\mathcal{O}(R_{\mathrm{base}}^2)$, by applying the Poisson equation and Pinelis' inequality.
Rearranging (11), we obtain

$$\begin{aligned}2\sum_{t=1}^{T}\eta_{t-1}\langle\bar{\boldsymbol{g}}(\boldsymbol{\theta}_{t-1}),\boldsymbol{\theta}^*-\boldsymbol{\theta}_{t-1}\rangle &= \|\boldsymbol{\theta}_0-\boldsymbol{\theta}^*\|^2-\|\boldsymbol{\theta}_T-\boldsymbol{\theta}^*\|^2\\ &\quad+\sum_{t=1}^{T}\eta_{t-1}^2\|\boldsymbol{g}_{t-1}\|^2+2\sum_{t=1}^{T}\eta_{t-1}\langle\boldsymbol{g}_{t-1}-\bar{\boldsymbol{g}}(\boldsymbol{\theta}_{t-1}),\boldsymbol{\theta}_{t-1}-\boldsymbol{\theta}^*\rangle\\ &=\mathcal{O}(R_{\mathrm{base}}^2)\,.\end{aligned}$$

The robust rate then follows from the convexity of f together with the equation (4):

$$f(\bar{\boldsymbol{\theta}}_T)-f(\boldsymbol{\theta}^*)\leq\frac{1}{\sum_{t=1}^{T}\eta_{t-1}}\sum_{t=1}^{T}\eta_{t-1}\langle\bar{\boldsymbol{g}}(\boldsymbol{\theta}_{t-1}),\boldsymbol{\theta}^*-\boldsymbol{\theta}_{t-1}\rangle=\mathcal{O}\left(\frac{R_{\mathrm{base}}^2}{S_T}\right)\,.$$

To derive the fast rate, recall that the TD update can be written as $\boldsymbol{g}(\boldsymbol{\theta}_{t-1},Z_{t-1})=-\boldsymbol{A}_{Z_{t-1}}\boldsymbol{\theta}_{t-1}+\boldsymbol{b}_{Z_{t-1}}$. Then the centered error $\boldsymbol{e}_t:=\boldsymbol{\theta}_t-\boldsymbol{\theta}^*$ can be shown (see (24)) to evolve as

$$\boldsymbol{e}_t=(\mathbf{I}_d-\eta_{t-1}\boldsymbol{A})\boldsymbol{e}_{t-1}+\eta_{t-1}\boldsymbol{\xi}_{t-1}+\eta_{t-1}\boldsymbol{\delta}_{t-1}\boldsymbol{e}_{t-1}\,. \tag{16}$$

Rearranging yields

$$\boldsymbol{A}\boldsymbol{e}_{t-1}=\frac{\boldsymbol{e}_{t-1}-\boldsymbol{e}_t}{\eta_{t-1}}+\boldsymbol{\xi}_{t-1}+\boldsymbol{\delta}_{t-1}\boldsymbol{e}_{t-1}\,.$$

Define $\bar{\boldsymbol{e}}_T:=\frac{1}{S_T}\sum_{t=1}^{T}\eta_{t-1}\boldsymbol{e}_{t-1}$. Then $\boldsymbol{A}\bar{\boldsymbol{e}}_T=I_1+I_2+I_3$, where

$$I_1:=\frac{1}{S_T}\sum_{t=1}^{T}(\boldsymbol{e}_{t-1}-\boldsymbol{e}_t),\qquad I_2:=\frac{1}{S_T}\sum_{t=1}^{T}\eta_{t-1}\boldsymbol{\xi}_{t-1},\qquad I_3:=\frac{1}{S_T}\sum_{t=1}^{T}\eta_{t-1}\boldsymbol{\delta}_{t-1}\boldsymbol{e}_{t-1}\,.$$

By Lemma 4, we have $(\boldsymbol{A}+\boldsymbol{A}^\top)/2\succeq\omega\mathbf{I}_d$. Combining this with the equation (4) again, we obtain

$$f(\bar{\boldsymbol{\theta}}_T)-f(\boldsymbol{\theta}^*)=\langle\boldsymbol{A}\bar{\boldsymbol{e}}_T,\bar{\boldsymbol{e}}_T\rangle\leq\|\boldsymbol{A}\bar{\boldsymbol{e}}_T\|\,\|\bar{\boldsymbol{e}}_T\|\leq\frac{1}{\omega}\,\|\boldsymbol{A}\bar{\boldsymbol{e}}_T\|^2\,.$$

The fast rate follows by controlling $\|I_1\|$, $\|I_2\|$, and $\|I_3\|$ separately. The term $\|I_1\|$ is bounded using the telescoping sum. For $\|I_2\|$ and $\|I_3\|$, we again apply Lemma 8 with forcing terms $\boldsymbol{\xi}_{t-1}$ and $\boldsymbol{\delta}_{t-1}\boldsymbol{e}_{t-1}$, respectively. Therefore, by Pinelis' inequality and the normalization by S_T, both $\|I_2\|$ and $\|I_3\|$ are of order $\mathcal{O}(R_{\mathrm{base}}/S_T)$ with high probability, which completes the proof sketch for the fast rate.

5. Detailed Comparison with Prior Results

In this section, we compare Theorems 6 and 7 with prior finite-time results for TD(0) with linear function approximation. This comparison is delicate because prior papers use slightly different assumptions on the problem and the algorithm. We therefore compare the results on a common footing, while emphasizing which unknown quantities are needed to set the algorithmic hyperparameters. For simplicity, in this section we assume $\phi_\infty = 1$.

First, we recall our guarantees under the stepsize schedule in (15): $\eta_t = \frac{1}{c\,\tau_{\text{mix}}\sqrt{t+1}\log(t+3)}$, where c satisfies the assumptions of Theorem 6. Under this choice, TD(0) with Polyak–Ruppert averaging *simultaneously* achieves the following three guarantees:

- bounded iterates, independent of ω,
- the ω-robust rate $\mathcal{O}\left(\frac{\tau_{\text{mix}} R_{\text{base}}^2 \log T \sqrt{\log(1/\delta)}}{\sqrt{T}}\right)$,
- the fast ω-aware rate $\mathcal{O}\left(\frac{\tau_{\text{mix}}^2 R_{\text{base}}^2 \log^2 T \log(1/\delta)}{\omega T}\right)$, with probability at least $1-\delta$.

Importantly, all guarantees use a single stepsize schedule that is independent of the curvature parameter ω.

Prior work on bounded iterates. To the best of our knowledge, our result is the first to show that TD(0) with Markovian sampling has bounded iterates with high probability without using ω in the stepsize or in a curvature-based stability argument.

The closest result is Lee and Orabona (2025), but they only provide expectation bounds, whose rate we match. While one would expect that it is relatively easy to convert expectation results to high-probability ones for well-behaved random variables, here it requires completely different machinery based on the Poisson equation.

Our proof proceeds by establishing the boundedness of the iterates via a carefully designed inductive argument. This argument uses only the condition $\langle \bar{\boldsymbol{g}}(\boldsymbol{\theta}), \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle \leq 0$. We then control the Markovian bias pathwise using a Poisson-equation-based decomposition. This approach is fundamentally different from the Linear Stochastic Approximation (LSA)-stability-based analyses in Srikant and Ying (2019); Mou et al. (2020); Chen et al. (2022); Patil et al. (2023); Li et al. (2024); Samsonov et al. (2024); Durmus et al. (2025) that result in a bound on the norm of the iterates that depends on ω, hence incompatible with the robust rate.

Our approach is also very different from using projections, that is, projecting $\boldsymbol{\theta}_k$ onto the ball $\mathcal{B}(\mathbf{0}, \|\boldsymbol{\theta}^*\|)$ (see, e.g., Bhandari et al., 2018; Liu and Olshevsky, 2021; Sun et al., 2021; Prashanth et al., 2021; Patil et al., 2023).

Such projections play an analytic rather than algorithmic role: they guarantee that $|\boldsymbol{g}(\boldsymbol{\theta}, \cdot)|_\infty$ is always bounded by a constant depending on $\|\boldsymbol{\theta}^*\|$, so that the following terms can be controlled easily using concentration inequalities:

$$\sum_k \eta_k^2 \|\boldsymbol{g}_k\|^2, \qquad \sum_k \eta_k \langle \boldsymbol{g}_k - \bar{\boldsymbol{g}}(\boldsymbol{\theta}_k), \boldsymbol{\theta}_k - \boldsymbol{\theta}^* \rangle.$$

These terms scale with $\|\boldsymbol{\theta}^*\|^2$ instead of an unknown bound on $\|\boldsymbol{\theta}_k\|^2$. We remark that this kind of modification of TD(0) is not used in practice, and it would require an additional hyperparameter. Moreover, simply bounding $\|\boldsymbol{\theta}^*\|$ with the loose bound (Bhandari et al., 2018, Lemma 7) $\|\boldsymbol{\theta}^*\| \leq \frac{2r_\infty}{\sqrt{\omega}(1-\lambda)}$ once again breaks the ω-independent result required for the robust rate.

Prior work on robust rates. To the best of our knowledge, ours is the first result to achieve the robust rate $\widetilde{\mathcal{O}}(\frac{\|\boldsymbol{\theta}^*\|^2\tau_{\text{mix}}}{\sqrt{T}})$ with high probability, without any algorithmic modification and with a simple stepsize schedule independent of ω. Once one considers the implicit dependence of T on τ_{mix} in some prior work, we match the rate in expectation from prior results (see, e.g., Bhandari et al., 2018; Liu and Olshevsky, 2021; Lee and Orabona, 2025), up to polylog factors. On the other hand, we require knowledge of τ_{mix} to set the stepsizes. Instead, prior work, either implicitly or explicitly, requires the number of iterations T to be large enough with respect to some function of τ_{mix} for the rate to be $\widetilde{\mathcal{O}}(\frac{1}{\sqrt{T}})$.

Our analysis is fundamentally different from previous ones,with the notable exception of the expectation result of Lee and Orabona (2025) that we already discussed. Indeed, the stability-based approach commonly used in the literature, which relies on (16) to unroll $\boldsymbol{\theta}_t - \boldsymbol{\theta}^*$, would typically result in a bound on $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|$ that scales like $\sum_{i=1}^{t}(1-\lambda_{\min}\eta)^i$ using inequalities similar to (17), which only controls $\|\boldsymbol{\theta}_t\|$ at the scale $\mathcal{O}(1/\lambda_{\min})$. This dependence is incompatible with the ω-independent bounded-iterates guarantee in Theorem 6.

Prior work on fast rates. The most relevant prior works on fast rates are Samsonov et al. (2024); Durmus et al. (2025).[3] These papers view TD(0) with linear function approximation as a special case of LSA. They prove that there exist constants $a > 0$, $\varkappa_p > 0$, and $\alpha_{p,\infty} > 0$ (depending on p and possibly on τ_{mix}^{-1}) such that $p\alpha_{p,\infty} \le 1/2$ and TD(0) satisfies the following exponential stability condition: for any moment order $p > 0$ and any stepsize $\alpha \in (0, \alpha_{p,\infty})$

$$\mathbb{E}^{1/p}\left[\|\Gamma^{\alpha}_{m:n}u\|^p\right] \le \varkappa_p(1-\alpha a)^n \|u\|\,, \quad \text{for any } u \in \mathbb{R}^d, \quad 1 \le m \le n\,, \tag{17}$$

where $\Gamma^{\alpha}_{m:n} = \prod_{k=m}^{n}(\mathbf{I} - \alpha \boldsymbol{A}_{Z_k})$ is a product of random matrices. This condition allows one to unroll the recursion and decompose $\boldsymbol{\theta}_t - \boldsymbol{\theta}^*$ as

$$\boldsymbol{\theta}_t - \boldsymbol{\theta}^* = (\mathbf{I} - \alpha \boldsymbol{A}_{Z_t})\,(\boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^*) + \alpha \boldsymbol{\xi}_{Z_t} = \Gamma^{\alpha}_{1:t}(\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*) + \alpha \sum_{i=1}^{t} \Gamma^{\alpha}_{i+1:t}\boldsymbol{\xi}_{Z_i}\,.$$

Then, one can use p-th moment bounds on $\Gamma^{\alpha}_{m:n}$ to control $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}$.

Under this approach, and recalling that $\omega \ge (1-\gamma)\lambda_{\min}(\boldsymbol{\Phi}^\top \boldsymbol{D}\boldsymbol{\Phi})$, Samsonov et al. (2024, Theorem 6) shows that, with high probability, a data-dropping modification of TD(0) (Nagaraj et al., 2020) achieves the rate $\mathcal{O}(\frac{\tau_{\text{mix}}\|\boldsymbol{\theta}^*\|^2\log^2(T/\delta)}{(1-\gamma)^2\lambda_{\min}^2 T})$ for $\|\bar{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*\|^2_{\boldsymbol{\Sigma}}$. Their bound also includes an additional exponentially decaying term that depends on $\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|^2$. The fixed stepsize α in Samsonov et al. (2024) is ω-agnostic, but the data-dropping strategy requires knowledge of τ_{mix}, and it is not commonly used in practice.

In parallel, Durmus et al. (2025, Corollary 2) considers a varying stepsize $\eta_t = \mathcal{O}(t^{-2/3})$ that depends on $\lambda_{\min}$ and τ_{mix} and obtains a better high-probability guarantee in which $\|\bar{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*\|^2_{\boldsymbol{\Sigma}}$ decays at rate $\mathcal{O}(\frac{\tau_{\text{mix}}\|\boldsymbol{\theta}^*\|^2\log(1/\delta)}{(1-\gamma)^2\lambda_{\min}T})$, plus an exponentially decaying term depending on $\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|^2$.

To compare these results with our bounds, we can use equation (5) to translate our bound on f into a bound on $\|\bar{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*\|^2_{\boldsymbol{\Sigma}}$. Hence, we obtain the fast rate $\widetilde{\mathcal{O}}(\frac{\tau_{\text{mix}}^2 R_{\text{base}}^2}{(1-\gamma)^2\lambda_{\min}T})$ in Theorem 7. Up to logarithmic factors, this rate matches the best-known high-probability results in the Markovian sampling regime (Samsonov et al., 2024; Durmus et al., 2025) and the i.i.d. stationary sampling

---

[3] As observed by Durmus et al. (2025, p. 7), the results in Mou et al. (2020) are based on restrictive assumptions that do not allow a fair comparison with ours and previous results, so we omit them here.

regime (Li et al., 2024) for all relevant non-mixing quantities. Indeed, our R^2_{base} has the same order of magnitude as $|\boldsymbol{\theta}^*|^2$ that appears in the other bounds. However, our τ^2_{mix} dependence is worse by one factor of τ_{mix} than the linear τ_{mix} dependence obtained in the Markovian bounds of Samsonov et al. (2024); Durmus et al. (2025). In addition, our initial-error term $\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|^2$ decays at a $1/T$ rate rather than exponentially, since we avoid relying on a contraction-based argument in our analysis.

The worse dependence on τ_{mix} could be a side effect of our Poisson-equation-based analysis. On the other hand, it could also be due to the fact that we do not use a data-dropping method (Samsonov et al., 2024) or knowledge of $\lambda_{\min}$ (Durmus et al., 2025).

More generally, it is unclear whether any of the above results are optimal in their respective settings. To the best of our knowledge, the closest lower bound is $\Omega(\frac{\|\boldsymbol{\theta}^*\|^2}{(1-\gamma)\lambda_{\min}T})$ due to Li et al. (2024, Theorem 2), which considers TD(0) with linear function approximation under the easier setting of i.i.d. sampling. Thus, when data dropping is not allowed, projections are not used, and $\lambda_{\min}$ is unknown, the optimal dependence of our rate on τ_{mix} and $1-\gamma$ remains open.

A reason one might expect such a dependence on τ_{mix} in the convergence rate or stepsize is that the TD fixed point $\boldsymbol{\theta}^*$ is defined with respect to the stationary distribution, and a slowly mixing Markov chain can require more samples to accurately estimate stationary expectations (Nagaraj et al., 2020). At the same time, tabular results show that one should not expect a multiplicative mixing-time dependence to be necessary in all settings: Li et al. (2020) show that, for asynchronous Q-learning and hence tabular TD, the leading statistical term can match the i.i.d. sampling setting, with only an additive transient τ_{mix} cost. In the linear function approximation setting, existing results still account for the mixing time in various ways: stepsize conditions that explicitly depend on τ_{mix} appear in (Srikant and Ying, 2019; Mitra, 2025; Durmus et al., 2025); data-dropping approaches that retain only one out of $\mathcal{O}(\tau_{\text{mix}})$ samples are studied in (Patil et al., 2023; Samsonov et al., 2024); and several results require the total sample budget T to be sufficiently large relative to τ_{mix} (Lee and Orabona, 2025; Li et al., 2026).

Acknowledgments

We thank Chi-Jen Lu, Li Chen, and our anonymous reviewers for their insightful comments. We also acknowledge the use of large language models, including GPT and Gemini, for editorial assistance and for suggesting that Poisson-equation-based methods may provide a useful way to handle randomness in our analysis. The authors independently verified, developed, and finalized all mathematical arguments, results, and references.

References

Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. In *Conference on Learning Theory*, pages 1691–1692. PMLR, 2018.

Ethan Blaser and Shangtong Zhang. Asymptotic and finite sample analysis of nonexpansive stochastic approximations with Markovian noise. *arXiv preprint arXiv:2409.19546*, 2024.

Siddharth Chandak and Vivek S. Borkar. A concentration bound for TD(0) with function approximation. *Stochastic Systems*, 2025. doi: 10.1287/stsy.2023.0055.

Zaiwei Chen, Sheng Zhang, Thinh T Doan, John-Paul Clarke, and Siva Theja Maguluri. Finite-sample analysis of nonlinear stochastic approximation with applications in reinforcement learning. *Automatica*, 146:110623, 2022.

Gal Dalal, Balázs Szörényi, Gugan Thoppe, and Shie Mannor. Finite sample analyses for TD(0) with function approximation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Alain Durmus, Eric Moulines, Alexey Naumov, and Sergey Samsonov. Finite-time high-probability bounds for Polyak–Ruppert averaged iterates of linear stochastic approximation. *Mathematics of Operations Research*, 50(2):935–964, 2025.

Vijaymohan R. Konda. *Actor-critic algorithms*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 2002. URL `https://dspace.mit.edu/handle/1721.1/8120`. Ph.D. thesis. Available via DSpace@MIT.

Nathaniel Korda and Prashanth La. On TD(0) with function approximation: Concentration bounds and a centered variant with exponential convergence. In *International Conference on Machine Learning*, pages 626–634. PMLR, 2015.

Harold Kushner. Stochastic approximation: a survey. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):87–96, 2010.

Chandrashekar Lakshminarayanan and Csaba Szepesvári. Linear stochastic approximation: How far does constant step-size and iterate averaging go? In *International Conference on Artificial Intelligence and Statistics*, pages 1347–1355. PMLR, 2018.

Wei-Cheng Lee and Francesco Orabona. A robust $\widetilde{\mathcal{O}}(1/\sqrt{T})$ rate for unprojected TD learning with linear function approximation. *arXiv preprint arXiv:2506.01052*, 2025.

David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Society, 2017.

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction. *Advances in neural information processing systems*, 33:7031–7043, 2020.

Gen Li, Weichen Wu, Yuejie Chi, Cong Ma, Alessandro Rinaldo, and Yuting Wei. High-probability sample complexities for policy evaluation with linear function approximation. *IEEE Transactions on Information Theory*, 70(8):5969–5999, 2024.

Yunxiang Li, Mark Schmidt, Reza Babanezhad, and Sharan Vaswani. Towards parameter-free temporal difference learning. *arXiv preprint arXiv:2603.02577*, 2026.

Rui Liu and Alex Olshevsky. Temporal difference learning as gradient splitting. In *International Conference on Machine Learning*, pages 6905–6913. PMLR, 2021.

Shie Mannor, Yishay Mansour, and Aviv Tamar. *Reinforcement Learning: Foundations*. Cambridge University Press, 2026. URL `https://sites.google.com/view/rlfoundations/home`.

Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

Aritra Mitra. A simple finite-time analysis of TD learning with linear function approximation. *IEEE Transactions on Automatic Control*, 70(2):1388–1394, 2025. doi: 10.1109/TAC.2024.3469328.

Wenlong Mou, Chris Junchi Li, Martin J Wainwright, Peter L Bartlett, and Michael I Jordan. On linear stochastic approximation: Fine-grained Polyak-Ruppert and non-asymptotic concentration. In *Proceedings of Thirty Third Conference on Learning Theory*, pages 2947–2997, 2020.

Dheeraj Nagaraj, Xian Wu, Guy Bresler, Prateek Jain, and Praneeth Netrapalli. Least squares regression with Markovian data: Fundamental limits and algorithms. *Advances in Neural Information Processing Systems*, 33:16666–16676, 2020.

Yann Ollivier. Approximate temporal difference learning is a gradient descent for reversible policies. *arXiv preprint arXiv:1805.00869*, 2018.

Gandharv Patil, L. A. Prashanth, Dheeraj Nagaraj, and Doina Precup. Finite time analysis of temporal difference learning with linear function approximation: Tail averaging and regularisation. In *International Conference on Artificial Intelligence and Statistics*, pages 5438–5448. PMLR, 2023.

Iosif Pinelis. Optimum bounds for the distributions of martingales in Banach spaces. *The Annals of Probability*, pages 1679–1706, 1994.

L. A. Prashanth, Nathaniel Korda, and Rémi Munos. Concentration bounds for temporal difference learning with linear function approximation: The case of batch data and uniform sampling. *Machine Learning*, 110(3):559–618, 2021.

Anant Raj, Pooria Joulani, Andras Gyorgy, and Csaba Szepesvári. Faster rates, adaptive algorithms, and finite-time bounds for linear composition optimization and gradient TD learning. In *International Conference on Artificial Intelligence and Statistics*, pages 7176–7186. PMLR, 2022.

Sergey Samsonov, Daniil Tiapkin, Alexey Naumov, and Eric Moulines. Improved high-probability bounds for the temporal difference learning algorithm via exponential stability. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 4511–4547. PMLR, 2024.

Rayadurgam Srikant and Lei Ying. Finite-time error bounds for linear stochastic approximation and TD learning. In *Conference on Learning Theory*, pages 2803–2830. PMLR, 2019.

Tao Sun, Han Shen, Tianyi Chen, and Dongsheng Li. Adaptive temporal difference learning with linear function approximation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):8812–8824, 2021.

Tao Sun, Dongsheng Li, and Bao Wang. Finite-time analysis of adaptive temporal difference learning with deep neural networks. *Advances in Neural Information Processing Systems*, 35:19592–19604, 2022.

Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3:9–44, 1988.

Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*, volume 1. MIT Press, 1998.

John Tsitsiklis and Benjamin Van Roy. Analysis of temporal-difference learning with function approximation. *Advances in Neural Information Processing Systems*, 9, 1996.

Appendix A. Poisson Toolkit for Markov Noise

In this section, we develop a Poisson-equation toolkit for Markov concentration. Throughout, let $(Z_t)_{t\geq 0}$ be the Markov chain on

$$\mathcal{Z} := \{(i,j) \in \mathcal{S} \times \mathcal{S} : P^{\mu}(i,j) > 0\}$$

with kernel P_Z and stationary distribution π_Z, as defined in Section 3. The state space $\mathcal{Z}$ is finite.

A.1. Poisson Equation under Geometric Mixing

We start with a standard Poisson equation lemma for geometrically mixing chains in $\mathbb{R}^d$ (Meyn and Tweedie, 2012, Theorem 17.4.3).

Lemma 8 (Poisson solution under geometric mixing) *Let $h : \mathcal{Z} \to \mathbb{R}^d$ be a bounded measurable function with stationary mean zero, i.e., $\mathbb{E}_{\pi_Z}[h(Z)] = 0$. Then, the Poisson equation*

$$u - P_Z u = h, \qquad \text{on } \mathcal{Z}, \tag{18}$$

where $(P_Z u)(z) := \mathbb{E}[u(Z_1) \mid Z_0 = z]$, admits a solution

$$u^*(z) := \sum_{k=0}^{\infty} (P_Z^k h)(z),$$

where $(P_Z^k h)(z) := \mathbb{E}[h(Z_k) \mid Z_0 = z]$. Moreover,

$$\|u^*\|_\infty \leq 16\tau_{\text{mix}} \|h\|_\infty, \qquad \text{where } \|h\|_\infty := \sup_{z\in\mathcal{Z}} \|h(z)\| .$$

Proof Define $u^*(z) := \sum_{k=0}^{\infty} (P_Z^k h)(z)$. Since $\mathbb{E}_{\pi_Z}[h(Z)] = 0$, for every $k \geq 0$ and every $z \in \mathcal{Z}$,

$$(P_Z^k h)(z) = \int h(z')\{P_Z^k(z, dz') - \pi_Z(dz')\} .$$

Using the geometric mixing bound from Lemma 3, we have for all $k \geq 0$,

$$\|(P_Z^k h)\|_\infty = \sup_{z\in\mathcal{Z}} \left\|\mathbb{E}\big[h(Z_k) \mid Z_0 = z\big]\right\| \leq \sup_{z\in\mathcal{Z}} 2\|h\|_\infty \|P_Z^k(z,\cdot) - \pi_Z\|_{\text{TV}} \leq 8 \cdot 2^{-k/\tau_{\text{mix}}} \|h\|_\infty .$$

Thus, the series defining u^* converges absolutely and uniformly with

$$\|u^*\|_\infty \leq \sum_{k=0}^{\infty} \|(P_Z^k h)\|_\infty \leq \frac{8}{1 - 2^{-1/\tau_{\text{mix}}}} \|h\|_\infty := C(\tau_{\text{mix}}) \|h\|_\infty .$$

Moreover,

$$(P_Z u^*)(z) = \sum_{k=0}^{\infty} P_Z^{k+1} h(z),$$

so

$$u^*(z) - (P_Z u^*)(z) = (P_Z^0 h)(z) = h(z),$$

that is, u^* solves (18).

Let $x = 1/\tau_{\text{mix}}$. We scale the function $C(\tau_{\text{mix}})$ by $1/\tau_{\text{mix}}$ and consider the new function $\tilde{C}(x)$:

$$\tilde{C}(x) := \frac{1}{\tau_{\text{mix}}} C\left(\frac{1}{x}\right) = x \cdot \frac{8 \cdot 2^x}{2^x - 1} = \frac{8x}{1 - 2^{-x}} \,.$$

We claim that $\tilde{C}(x)$ is monotonically increasing on $(0, 1]$. Let $t = x \ln 2$. Then $\tilde{C}(x)$ can be written as $\frac{8}{\ln 2} \frac{t}{1-e^{-t}}$. Consider the derivative of $f(t) = \frac{t}{1-e^{-t}}$ for $t > 0$:

$$f'(t) = \frac{1(1 - e^{-t}) - t(e^{-t})}{(1 - e^{-t})^2} = \frac{1 - (1 + t)e^{-t}}{(1 - e^{-t})^2} \,.$$

Using the elementary inequality $e^t > 1 + t$ for $t > 0$, we have $1 > (1 + t)e^{-t}$, which implies $f'(t) > 0$. Thus, $\tilde{C}(x)$ is strictly increasing on $(0, 1]$. Consequently, the maximum value of $\tilde{C}(x)$ on the interval $(0, 1]$ occurs at $x = 1$:

$$\tilde{C}(x) \le \tilde{C}(1) = \frac{8 \cdot 1}{1 - 2^{-1}} = \frac{8}{0.5} = 16 \,.$$

Therefore, we have $C(\tau_{\text{mix}}) \le 16\tau_{\text{mix}}$, which completes the proof. ■

A.2. Bounds for TD-type Poisson solutions

We now specialize Lemma 8 to the functions arising from TD. For $z = (s, s') \in \mathcal{Z}$, define the scalar function

$$h_{\boldsymbol{\theta},\boldsymbol{\theta}-\boldsymbol{\theta}^*}(z) := \big\langle \boldsymbol{g}(\boldsymbol{\theta}, z) - \bar{\boldsymbol{g}}(\boldsymbol{\theta}),\, \boldsymbol{\theta} - \boldsymbol{\theta}^* \big\rangle,$$

and the vector-valued functions

$$\begin{aligned} h(z) &:= \varepsilon_z \boldsymbol{\phi}(s) \in \mathbb{R}^d, && \text{where we recall } \varepsilon_z = r(s, s') + \gamma \boldsymbol{\phi}(s')^\top \boldsymbol{\theta}^* - \boldsymbol{\phi}(s)^\top \boldsymbol{\theta}^*, \\ h_{\boldsymbol{\theta}}(z) &:= \boldsymbol{\delta}_z \big(\boldsymbol{\theta} - \boldsymbol{\theta}^*\big) \in \mathbb{R}^d, && \text{where we define } \boldsymbol{\delta}_z := \boldsymbol{A} - \boldsymbol{A}_z \in \mathbb{R}^{d \times d} \,. \end{aligned}$$

By construction, $h_{\boldsymbol{\theta},\boldsymbol{\theta}-\boldsymbol{\theta}^*}$, h, and $h_{\boldsymbol{\theta}}$ all have stationary mean zero under π_Z. We study properties of the corresponding Poisson-equation solutions in the following lemma.

Lemma 9 (Bounds and Lipschitz properties of TD Poisson equation solutions) *In the setting of Section 3, let ϕ_∞, r_∞ and τ_{mix} be the constants introduced in Lemmas 2 and 3. Then the following statements hold.*

(i) **Growth bounds.** *For all $z \in \mathcal{Z}$,*

$$\begin{aligned} |h_{\boldsymbol{\theta},\boldsymbol{\theta}-\boldsymbol{\theta}^*}(z)| &\le (2r_\infty \phi_\infty + 4\phi_\infty^2 \|\boldsymbol{\theta}\|)\, \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|, \\ \|h(z)\| &\le r_\infty \phi_\infty + 2\phi_\infty^2 \|\boldsymbol{\theta}^*\|, \\ \|h_{\boldsymbol{\theta}}(z)\| &\le 4\phi_\infty^2 \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \,. \end{aligned}$$

(ii) **Bounded Poisson solutions.** *Let* $u^*_{\boldsymbol{\theta},\boldsymbol{\theta}-\boldsymbol{\theta}^*} : \mathcal{Z} \to \mathbb{R}$ *be the infinite series solution of the Poisson equation*

$$u^*_{\boldsymbol{\theta},\boldsymbol{\theta}-\boldsymbol{\theta}^*} - P_Z u^*_{\boldsymbol{\theta},\boldsymbol{\theta}-\boldsymbol{\theta}^*} = h_{\boldsymbol{\theta},\boldsymbol{\theta}-\boldsymbol{\theta}^*},$$

and let $u^* : \mathcal{Z} \to \mathbb{R}^d$ *be the infinite series solution of*

$$u^* - P_Z u^* = h,$$

and let $u^*_{\boldsymbol{\theta}} : \mathcal{Z} \to \mathbb{R}^d$ *be the infinite series solution of*

$$u^*_{\boldsymbol{\theta}} - P_Z u^*_{\boldsymbol{\theta}} = h_{\boldsymbol{\theta}}\,.$$

Then,

$$\begin{aligned}\|u^*_{\boldsymbol{\theta},\boldsymbol{\theta}-\boldsymbol{\theta}^*}\|_\infty &\le 16\tau_{\text{mix}}\|h_{\boldsymbol{\theta},\boldsymbol{\theta}-\boldsymbol{\theta}^*}\|_\infty \le 16\tau_{\text{mix}}(2r_\infty\phi_\infty + 4\phi_\infty^2\|\boldsymbol{\theta}\|)\,\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|,\\ \|u^*\|_\infty &\le 16\tau_{\text{mix}}\|h\|_\infty \le 16\tau_{\text{mix}}(r_\infty\phi_\infty + 2\phi_\infty^2\|\boldsymbol{\theta}^*\|),\\ \|u^*_{\boldsymbol{\theta}}\|_\infty &\le 16\tau_{\text{mix}}\|h_{\boldsymbol{\theta}}\|_\infty \le 64\tau_{\text{mix}}\phi_\infty^2\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|\,.\end{aligned}$$

(iii) **Local Lipschitz continuity in $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$.** *With the definitions in the previous point, we also have*

$$\begin{aligned}\|u^*_{\boldsymbol{\theta},\boldsymbol{\theta}-\boldsymbol{\theta}^*} - u^*_{\boldsymbol{\theta}',\boldsymbol{\theta}'-\boldsymbol{\theta}^*}\|_\infty &\le 16\tau_{\text{mix}}\|h_{\boldsymbol{\theta},\boldsymbol{\theta}-\boldsymbol{\theta}^*} - h_{\boldsymbol{\theta}',\boldsymbol{\theta}'-\boldsymbol{\theta}^*}\|_\infty\\ &\le 16\tau_{\text{mix}}(2r_\infty\phi_\infty + 4\phi_\infty^2\,\|\boldsymbol{\theta}\|)\,\left\|\boldsymbol{\theta}-\boldsymbol{\theta}'\right\|\\ &\quad + 64\tau_{\text{mix}}\phi_\infty^2\,\left\|\boldsymbol{\theta}'-\boldsymbol{\theta}^*\right\|\,\left\|\boldsymbol{\theta}-\boldsymbol{\theta}'\right\|,\end{aligned}$$

and

$$\|u^*_{\boldsymbol{\theta}} - u^*_{\boldsymbol{\theta}'}\|_\infty \le 64\tau_{\text{mix}}\phi_\infty^2\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|\,.$$

Proof *(i) Growth bounds.* By the feature and reward bounds in Lemma 2, for $z = (s, s')$ we have

$$\begin{aligned}\|\boldsymbol{g}(\boldsymbol{\theta}, z)\| &= \left|r(s,s') + \gamma\boldsymbol{\phi}(s')^\top\boldsymbol{\theta} - \boldsymbol{\phi}(s)^\top\boldsymbol{\theta}\right| \cdot \|\boldsymbol{\phi}(s)\|\\ &\le \big(r_\infty + (\gamma+1)\phi_\infty\|\boldsymbol{\theta}\|\big)\,\phi_\infty \le \big(r_\infty + 2\phi_\infty\|\boldsymbol{\theta}\|\big)\,\phi_\infty\,.\end{aligned}$$

Taking expectations under π_Z yields the same type of bound for $\|\bar{\boldsymbol{g}}(\boldsymbol{\theta})\|$, hence

$$\|\boldsymbol{g}(\boldsymbol{\theta}, z) - \bar{\boldsymbol{g}}(\boldsymbol{\theta})\| \le 2(r_\infty + 2\phi_\infty\|\boldsymbol{\theta}\|)\,\phi_\infty = 2r_\infty\phi_\infty + 4\phi_\infty^2\|\boldsymbol{\theta}\|\,.$$

Therefore,

$$|h_{\boldsymbol{\theta},\boldsymbol{\theta}-\boldsymbol{\theta}^*}(z)| = \left|\langle \boldsymbol{g}(\boldsymbol{\theta}, z) - \bar{\boldsymbol{g}}(\boldsymbol{\theta}), \boldsymbol{\theta}-\boldsymbol{\theta}^*\rangle\right| \le (2r_\infty\phi_\infty + 4\phi_\infty^2\|\boldsymbol{\theta}\|)\,\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|\,.$$

For h, recall that

$$\varepsilon_z = r(s,s') + \gamma\boldsymbol{\phi}(s')^\top\boldsymbol{\theta}^* - \boldsymbol{\phi}(s)^\top\boldsymbol{\theta}^*\,.$$

Then,

$$\|h(z)\| = \|\varepsilon_z\boldsymbol{\phi}(s)\| \le r_\infty\phi_\infty + 2\phi_\infty^2\|\boldsymbol{\theta}^*\|\,.$$

For $h_{\boldsymbol{\theta}}$, recall that

$$\boldsymbol{A}_z = \boldsymbol{\phi}(s)\big(\boldsymbol{\phi}(s) - \gamma\boldsymbol{\phi}(s')\big)^\top\,.$$

Then,

$$\|\boldsymbol{A}_z\|_{\mathrm{op}} \le \|\boldsymbol{\phi}(s)\|\big(\|\boldsymbol{\phi}(s)\| + \gamma\|\boldsymbol{\phi}(s')\|\big) \le (1+\gamma)\phi_\infty^2 \le 2\phi_\infty^2\,.$$

Consequently, $\|\boldsymbol{A}\|_{\mathrm{op}} \le 2\phi_\infty^2$ and

$$\|\boldsymbol{\delta}_z\|_{\mathrm{op}} \le \|\boldsymbol{A}\|_{\mathrm{op}} + \|\boldsymbol{A}_z\|_{\mathrm{op}} \le 4\phi_\infty^2\,.$$

Thus,

$$\|h_{\boldsymbol{\theta}}(z)\| = \|\boldsymbol{\delta}_z(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\| \le 4\phi_\infty^2\,\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|\ .$$

(ii) Poisson solution sup-norm bounds. It follows from Lemma 8 applied to the real-valued function $h_{\boldsymbol{\theta},\boldsymbol{\theta}-\boldsymbol{\theta}^*}$ and the $\mathbb{R}^d$ functions h and $h_{\boldsymbol{\theta}}$.

(iii) Lipschitz continuity in $\boldsymbol{\theta}$. Scalar case $h_{\boldsymbol{\theta},\boldsymbol{\theta}-\boldsymbol{\theta}^*}$. For each $z=(s,s')$ and $\boldsymbol{\theta}\in\mathbb{R}^d$ recall that

$$\boldsymbol{A}_z := \boldsymbol{\phi}(s)\big(\boldsymbol{\phi}(s)-\gamma\boldsymbol{\phi}(s')\big)^\top, \qquad \boldsymbol{b}_z := r(s,s')\boldsymbol{\phi}(s),$$

and

$$\boldsymbol{A} = \mathbb{E}_{\pi_Z}[\boldsymbol{A}_Z], \qquad \boldsymbol{b} = \mathbb{E}_{\pi_Z}[\boldsymbol{b}_Z],$$

so that $\bar{\boldsymbol{g}}(\boldsymbol{\theta}) = -\boldsymbol{A}\boldsymbol{\theta}+\boldsymbol{b}$ and $\boldsymbol{g}(\boldsymbol{\theta},z) = -\boldsymbol{A}_z\boldsymbol{\theta}+\boldsymbol{b}_z$. Also, recall that

$$\boldsymbol{\delta}_z = \boldsymbol{A}-\boldsymbol{A}_z,\ \text{ and let } \qquad \boldsymbol{\zeta}_z := \boldsymbol{b}_z - \boldsymbol{b}\,.$$

Then

$$\boldsymbol{g}(\boldsymbol{\theta},z) - \bar{\boldsymbol{g}}(\boldsymbol{\theta}) = \boldsymbol{\delta}_z\boldsymbol{\theta} + \boldsymbol{\zeta}_z,$$

and hence

$$h_{\boldsymbol{\theta},\boldsymbol{\theta}-\boldsymbol{\theta}^*}(z) = \big\langle \boldsymbol{\delta}_z\boldsymbol{\theta}+\boldsymbol{\zeta}_z,\, \boldsymbol{\theta}-\boldsymbol{\theta}^*\big\rangle\,.$$

As above, $\|\boldsymbol{\delta}_z\|_{\mathrm{op}} \le 4\phi_\infty^2$, and $\|\boldsymbol{b}_z\| \le r_\infty\phi_\infty$ implies $\|\boldsymbol{\zeta}_z\| \le 2r_\infty\phi_\infty$. We have

$$\begin{aligned}
|h_{\boldsymbol{\theta},\boldsymbol{\theta}-\boldsymbol{\theta}^*}(z) - h_{\boldsymbol{\theta}',\boldsymbol{\theta}'-\boldsymbol{\theta}^*}(z)| &= |\big\langle \boldsymbol{\delta}_z\boldsymbol{\theta}+\boldsymbol{\zeta}_z,\, \boldsymbol{\theta}-\boldsymbol{\theta}^*\big\rangle - \big\langle \boldsymbol{\delta}_z\boldsymbol{\theta}'+\boldsymbol{\zeta}_z,\, \boldsymbol{\theta}'-\boldsymbol{\theta}^*\big\rangle| \\
&= |\big\langle \boldsymbol{\delta}_z\boldsymbol{\theta}+\boldsymbol{\zeta}_z,\, \boldsymbol{\theta}-\boldsymbol{\theta}^*-\boldsymbol{\theta}'+\boldsymbol{\theta}^*\big\rangle + \big\langle \boldsymbol{\delta}_z\boldsymbol{\theta}+\boldsymbol{\zeta}_z-(\boldsymbol{\delta}_z\boldsymbol{\theta}'+\boldsymbol{\zeta}_z),\, \boldsymbol{\theta}'-\boldsymbol{\theta}^*\big\rangle| \\
&\le \|\boldsymbol{\delta}_z\boldsymbol{\theta}+\boldsymbol{\zeta}_z\|\,\|\boldsymbol{\theta}-\boldsymbol{\theta}'\| + \|\boldsymbol{\delta}_z(\boldsymbol{\theta}-\boldsymbol{\theta}')\|\,\|\boldsymbol{\theta}'-\boldsymbol{\theta}^*\| \\
&\le \big(4\phi_\infty^2\,\|\boldsymbol{\theta}\| + 2r_\infty\phi_\infty\big)\,\|\boldsymbol{\theta}-\boldsymbol{\theta}'\| + 4\phi_\infty^2\,\|\boldsymbol{\theta}'-\boldsymbol{\theta}^*\|\,\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|\ .
\end{aligned}$$

The second inequality follows from applying Lemma 8 and linearity of the Poisson equation:

$$\|u^*_{\boldsymbol{\theta},\boldsymbol{\theta}-\boldsymbol{\theta}^*} - u^*_{\boldsymbol{\theta}',\boldsymbol{\theta}'-\boldsymbol{\theta}^*}\|_\infty \le 16\tau_{\mathrm{mix}}\|h_{\boldsymbol{\theta},\boldsymbol{\theta}-\boldsymbol{\theta}^*} - h_{\boldsymbol{\theta}',\boldsymbol{\theta}'-\boldsymbol{\theta}^*}\|_\infty\,.$$

Vector case $h_{\boldsymbol{\theta}}$. For $h_{\boldsymbol{\theta}}$,

$$h_{\boldsymbol{\theta}}(z) - h_{\boldsymbol{\theta}'}(z) = \boldsymbol{\delta}_z(\boldsymbol{\theta}-\boldsymbol{\theta}'),$$

so

$$\|h_{\boldsymbol{\theta}}(z) - h_{\boldsymbol{\theta}'}(z)\| \le \|\boldsymbol{\delta}_z\|_{\mathrm{op}}\|\boldsymbol{\theta}-\boldsymbol{\theta}'\| \le 4\phi_\infty^2\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|\,.$$

Taking the supremum over z gives

$$\|h_{\boldsymbol{\theta}} - h_{\boldsymbol{\theta}'}\|_\infty \le 4\phi_\infty^2\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|\,.$$

Thus, by Lemma 8 and linearity of the Poisson equation,

$$\|u^*_{\boldsymbol{\theta}} - u^*_{\boldsymbol{\theta}'}\|_\infty \le 64\tau_{\mathrm{mix}}\phi_\infty^2\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|\,.$$

■

A.3. Poisson Martingale decomposition of additive Markov noise

Recall $\mathcal{F}_t := \sigma(Z_0, \dots, Z_t)$. Let u_{t-1} be the infinite series solution of the Poisson equation with forcing function h_{t-1} defined as one of $h_{\boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_{t-1}-\boldsymbol{\theta}^*}$, h, or $h_{\boldsymbol{\theta}_{t-1}}$, that is

$$u_{t-1} - P_Z u_{t-1} = h_{t-1} \,.$$

Then, we have

$$\begin{aligned}
\sum_{t=1}^{T} \eta_{t-1} h_{t-1}(Z_{t-1}) &= \sum_{t=1}^{T} \eta_{t-1}\big(u_{t-1}(Z_{t-1}) - (P_Z u_{t-1})\,(Z_{t-1})\big) \\
&= \sum_{t=1}^{T} \eta_{t-1}\big(u_{t-1}(Z_{t-1}) - u_{t-1}(Z_t) + u_{t-1}(Z_t) - (P_Z u_{t-1})\,(Z_{t-1})\big) \\
&= \sum_{t=1}^{T} \eta_{t-1}\big(u_{t-1}(Z_{t-1}) - u_{t-1}(Z_t)\big) + \sum_{t=1}^{T} \eta_{t-1} \Delta M_t,
\end{aligned}$$

where $\Delta M_t := u_{t-1}(Z_t) - (P_Z u_{t-1})\,(Z_{t-1})$ is a martingale difference with respect to $\mathcal{F}_t := \sigma(Z_0, \dots, Z_t)$ since $u_{t-1} \in \mathcal{F}_{t-1}$ and

$$\mathbb{E}[u_{t-1}(Z_t) \mid \mathcal{F}_{t-1}] = \mathbb{E}[u_{t-1}(Z_t) \mid \mathcal{F}_{t-1}, u_{t-1}] = (P_Z u_{t-1})\,(Z_{t-1}) \,.$$

We will use this Poisson decomposition repeatedly throughout the paper.
Thus, for all $T \geq 1$, we obtain

$$M_T := \sum_{t=1}^{T} \eta_{t-1} \Delta M_t, \quad R_T := \sum_{t=1}^{T} \eta_{t-1}\big(u_{t-1}(Z_{t-1}) - u_{t-1}(Z_t)\big), \tag{19}$$

and $(M_T)_{T\geq 1}$ is a $\mathbb{R}^d$-valued martingale with respect to $(\mathcal{F}_T)_{T\geq 1}$. Moreover, R_T enjoys the following representation:

Lemma 10 (Summation by parts)

$$R_T = \eta_0 u_0(Z_0) - \eta_T u_T(Z_T) - \sum_{t=1}^{T} (\eta_{t-1} - \eta_t) u_{t-1}(Z_t) - \sum_{t=1}^{T} \eta_t\big(u_{t-1}(Z_t) - u_t(Z_t)\big) \,.$$

Proof Let $R_T = \sum_{t=1}^{T} \eta_{t-1}(u_{t-1}(Z_{t-1}) - u_{t-1}(Z_t))$ and write

$$R_T = \underbrace{\sum_{t=1}^{T} \eta_{t-1} u_{t-1}(Z_{t-1})}_{T_T^{(1)}} - \underbrace{\sum_{t=1}^{T} \eta_{t-1} u_{t-1}(Z_t)}_{T_T^{(2)}} \,.$$

Reindex the first sum as $T_T^{(1)} = \sum_{j=0}^{T-1} \eta_j u_j(Z_j)$. For the second sum, add and subtract the terms $\eta_t u_t(Z_t)$ and $\eta_t u_{t-1}(Z_t)$:

$$\eta_{t-1} u_{t-1}(Z_t) = \eta_t u_t(Z_t) + (\eta_{t-1} - \eta_t) u_{t-1}(Z_t) + \eta_t (u_{t-1}(Z_t) - u_t(Z_t)) \,.$$

Summing over $t = 1, \dots, T$ yields

$$T_T^{(2)} = \sum_{t=1}^{T} \eta_t u_t(Z_t) + \sum_{t=1}^{T} (\eta_{t-1} - \eta_t) u_{t-1}(Z_t) + \sum_{t=1}^{T} \eta_t \big(u_{t-1}(Z_t) - u_t(Z_t)\big) \, .$$

Hence,

$$R_T = \eta_0 u_0(Z_0) - \eta_T u_T(Z_T) - \sum_{t=1}^{T} (\eta_{t-1} - \eta_t) u_{t-1}(Z_t) - \sum_{t=1}^{T} \eta_t \big(u_{t-1}(Z_t) - u_t(Z_t)\big) \, .$$

■

Appendix B. Bounded Iterates: Proof of Theorem 6

B.1. Markov noise and localization

Let $\boldsymbol{e}_t := \boldsymbol{\theta}_t - \boldsymbol{\theta}^*$ denote the error at time t. From the TD update, expanding $\boldsymbol{e}_t$ gives

$$\begin{aligned} \|\boldsymbol{e}_t\|^2 &= \|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|^2 = \|\boldsymbol{\theta}_{t-1} + \eta_{t-1} \boldsymbol{g}(\boldsymbol{\theta}_{t-1}, Z_{t-1}) - \boldsymbol{\theta}^*\|^2 \\ &= \|\boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^*\|^2 + \eta_{t-1}^2 \|\boldsymbol{g}(\boldsymbol{\theta}_{t-1}, Z_{t-1})\|^2 + 2\eta_{t-1} \langle \boldsymbol{g}(\boldsymbol{\theta}_{t-1}, Z_{t-1}), \boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^* \rangle \\ &= \|\boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^*\|^2 + \eta_{t-1}^2 \|\boldsymbol{g}(\boldsymbol{\theta}_{t-1}, Z_{t-1})\|^2 + 2\eta_{t-1} \langle \boldsymbol{g}(\boldsymbol{\theta}_{t-1}, Z_{t-1}) - \bar{\boldsymbol{g}}(\boldsymbol{\theta}_{t-1}), \boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^* \rangle \\ &\quad + 2\eta_{t-1} \langle \bar{\boldsymbol{g}}(\boldsymbol{\theta}_{t-1}), \boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^* \rangle \, . \end{aligned}$$

Summing from $t = 1$ to T yields

$$\|\boldsymbol{e}_T\|^2 = \|\boldsymbol{e}_0\|^2 + \sum_{t=1}^{T} \eta_{t-1}^2 \|\boldsymbol{g}(\boldsymbol{\theta}_{t-1}, Z_{t-1})\|^2 + 2 \sum_{t=1}^{T} \eta_{t-1} \langle \bar{\boldsymbol{g}}(\boldsymbol{\theta}_{t-1}), \boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^* \rangle + B_T, \tag{20}$$

where the Markov bias term is

$$B_T := 2 \sum_{t=1}^{T} \eta_{t-1} \, \langle \boldsymbol{g}(\boldsymbol{\theta}_{t-1}, Z_{t-1}) - \bar{\boldsymbol{g}}(\boldsymbol{\theta}_{t-1}), \, \boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^* \rangle \, . \tag{21}$$

By (4) and the nonnegativity of $f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}^*)$,

$$\langle \bar{\boldsymbol{g}}(\boldsymbol{\theta}_{t-1}), \boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^* \rangle = -\big(f(\boldsymbol{\theta}_{t-1}) - f(\boldsymbol{\theta}^*)\big) \leq 0.$$

Thus, we have

$$\|\boldsymbol{e}_T\|^2 \leq \|\boldsymbol{e}_0\|^2 + \sum_{t=1}^{T} \eta_{t-1}^2 \|\boldsymbol{g}(\boldsymbol{\theta}_{t-1}, Z_{t-1})\|^2 + B_T \, .$$

To localize the process, we define

$$T_{\text{exit}} := \inf\{t \geq 0 : \, \|\boldsymbol{\theta}_t\| > \rho R_{\text{base}}\},$$

where T_{exit} is a stopping time with respect to $(\mathcal{F}_t)$ because $\boldsymbol{\theta}_t$ is $\mathcal{F}_t$-measurable. The number $\rho > 2$ will be specified later. Define

$$R_{\text{base}} \coloneqq \max\left\{\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|, \|\boldsymbol{\theta}^*\|, \frac{r_\infty}{\phi_\infty}\right\}, \qquad R_{\max} \coloneqq \rho R_{\text{base}},$$

and the stopped iterates and stepsizes by

$$\tilde{\boldsymbol{\theta}}_t \coloneqq \boldsymbol{\theta}_{t \wedge T_{\text{exit}}}, \qquad \tilde{\eta}_t \coloneqq \begin{cases} \eta_t, & t < T_{\text{exit}}, \\ 0, & t \geq T_{\text{exit}}. \end{cases}$$

We define the stopped Markov bias term

$$\tilde{B}_T \coloneqq 2 \sum_{t=1}^{T} \tilde{\eta}_{t-1} \langle \boldsymbol{g}(\tilde{\boldsymbol{\theta}}_{t-1}, Z_{t-1}) - \bar{\boldsymbol{g}}(\tilde{\boldsymbol{\theta}}_{t-1}), \tilde{\boldsymbol{\theta}}_{t-1} - \boldsymbol{\theta}^* \rangle .$$

In particular, on the event $\{T_{\text{exit}} = \infty\}$ we have $\tilde{\boldsymbol{\theta}}_t = \boldsymbol{\theta}_t$, $\tilde{\eta}_t = \eta_t$, and $\tilde{B}_t = B_t$ for all t.

B.2. Poisson representation and bounds for the localized Markov noise

Let u_{t-1} be the infinite series solution of the Poisson equation with forcing function $h_{t-1} \coloneqq h_{\tilde{\boldsymbol{\theta}}_{t-1}, \tilde{\boldsymbol{\theta}}_{t-1} - \boldsymbol{\theta}^*} \in \mathcal{F}_{t-2}$, that is,

$$u_{t-1} - P_Z u_{t-1} = h_{t-1} .$$

Then, from Section A.3, we have

$$\begin{aligned}
\sum_{t=1}^{T} \tilde{\eta}_{t-1} h_{t-1}(Z_{t-1}) &= \sum_{t=1}^{T} \tilde{\eta}_{t-1} \big(u_{t-1}(Z_{t-1}) - (P_Z u_{t-1})(Z_{t-1})\big) \\
&= \sum_{t=1}^{T} \tilde{\eta}_{t-1} \big(u_{t-1}(Z_{t-1}) - u_{t-1}(Z_t) + u_{t-1}(Z_t) - (P_Z u_{t-1})(Z_{t-1})\big) \\
&= \sum_{t=1}^{T} \tilde{\eta}_{t-1} \big(u_{t-1}(Z_{t-1}) - u_{t-1}(Z_t)\big) + \sum_{t=1}^{T} \tilde{\eta}_{t-1} \Delta M_t,
\end{aligned}$$

where $\Delta M_t \coloneqq u_{t-1}(Z_t) - (P_Z u_{t-1})(Z_{t-1})$ is a martingale difference with respect to $\mathcal{F}_t \coloneqq \sigma(Z_0, \ldots, Z_t)$ since $u_{t-1} \in \mathcal{F}_{t-1}$ and

$$\mathbb{E}[u_{t-1}(Z_t) \mid \mathcal{F}_{t-1}] = \mathbb{E}[u_{t-1}(Z_t) \mid \mathcal{F}_{t-1}, u_{t-1}] = (P_Z u_{t-1})(Z_{t-1}) .$$

Thus, for all $T \geq 1$, we obtain

$$\tilde{B}_T = 2\tilde{M}_T + 2\tilde{R}_T, \qquad \tilde{M}_T \coloneqq \sum_{t=1}^{T} \tilde{\eta}_{t-1} \Delta M_t, \quad \tilde{R}_T \coloneqq \sum_{t=1}^{T} \tilde{\eta}_{t-1} \big(u_{t-1}(Z_{t-1}) - u_{t-1}(Z_t)\big), \tag{22}$$

and $(\tilde{M}_T)_{T \geq 1}$ is a real-valued martingale with respect to $(\mathcal{F}_T)_{T \geq 1}$.

Moreover, if $\tilde{\eta}_{t-1} > 0$, we have $\|\tilde{\boldsymbol{\theta}}_{t-1}\| \le R_{\max}$ and $\|\tilde{\boldsymbol{\theta}}_{t-1} - \boldsymbol{\theta}^*\| \le 2R_{\max}$. Thus, Lemma 9 implies that

$$\begin{aligned}\|u_{t-1}\|_\infty &\le 16\tau_{\text{mix}}\left(2r_\infty\phi_\infty + 4\phi_\infty^2\left\|\tilde{\boldsymbol{\theta}}_{t-1}\right\|\right)\left(\left\|\tilde{\boldsymbol{\theta}}_{t-1} - \boldsymbol{\theta}^*\right\|\right) \le 16\tau_{\text{mix}}\phi_\infty^2(4R_{\max}^2 + 8R_{\max}^2) \\ &\le 192\tau_{\text{mix}}\phi_\infty^2 R_{\max}^2\,.\end{aligned}$$

Also,

$$\|\Delta M_t\|_\infty \le \|u_{t-1}\|_\infty + \|P_Z u_{t-1}\|_\infty \le 2\|u_{t-1}\|_\infty \le 384\tau_{\text{mix}}\phi_\infty^2 R_{\max}^2\,.$$

Since $\tilde{\eta}_{t-1} = 0$ for all $t > T_{\text{exit}}$, the increments $\tilde{\eta}_{t-1}\Delta M_t$ satisfy

$$\|\tilde{\eta}_{t-1}\Delta M_t\|_\infty \le \tilde{\eta}_{t-1}384\tau_{\text{mix}}\phi_\infty^2 R_{\max}^2\,. \tag{23}$$

We now use Pinelis' inequality in Lemma 22. Applying it to the increments

$$X_t := \tilde{\eta}_{t-1}\Delta M_t, \qquad t \ge 1,$$

with (23), we obtain the following anytime high-probability guarantee.

Lemma 11 (Pinelis' inequality for the stopped martingale) *Let $\tilde{M}_T := \sum_{t=1}^T X_t$ with $X_t := \tilde{\eta}_{t-1}\Delta M_t$ as above. Then, for all $\delta \in (0,1)$,*

$$\mathbb{P}\left\{\sup_{T\ge 1}|\tilde{M}_T| \;\le\; 384\tau_{\text{mix}}\phi_\infty^2 R_{\max}^2\sqrt{\sum_{t=0}^{\infty}\eta_t^2}\sqrt{2\log\frac{2}{\delta}}\right\} \;\ge\; 1-\delta\,.$$

Proof Setting $r \ge Db_*\sqrt{2\log\frac{2}{\delta}}$ in Lemma 22, we have $2\exp\left(-\frac{r^2}{2D^2b_*^2}\right) \le \delta$. Since $\mathbb{R}^d$ with ℓ_2 is a $(2,1)$-smooth Banach space, choosing $b_*^2 = \sum_{t=0}^\infty \eta_t^2(384\tau_{\text{mix}}\phi_\infty^2 R_{\max}^2)^2$ finishes the proof. ■

B.3. Bound on the remainder term

We now bound the deterministic remainder $\tilde{R}_T$ defined in (22) using Lemma 10 and the Lipschitz continuity of u_{t-1}.

Lemma 12 (Bound on the localized remainder) *In the setting of Section 3, with the constants from Lemmas 2 and 3, for any $T \ge 1$,*

$$|\tilde{R}_T| \le 576\eta_0\tau_{\text{mix}}\phi_\infty^2 R_{\max}^2 + 672\tau_{\text{mix}}\phi_\infty^4 R_{\max}^2\sum_{t=1}^T\eta_{t-1}^2\,.$$

Proof Using Lemma 10, we have

$$\tilde{R}_T = \tilde{\eta}_0 u_0(Z_0) - \tilde{\eta}_T u_T(Z_T) - \sum_{t=1}^T(\tilde{\eta}_{t-1} - \tilde{\eta}_t)u_{t-1}(Z_t) - \sum_{t=1}^T\tilde{\eta}_t\big(u_{t-1}(Z_t) - u_t(Z_t)\big)\,.$$

If $\tilde{\eta}_{t-1} > 0$, we have $\|\tilde{\boldsymbol{\theta}}_{t-1}\| \le R_{\max}$, hence $\|\tilde{\boldsymbol{\theta}}_{t-1} - \boldsymbol{\theta}^*\| \le 2R_{\max}$. Lemma 9 implies

$$\|u_{t-1}\|_\infty \le 192\tau_{\text{mix}}\phi_\infty^2 R_{\max}^2\,.$$

Therefore

$$|\tilde{\eta}_0 u_0(Z_0)| + |\tilde{\eta}_T u_T(Z_T)| \le 384\eta_0 \tau_{\text{mix}} \phi_\infty^2 R_{\max}^2 \,.$$

Since (η_{t-1}) is non-increasing, we have $\sum_{t=1}^T |\tilde{\eta}_{t-1} - \tilde{\eta}_t| = \sum_{t=1}^T (\tilde{\eta}_{t-1} - \tilde{\eta}_t) \le \eta_0$, so

$$\left|\sum_{t=1}^T (\tilde{\eta}_{t-1} - \tilde{\eta}_t) u_{t-1}(Z_t)\right| \le 192\eta_0 \tau_{\text{mix}} \phi_\infty^2 R_{\max}^2 \,.$$

For the $u_t - u_{t-1}$ term, when t satisfies $\tilde{\eta}_t > 0$, we expand the TD recursion $\tilde{\boldsymbol{\theta}}_t - \tilde{\boldsymbol{\theta}}_{t-1} = \tilde{\eta}_{t-1} \boldsymbol{g}(\tilde{\boldsymbol{\theta}}_{t-1}, Z_{t-1})$ and use Lemma 9(iii) to conclude

$$\begin{aligned}
\|u_t - u_{t-1}\|_\infty &\le 16\tau_{\text{mix}} \left(2r_\infty \phi_\infty + 4\phi_\infty^2 \left\|\tilde{\boldsymbol{\theta}}_t\right\|\right) \left\|\tilde{\boldsymbol{\theta}}_t - \tilde{\boldsymbol{\theta}}_{t-1}\right\| + 64\tau_{\text{mix}} \phi_\infty^2 \left\|\tilde{\boldsymbol{\theta}}_{t-1} - \boldsymbol{\theta}^*\right\| \left\|\tilde{\boldsymbol{\theta}}_t - \tilde{\boldsymbol{\theta}}_{t-1}\right\| \\
&\le \tilde{\eta}_{t-1} 16\tau_{\text{mix}} (2r_\infty \phi_\infty + 4\phi_\infty^2 R_{\max})(r_\infty \phi_\infty + 2\phi_\infty^2 R_{\max}) \\
&\quad + 64\tilde{\eta}_{t-1} \tau_{\text{mix}} \phi_\infty^2 2R_{\max} (r_\infty \phi_\infty + 2\phi_\infty^2 R_{\max}) \\
&\le \tilde{\eta}_{t-1} \phi_\infty^4 R_{\max}^2 (16 \times 18 + 64 \times 2 \times 3) \tau_{\text{mix}} \\
&\le 672 \tilde{\eta}_{t-1} \tau_{\text{mix}} \phi_\infty^4 R_{\max}^2 \,.
\end{aligned}$$

Thus,

$$\begin{aligned}
\left|\sum_{t=1}^T \tilde{\eta}_t \big(u_{t-1}(Z_t) - u_t(Z_t)\big)\right| &\le \sum_{t=1}^T \tilde{\eta}_t \|u_t - u_{t-1}\|_\infty \le 672\tau_{\text{mix}} \phi_\infty^4 R_{\max}^2 \sum_{t=1}^T \tilde{\eta}_t \tilde{\eta}_{t-1} \\
&\le 672\tau_{\text{mix}} \phi_\infty^4 R_{\max}^2 \sum_{t=1}^T \eta_{t-1}^2 \,.
\end{aligned}$$

Combining the three bounds, we have

$$|\tilde{R}_T| \le 576\eta_0 \tau_{\text{mix}} \phi_\infty^2 R_{\max}^2 + 672\tau_{\text{mix}} \phi_\infty^4 R_{\max}^2 \sum_{t=1}^T \eta_{t-1}^2 \,.$$

■

B.4. High-probability bound on the localized Markov noise

Combining Lemmas 11 and 12 with the decomposition (22) yields the following high-probability control of $\tilde{B}_T$.

Lemma 13 (Localized high-probability bound for B_T) *In the setting of Section 3, with the constants from Lemmas 2 and 3, for any $\delta \in (0, 1)$, we have*

$$\mathbb{P}\left\{\frac{1}{2} \sup_{T \ge 1} |\tilde{B}_T| \;\le\; 384\tau_{\text{mix}} \phi_\infty^2 R_{\max}^2 \sqrt{\sum_{t=0}^\infty \eta_t^2} \sqrt{2\log\frac{2}{\delta}} + C\right\} \;\ge\; 1 - \delta,$$

where $C = 576\eta_0 \tau_{\text{mix}} \phi_\infty^2 R_{\max}^2 + 672\tau_{\text{mix}} \phi_\infty^4 R_{\max}^2 \sum_{t=0}^\infty \eta_t^2$.

Proof From (22), $\tilde{B}_T = 2\tilde{M}_T + 2\tilde{R}_T$. By Lemma 11, with probability at least $1-\delta$,

$$\sup_{T\geq 1} |\tilde{M}_T| \leq 384\tau_{\mathrm{mix}}\phi_\infty^2 R_{\mathrm{max}}^2 \sqrt{\sum_{t=0}^{\infty} \eta_t^2}\sqrt{2\log\frac{2}{\delta}}\,.$$

By Lemma 12, we also have the deterministic bound

$$\sup_{T\geq 1} |\tilde{R}_T| \leq 576\eta_0\tau_{\mathrm{mix}}\phi_\infty^2 R_{\mathrm{max}}^2 + 672\tau_{\mathrm{mix}}\phi_\infty^4 R_{\mathrm{max}}^2 \sum_{t=0}^{\infty}\eta_t^2\,.$$

Combining these two bounds yields the desired result. ■

B.5. High-probability bounded iterates via bootstrap

We can now bootstrap on R_{max} to prove high-probability bounded iterates, completing the proof of Theorem 6.

Lemma 14 (Bootstrap inequality for the radius) *In the setting of Section 3, with the constants from Lemmas 2 and 3, for all $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\begin{aligned}\sup_{T\geq 0} \|\tilde{\boldsymbol{e}}_T\|^2 \;\leq\; &\|\boldsymbol{e}_0\|^2 + 768\tau_{\mathrm{mix}}\phi_\infty^2 R_{\mathrm{max}}^2 \sqrt{\sum_{t=0}^{\infty}\eta_t^2}\sqrt{2\log\frac{2}{\delta}} + 1152\eta_0\tau_{\mathrm{mix}}\phi_\infty^2 R_{\mathrm{max}}^2 \\ &+ 1344\tau_{\mathrm{mix}}\phi_\infty^4 R_{\mathrm{max}}^2 \sum_{t=0}^{\infty}\eta_t^2 + 9\phi_\infty^4 R_{\mathrm{max}}^2 \sum_{t=0}^{\infty}\eta_t^2,\end{aligned}$$

where $\tilde{\boldsymbol{e}}_T := \tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*$.

Proof Using inequality (20) and dropping the nonpositive $2\sum_{t=1}^{T}\eta_{t-1}\langle \bar{\boldsymbol{g}}(\boldsymbol{\theta}_{t-1}), \boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^*\rangle$ term by (4), we obtain

$$\|\tilde{\boldsymbol{e}}_T\|^2 \leq \|\boldsymbol{e}_0\|^2 + \sum_{t=1}^{T}\tilde{\eta}_{t-1}^2 \|\boldsymbol{g}(\tilde{\boldsymbol{\theta}}_{t-1}, Z_{t-1})\|^2 + \tilde{B}_T\,.$$

Lemma 2 implies

$$\|\boldsymbol{g}(\tilde{\boldsymbol{\theta}}_{t-1}, Z_{t-1})\| \leq r_\infty\phi_\infty + 2\phi_\infty^2 R_{\mathrm{max}} \leq 3\phi_\infty^2 R_{\mathrm{max}}\,.$$

Thus,

$$\sum_{t=1}^{T}\tilde{\eta}_{t-1}^2 \|\boldsymbol{g}(\tilde{\boldsymbol{\theta}}_{t-1}, Z_{t-1})\|^2 \leq 9\phi_\infty^4 R_{\mathrm{max}}^2 \sum_{t=0}^{\infty}\eta_t^2\,.$$

Combining this with Lemma 13 finishes the proof. ■

We are now ready to give the proof of Theorem 6.

Proof By the triangle inequality,

$$\begin{aligned}
\sup_{T\geq 0}\ \|\tilde{\boldsymbol{\theta}}_T\|^2 &\leq 2\|\boldsymbol{\theta}^*\|^2 + 2\sup_T\ \|\tilde{\boldsymbol{e}}_T\|^2 \\
&\leq 2\|\boldsymbol{\theta}^*\|^2 + 2\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_0\|^2 + 1536\tau_{\text{mix}}\phi_\infty^2 R_{\max}^2 \sqrt{\sum_{t=0}^{\infty}\eta_t^2}\sqrt{2\log\frac{2}{\delta}} \\
&\quad + 2304\eta_0\tau_{\text{mix}}\phi_\infty^2 R_{\max}^2 + (2688\tau_{\text{mix}} + 18)\phi_\infty^4 R_{\max}^2 \sum_{t=0}^{\infty}\eta_t^2 \\
&\leq R_{\text{base}}^2\left(2 + 2 + 1536\tau_{\text{mix}}\phi_\infty^2\rho^2\sqrt{\sum_{t=0}^{\infty}\eta_t^2}\sqrt{2\log\frac{2}{\delta}}\right) \\
&\quad + R_{\text{base}}^2\left(2304\eta_0\tau_{\text{mix}}\phi_\infty^2\rho^2 + (2688\tau_{\text{mix}} + 18)\phi_\infty^4\rho^2\sum_{t=0}^{\infty}\eta_t^2\right)\ .
\end{aligned}$$

Recall that $\eta_{\text{base}} = \frac{1}{c\tau_{\text{mix}}\phi_\infty^2}$ and $\sqrt{\sum_{t=0}^{\infty}\eta_t^2} = \frac{1}{c\tau_{\text{mix}}\phi_\infty^2}\sqrt{\sum_{t=0}^{\infty}a_t^2}$. It follows that

$$\sup_{T\geq 0}\ \|\tilde{\boldsymbol{\theta}}_T\|^2 \leq R_{\text{base}}^2\left(4 + 1536\frac{\rho^2}{c}\sqrt{\sum_{t=0}^{\infty}a_t^2}\sqrt{2\log\frac{2}{\delta}} + 2304\frac{\rho^2}{c} + 2706\left(\sum_{t=0}^{\infty}a_t^2\right)\frac{\rho^2}{c^2}\right)\ .$$

There exists a sufficiently large c such that

$$4 + 1536\frac{\rho^2}{c}\sqrt{\sum_{t=0}^{\infty}a_t^2}\sqrt{2\log\frac{2}{\delta}} + 2304\frac{\rho^2}{c} + 2706\left(\sum_{t=0}^{\infty}a_t^2\right)\frac{\rho^2}{c^2} \leq \rho^2.$$

For such c, we have

$$\sup_{T\geq 0}\ \|\tilde{\boldsymbol{\theta}}_T\|^2 \leq \rho^2 R_{\text{base}}^2 = R_{\max}^2\ .$$

On the event $\{\sup_T\ \|\tilde{\boldsymbol{\theta}}_T\|^2 \leq R_{\max}^2\}$ we must have $T_{\text{exit}} = \infty$ by the definition of T_{exit}. Thus $\tilde{\boldsymbol{\theta}}_T = \boldsymbol{\theta}_T$ for all T, and

$$\sup_{T\geq 0}\ \|\boldsymbol{\theta}_T\|^2 \leq R_{\max}^2\ .$$

■

Appendix C. Convergence Rates: Proof of Theorem 7

Recall that the TD update can be written as

$$\boldsymbol{g}(\boldsymbol{\theta}_{t-1}, Z_{t-1}) = -\boldsymbol{A}_{Z_{t-1}}\boldsymbol{\theta}_{t-1} + \boldsymbol{b}_{Z_{t-1}}.$$

Therefore, the centered error $\boldsymbol{e}_t := \boldsymbol{\theta}_t - \boldsymbol{\theta}^*$ evolves as

$$\begin{aligned}\boldsymbol{e}_t &= \boldsymbol{e}_{t-1} + \eta_{t-1}\big(\boldsymbol{b}_{Z_{t-1}} - \boldsymbol{A}_{Z_{t-1}}\boldsymbol{\theta}_{t-1}\big)\\ &= \boldsymbol{e}_{t-1} + \eta_{t-1}\big(-\boldsymbol{A}_{Z_{t-1}}(\boldsymbol{\theta}^* + \boldsymbol{e}_{t-1}) + \boldsymbol{b}_{Z_{t-1}}\big)\\ &= \boldsymbol{e}_{t-1} + \eta_{t-1}\big(-\boldsymbol{A}_{Z_{t-1}}\boldsymbol{e}_{t-1} + \boldsymbol{\xi}_{t-1}\big)\\ &= (\mathbf{I}_d - \eta_{t-1}\boldsymbol{A})\boldsymbol{e}_{t-1} + \eta_{t-1}\boldsymbol{\xi}_{t-1} + \eta_{t-1}\boldsymbol{\delta}_{t-1}\boldsymbol{e}_{t-1}, \qquad (24)\end{aligned}$$

using $\boldsymbol{A}_{Z_{t-1}} = \boldsymbol{A} - \boldsymbol{\delta}_{t-1}$ and $\boldsymbol{\xi}_{t-1} = \boldsymbol{b}_{Z_{t-1}} - \boldsymbol{A}_{Z_{t-1}}\boldsymbol{\theta}^*$. Rearranging yields

$$\boldsymbol{A}\boldsymbol{e}_{t-1} = \frac{\boldsymbol{e}_{t-1} - \boldsymbol{e}_t}{\eta_{t-1}} + \boldsymbol{\xi}_{t-1} + \boldsymbol{\delta}_{t-1}\boldsymbol{e}_{t-1}\,. \qquad (25)$$

Define the following quantities:

$$S_T := \sum_{t=1}^{T} \eta_{t-1}, \qquad \bar{\boldsymbol{e}}_T := \frac{1}{S_T}\sum_{t=1}^{T} \eta_{t-1}\boldsymbol{e}_{t-1}, \qquad \bar{\boldsymbol{\theta}}_T := \boldsymbol{\theta}^* + \bar{\boldsymbol{e}}_T\,.$$

Multiplying (25) by η_{t-1} and summing from $t = 1$ to T yields

$$\sum_{t=1}^{T} \eta_{t-1}\boldsymbol{A}\boldsymbol{e}_{t-1} = \sum_{t=1}^{T}(\boldsymbol{e}_{t-1} - \boldsymbol{e}_t) + \sum_{t=1}^{T}\eta_{t-1}\boldsymbol{\xi}_{t-1} + \sum_{t=1}^{T}\eta_{t-1}\boldsymbol{\delta}_{t-1}\boldsymbol{e}_{t-1}\,.$$

Dividing by S_T and recalling the definition of $\bar{\boldsymbol{e}}_T$ yields

$$\boldsymbol{A}\bar{\boldsymbol{e}}_T = I_1 + I_2 + I_3, \qquad (26)$$

where

$$I_1 := \frac{1}{S_T}\sum_{t=1}^{T}(\boldsymbol{e}_{t-1} - \boldsymbol{e}_t), \quad I_2 := \frac{1}{S_T}\sum_{t=1}^{T}\eta_{t-1}\boldsymbol{\xi}_{t-1}, \quad I_3 := \frac{1}{S_T}\sum_{t=1}^{T}\eta_{t-1}\boldsymbol{\delta}_{t-1}\boldsymbol{e}_{t-1}\,.$$

Notice that, by Lemma 4, $(\boldsymbol{A} + \boldsymbol{A}^\top)/2 \succeq \omega\mathbf{I}_d$, so for any $\boldsymbol{x} \in \mathbb{R}^d$,

$$\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x}\rangle = \boldsymbol{x}^\top \frac{\boldsymbol{A} + \boldsymbol{A}^\top}{2}\boldsymbol{x} \geq \omega\|\boldsymbol{x}\|^2\,.$$

Combining this with Cauchy–Schwarz,

$$\omega\|\boldsymbol{x}\|^2 \leq \langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x}\rangle \leq \|\boldsymbol{A}\boldsymbol{x}\|\|\boldsymbol{x}\|,$$

we obtain, for $\boldsymbol{x} \neq 0$,

$$\|\boldsymbol{x}\| \leq \frac{1}{\omega}\|\boldsymbol{A}\boldsymbol{x}\|.$$

In particular, for $\boldsymbol{x} = \bar{\boldsymbol{e}}_T$,

$$f(\bar{\boldsymbol{\theta}}_T) - f(\boldsymbol{\theta}^*) = \langle \boldsymbol{A}\bar{\boldsymbol{e}}_T, \bar{\boldsymbol{e}}_T\rangle \leq \|\,\boldsymbol{A}\bar{\boldsymbol{e}}_T\|_2\,\|\bar{\boldsymbol{e}}_T\| \leq \frac{1}{\omega}\,\|\boldsymbol{A}\bar{\boldsymbol{e}}_T\|^2\,. \qquad (27)$$

We now bound $\|I_1\|$, $\|I_2\|$, and $\|I_3\|$ in turn to control $f(\bar{\boldsymbol{\theta}}_T) - f(\boldsymbol{\theta}^*)$.

C.1. Bound on the telescoping term I_1

The first term is purely deterministic.

Lemma 15 (Bound on I_1) *On the bounded-iterates event $\mathcal{E}_R := \{\sup_t \|\boldsymbol{\theta}_t\| \le R_{\max}\}$ we have,*

$$\|I_1\|_2 \le \frac{2R_{\max}}{\sum_{t=1}^T \eta_{t-1}}$$

for all $T \ge 1$.

Proof By telescoping,

$$\sum_{t=1}^T (\boldsymbol{e}_{t-1} - \boldsymbol{e}_t) = \boldsymbol{e}_0 - \boldsymbol{e}_T,$$

so

$$I_1 = \frac{\boldsymbol{e}_0 - \boldsymbol{e}_T}{S_T}.$$

On $\mathcal{E}_R$, $\|\boldsymbol{e}_0 - \boldsymbol{e}_T\| = \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^* - \boldsymbol{\theta}_T + \boldsymbol{\theta}^*\| \le 2R_{\max}$. Therefore, $\|I_1\| \le \frac{2R_{\max}}{\sum_{t=1}^T \eta_{t-1}}$. ■

C.2. Bound on I_2 via vector-valued Martingale Inequality

Lemma 16 (Bounds for I_2) *In the setting of Section 3, with the constants from Lemmas 2 and 3, let*

$$I_2 := \frac{1}{S_T} \sum_{t=1}^T \eta_{t-1} \boldsymbol{\xi}_{t-1},$$

where $\boldsymbol{\xi}_{t-1} := \boldsymbol{\xi}(Z_{t-1})$ and the stepsize sequence (η_{t-1}) is non-increasing. Then, with probability at least $1-\delta$, for all $T \ge 1$, we have

$$\|I_2\| \le \frac{1}{\sum_{t=1}^T \eta_{t-1}} \left(2\sqrt{\sum_{t=0}^\infty \eta_t^2} \sqrt{2\log\frac{2}{\delta}} + 3\eta_0 \right) 16\tau_{\text{mix}}(r_\infty \phi_\infty + 2\phi_\infty^2 \|\boldsymbol{\theta}^*\|)\,.$$

Proof By Lemma 8, we have

$$u - P_Z u = \boldsymbol{\xi}, \quad \text{and} \quad \|u\|_\infty \le 16\tau_{\text{mix}}(r_\infty \phi_\infty + 2\phi_\infty^2 \|\boldsymbol{\theta}^*\|)\,.$$

Then, reasoning as in Section A.3, we obtain

$$\begin{aligned}
\sum_{t=1}^T \eta_{t-1}\boldsymbol{\xi}_{t-1} &= \sum_{t=1}^T \eta_{t-1} \left(u(Z_{t-1}) - (P_Z u)(Z_{t-1})\right) \\
&= \sum_{t=1}^T \eta_{t-1} \left(u(Z_{t-1}) - u(Z_t) + u(Z_t) - (P_Z u)(Z_{t-1})\right) \\
&= M_T + R_T,
\end{aligned}$$

where

$$M_T := \sum_{t=1}^{T} \eta_{t-1} \Delta M_t, \quad R_T := \sum_{t=1}^{T} \eta_{t-1} \left(u(Z_{t-1}) - u(Z_t) \right) .$$

Since $\|u\|_\infty \leq 16\tau_{\text{mix}}(r_\infty \phi_\infty + 2\phi_\infty^2 \|\boldsymbol{\theta}^*\|)$, we have

$$\|\Delta M_t\|_\infty \leq \|u\|_\infty + \left\|P_Z u\right\|_\infty \leq 2\|u\|_\infty \leq 32\tau_{\text{mix}}(r_\infty \phi_\infty + 2\phi_\infty^2 \|\boldsymbol{\theta}^*\|) .$$

Using Lemma 22, for any $\delta \in (0, 1)$, we have with probability at least $1 - \delta$,

$$\sup_{T \geq 1} \|M_T\| \leq 32\tau_{\text{mix}}(r_\infty \phi_\infty + 2\phi_\infty^2 \|\boldsymbol{\theta}^*\|_2) \sqrt{\sum_{t=0}^{\infty} \eta_t^2} \sqrt{2 \log \frac{2}{\delta}} .$$

For the remainder term R_T, we have

$$\|R_T\| = \left\| \sum_{t=1}^{T} \eta_{t-1} \big(u(Z_{t-1}) - u(Z_t) \big) \right\| \leq 3\eta_0 \|u\|_\infty \leq 48\tau_{\text{mix}}(r_\infty \phi_\infty + 2\phi_\infty^2 \|\boldsymbol{\theta}^*\|)\eta_0,$$

since η_{t-1} is non-increasing. Thus, with probability at least $1 - \delta$,

$$\sup_{T \geq 1} \|M_T + R_T\| \leq \left(2\sqrt{\sum_{t=0}^{\infty} \eta_t^2} \sqrt{2 \log \frac{2}{\delta}} + 3\eta_0 \right) 16\tau_{\text{mix}}(r_\infty \phi_\infty + 2\phi_\infty^2 \|\boldsymbol{\theta}^*\|) .$$

Hence, with probability at least $1 - \delta$,

$$\|I_2\| = \frac{\|M_T + R_T\|_2}{S_T} \leq \frac{1}{\sum_{t=1}^{T} \eta_{t-1}} \left(2\sqrt{\sum_{t=0}^{\infty} \eta_t^2} \sqrt{2 \log \frac{2}{\delta}} + 3\eta_0 \right) 16\tau_{\text{mix}}(r_\infty \phi_\infty + 2\phi_\infty^2 \|\boldsymbol{\theta}^*\|) .$$

■

C.3. Bound on I_3 via localized vector-valued Martingale Inequality

Recall the stopping time and stopped process:

$$T_{\text{exit}} := \inf\{t \geq 0 : \|\boldsymbol{\theta}_t\| > R_{\max}\}, \qquad \tilde{\boldsymbol{\theta}}_t := \boldsymbol{\theta}_{t \wedge T_{\text{exit}}}, \qquad \tilde{\boldsymbol{e}}_t := \tilde{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*,$$

and the stopped stepsizes

$$\tilde{\eta}_t := \begin{cases} \eta_t, & t < T_{\text{exit}}, \\ 0, & t \geq T_{\text{exit}} . \end{cases}$$

For $T \geq 1$, define the localized multiplicative-noise term

$$\tilde{I}_{3,T} := \frac{1}{S_T} \sum_{t=1}^{T} \tilde{\eta}_{t-1} \boldsymbol{\delta}_{t-1} \tilde{\boldsymbol{e}}_{t-1} . \tag{28}$$

On the event $\{T_{\text{exit}} > T - 1\}$ we have $\tilde{\eta}_{t-1} = \eta_{t-1}$ and $\tilde{\boldsymbol{e}}_{t-1} = \boldsymbol{e}_{t-1}$ for all $1 \leq t \leq T$, so $\tilde{I}_{3,T} = I_3$.

Lemma 17 (Localized bounds for I_3) *In the setting of Section 3, with the constants from Lemmas 2 and 3, consider the stopped process defined above. Fix $\delta \in (0,1)$. We have, with probability at least $1-\delta$, for any $T \geq 1$,*

$$\|\tilde{I}_{3,T}\|_2 \leq \frac{256\tau_{\text{mix}}\phi_\infty^2 R_{\max}\sqrt{\sum_{t=0}^\infty \eta_t^2}\sqrt{2\log\frac{2}{\delta}} + 384\eta_0\tau_{\text{mix}}\phi_\infty^2 R_{\max}}{S_T} + \frac{192\tau_{\text{mix}}\phi_\infty^4 R_{\max}\sum_{t=0}^\infty \eta_t^2}{S_T}\,.$$

Proof Recall

$$\boldsymbol{\delta}_z = \boldsymbol{A} - \boldsymbol{A}_z, \qquad h_{\tilde{\boldsymbol{\theta}}}(z) := \boldsymbol{\delta}_z(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*),$$

and $\|\boldsymbol{\delta}_z\|_{\text{op}} \leq 4\phi_\infty^2$ for all z. Thus, for each $\boldsymbol{\theta}$, there exists a Poisson solution $u_{\boldsymbol{\theta}} : \mathcal{Z} \to \mathbb{R}^d$ satisfying

$$u_{\boldsymbol{\theta}} - P_Z u_{\boldsymbol{\theta}} = h_{\boldsymbol{\theta}},$$

and Lemma 9 gives

$$\|u_{\boldsymbol{\theta}}\|_\infty \leq 64\tau_{\text{mix}}\phi_\infty^2\,\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|\,, \qquad \|u_{\boldsymbol{\theta}} - u_{\boldsymbol{\theta}'}\|_\infty \leq 64\tau_{\text{mix}}\phi_\infty^2\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|\,.$$

Reasoning as in Section A.3 with $h_{\tilde{\boldsymbol{\theta}}_{t-1}}$ yields, for all $T \geq 1$,

$$\begin{aligned}\sum_{t=1}^T \tilde{\eta}_{t-1}\boldsymbol{\delta}_{t-1}\tilde{\boldsymbol{e}}_{t-1} &= \sum_{t=1}^T \tilde{\eta}_{t-1}h_{\tilde{\boldsymbol{\theta}}_{t-1}}(Z_{t-1}) = \sum_{t=1}^T \tilde{\eta}_{t-1}\left(u_{\tilde{\boldsymbol{\theta}}_{t-1}}(Z_{t-1}) - \left(P_Z u_{\tilde{\boldsymbol{\theta}}_{t-1}}\right)(Z_{t-1})\right)\\ &= \sum_{t=1}^T \tilde{\eta}_{t-1}\left(u_{\tilde{\boldsymbol{\theta}}_{t-1}}(Z_{t-1}) - u_{\tilde{\boldsymbol{\theta}}_{t-1}}(Z_t) + u_{\tilde{\boldsymbol{\theta}}_{t-1}}(Z_t) - \left(P_Z u_{\tilde{\boldsymbol{\theta}}_{t-1}}\right)(Z_{t-1})\right)\,.\end{aligned}$$

Let $\Delta M_t := u_{\tilde{\boldsymbol{\theta}}_{t-1}}(Z_t) - \left(P_Z u_{\tilde{\boldsymbol{\theta}}_{t-1}}\right)(Z_{t-1})$. We have $\sum_{t=1}^T \tilde{\eta}_{t-1}\boldsymbol{\delta}_{t-1}\tilde{\boldsymbol{e}}_{t-1} = \tilde{M}_T + \tilde{R}_T$, where

$$\begin{aligned}\tilde{M}_T &:= \sum_{t=1}^T \tilde{\eta}_{t-1}\Delta M_t,\\ \tilde{R}_T &:= \sum_{t=1}^T \tilde{\eta}_{t-1}\big(u_{\tilde{\boldsymbol{\theta}}_{t-1}}(Z_{t-1}) - u_{\tilde{\boldsymbol{\theta}}_{t-1}}(Z_t)\big)\,.\end{aligned}$$

If $\tilde{\eta}_{t-1} > 0$, then $\left\|\tilde{\boldsymbol{\theta}}_{t-1}\right\| \leq R_{\max}$ and we have

$$\|\tilde{\eta}_{t-1}\Delta M_t\|_2 \leq 2\tilde{\eta}_{t-1}\left\|u_{\tilde{\boldsymbol{\theta}}_{t-1}}\right\|_\infty \leq 256\tilde{\eta}_{t-1}\tau_{\text{mix}}\phi_\infty^2 R_{\max}\,.$$

Using Lemma 22, for any $\delta \in (0,1)$, we have with probability at least $1-\delta$,

$$\sup_{T\geq 1}\left\|\tilde{M}_T\right\|_2 \leq 256\tau_{\text{mix}}\phi_\infty^2 R_{\max}\sqrt{\sum_{t=0}^\infty \eta_t^2}\sqrt{2\log\frac{2}{\delta}}\,.$$

For the remainder term $\tilde{R}_T$, using Lemma 10, we have

$$\tilde{R}_T = \tilde{\eta}_0 u_{\tilde{\boldsymbol{\theta}}_0}(Z_0) - \tilde{\eta}_T u_{\tilde{\boldsymbol{\theta}}_T}(Z_T) - \sum_{t=1}^{T}(\tilde{\eta}_{t-1} - \tilde{\eta}_t) u_{\tilde{\boldsymbol{\theta}}_{t-1}}(Z_t) - \sum_{t=1}^{T} \tilde{\eta}_t \big(u_{\tilde{\boldsymbol{\theta}}_{t-1}}(Z_t) - u_{\tilde{\boldsymbol{\theta}}_t}(Z_t)\big)\,.$$

Since $\eta_0 \geq \eta_{T-1}$, we obtain

$$\|\tilde{\eta}_0 u_{\tilde{\boldsymbol{\theta}}_0}(Z_0)\|_2 + \|\tilde{\eta}_T u_{\tilde{\boldsymbol{\theta}}_T}(Z_T)\|_2 \leq 256 \eta_0 \tau_{\mathrm{mix}} \phi_\infty^2 R_{\max}\,.$$

For the second term, we have

$$\left\|\sum_{t=1}^{T}(\tilde{\eta}_{t-1} - \tilde{\eta}_t) u_{\tilde{\boldsymbol{\theta}}_{t-1}}(Z_t)\right\|_2 \leq 128 \tau_{\mathrm{mix}} \phi_\infty^2 R_{\max} \sum_{t=1}^{T} |\tilde{\eta}_{t-1} - \tilde{\eta}_t|\,.$$

Since the stopped stepsizes $(\tilde{\eta}_t)$ are non-increasing and satisfy $\sum_{t=1}^{T}(\tilde{\eta}_{t-1} - \tilde{\eta}_t) = \tilde{\eta}_0 - \tilde{\eta}_T \leq \eta_0$, we have

$$\left\|\sum_{t=1}^{T}(\tilde{\eta}_{t-1} - \tilde{\eta}_t) u_{\tilde{\boldsymbol{\theta}}_{t-1}}(Z_t)\right\|_2 \leq 128 \eta_0 \tau_{\mathrm{mix}} \phi_\infty^2 R_{\max}\,.$$

For the last term, by the Lipschitz property of $u_{\tilde{\boldsymbol{\theta}}}$ in part (iii) of Lemma 9, we have

$$\|u_{\tilde{\boldsymbol{\theta}}_{t-1}} - u_{\tilde{\boldsymbol{\theta}}_t}\|_\infty \leq 64 \tau_{\mathrm{mix}} \phi_\infty^2 \|\tilde{\boldsymbol{\theta}}_{t-1} - \tilde{\boldsymbol{\theta}}_t\|_2\,.$$

Expanding the TD recursion, we have

$$\|\tilde{\boldsymbol{\theta}}_t - \tilde{\boldsymbol{\theta}}_{t-1}\|_2 = \tilde{\eta}_{t-1} \|\boldsymbol{g}(\tilde{\boldsymbol{\theta}}_{t-1}, Z_{t-1})\|_2 \leq \tilde{\eta}_{t-1}\big(r_\infty + 2\phi_\infty \|\tilde{\boldsymbol{\theta}}_{t-1}\|_2\big)\phi_\infty \leq 3 \tilde{\eta}_{t-1} \phi_\infty^2 R_{\max}\,.$$

Therefore,

$$\begin{aligned}
\left\|\sum_{t=1}^{T} \tilde{\eta}_t \big(u_{\tilde{\boldsymbol{\theta}}_{t-1}}(Z_t) - u_{\tilde{\boldsymbol{\theta}}_t}(Z_t)\big)\right\|_2 &\leq \sum_{t=1}^{T} \tilde{\eta}_t \|u_{\tilde{\boldsymbol{\theta}}_{t-1}} - u_{\tilde{\boldsymbol{\theta}}_t}\|_\infty \\
&\leq 192 \tau_{\mathrm{mix}} \phi_\infty^4 R_{\max} \sum_{t=1}^{T} \tilde{\eta}_t \tilde{\eta}_{t-1} \\
&\leq 192 \tau_{\mathrm{mix}} \phi_\infty^4 R_{\max} \sum_{t=0}^{\infty} \eta_t^2,
\end{aligned}$$

since $(\tilde{\eta}_t)$ is non-increasing. Combining all the estimates finishes the proof. ■

C.4. Proof of Theorem 7

Proof Fix $\delta \in (0,1)$ and choose (c, ρ) with $\rho > 2$ satisfying

$$4 + 1536 \frac{\rho^2}{c} \sqrt{\sum_{t=0}^{\infty} a_t^2} \sqrt{2 \log \frac{8}{\delta}} + 2304 \frac{\rho^2}{c} + 2706 \left(\sum_{t=0}^{\infty} a_t^2\right) \frac{\rho^2}{c^2} \leq \rho^2\,.$$

Then, with probability at least $1-\delta/4$, running TD(0) with the stepsize schedule (η_t) defined in Theorem 6 guarantees

$$\sup_{t\geq 0} \|\boldsymbol{\theta}_t\| \leq \rho \max\left\{\|\boldsymbol{\theta}_0-\boldsymbol{\theta}^*\|, \|\boldsymbol{\theta}^*\|, \frac{r_\infty}{\phi_\infty}\right\},$$

where we define the corresponding event $\mathcal{E}_R := \left\{\sup_{t\geq 0} \|\boldsymbol{\theta}_t\| \leq \rho \max\left\{\|\boldsymbol{\theta}_0-\boldsymbol{\theta}^*\|, \|\boldsymbol{\theta}^*\|, \frac{r_\infty}{\phi_\infty}\right\}\right\}$.

Next, we apply Lemma 16 with confidence parameter $\delta/4$ to obtain an event $\mathcal{E}_2$ such that

$$\mathbb{P}\{\mathcal{E}_2\} \geq 1-\frac{\delta}{4}, \text{ and on } \mathcal{E}_2: \|I_2\| \leq \frac{2}{S_T}\left(2\sqrt{\sum_{t=0}^{\infty}\eta_t^2}\sqrt{2\log\frac{8}{\delta}}+3\eta_0\right)8\tau_{\text{mix}}(r_\infty\phi_\infty+2\phi_\infty^2\|\boldsymbol{\theta}^*\|).$$

Similarly, apply Lemma 17 with the same $\delta/4$ to obtain an event $\mathcal{E}_3$ such that

$$\mathbb{P}\{\mathcal{E}_3\} \geq 1-\frac{\delta}{4}, \qquad \text{and on } \mathcal{E}_3:$$

$$\|\tilde{I}_{3,T}\| \leq \frac{2}{S_T}\left(128\tau_{\text{mix}}\phi_\infty^2 R_{\max}\sqrt{\sum_{t=0}^{\infty}\eta_t^2}\sqrt{2\log\frac{8}{\delta}}+192\eta_0\tau_{\text{mix}}\phi_\infty^2 R_{\max}\right) + \frac{2}{S_T}\left(96\tau_{\text{mix}}\phi_\infty^4 R_{\max}\sum_{t=0}^{\infty}\eta_t^2\right).$$

Moreover, from the proof of Lemma 14, we can also obtain an event $\mathcal{E}_4$ such that

$$\mathbb{P}\{\mathcal{E}_4\} \geq 1-\frac{\delta}{4}, \qquad \text{and on } \mathcal{E}_4:$$

$$\sum_{t=1}^{T}\tilde{\eta}_{t-1}^2\left\|\boldsymbol{g}(\tilde{\boldsymbol{\theta}}_{t-1},Z_{t-1})\right\|^2+\tilde{B}_T \leq 768\tau_{\text{mix}}\phi_\infty^2 R_{\max}^2\sqrt{\sum_{t=0}^{\infty}\eta_t^2}\sqrt{2\log\frac{8}{\delta}}+1152\eta_0\tau_{\text{mix}}\phi_\infty^2 R_{\max}^2 + 1344\tau_{\text{mix}}\phi_\infty^4 R_{\max}^2\sum_{t=0}^{\infty}\eta_t^2+9\phi_\infty^4 R_{\max}^2\sum_{t=0}^{\infty}\eta_t^2.$$

In particular, if we choose $R_{\max} = \rho\max\left\{\|\boldsymbol{\theta}_0-\boldsymbol{\theta}^*\|, \|\boldsymbol{\theta}^*\|, \frac{r_\infty}{\phi_\infty}\right\}$, we have $T_{\text{exit}} = \infty$ on $\mathcal{E}_R$. Hence, for all $T \geq 1$, $\tilde{\eta}_{t-1} = \eta_{t-1}$ and $\tilde{\boldsymbol{e}}_{t-1} = \boldsymbol{e}_{t-1}$ for $1 \leq t \leq T$, so $\tilde{I}_{3,T} = I_3$ and $\sum_{t=1}^{T}\tilde{\eta}_{t-1}^2\left\|\boldsymbol{g}(\tilde{\boldsymbol{\theta}}_{t-1},Z_{t-1})\right\|^2+\tilde{B}_T = \sum_{t=1}^{T}\eta_{t-1}^2\|\boldsymbol{g}(\boldsymbol{\theta}_{t-1},Z_{t-1})\|^2+B_T$. Define

$$\mathcal{E} := \mathcal{E}_R\cap\mathcal{E}_2\cap\mathcal{E}_3\cap\mathcal{E}_4.$$

By a union bound, we have

$$\mathbb{P}\{\mathcal{E}\} \geq 1-\left(\frac{\delta}{4}+\frac{\delta}{4}+\frac{\delta}{4}+\frac{\delta}{4}\right) = 1-\delta.$$

Moreover, on $\mathcal{E}$,

$$\begin{aligned}
\|\boldsymbol{A}\bar{\boldsymbol{e}}_T\| &= \|I_1 + I_2 + I_3\| \le \|I_1\| + \|I_2\| + \|I_3\| \\
&\le \frac{2R_{\max}}{\sum_{t=1}^{T}\eta_{t-1}} + \frac{2}{\sum_{t=1}^{T}\eta_{t-1}}\left(2\sqrt{\sum_{t=0}^{\infty}\eta_t^2}\sqrt{2\log\frac{8}{\delta}} + 3\eta_0\right)8\tau_{\mathrm{mix}}(r_\infty\phi_\infty + 2\phi_\infty^2\|\boldsymbol{\theta}^*\|) \\
&\quad + \frac{2}{\sum_{t=1}^{T}\eta_{t-1}}\left(128\tau_{\mathrm{mix}}\phi_\infty^2 R_{\max}\sqrt{\sum_{t=0}^{\infty}\eta_t^2}\sqrt{2\log\frac{8}{\delta}}\right. \\
&\qquad\left. + 192\eta_0\tau_{\mathrm{mix}}\phi_\infty^2 R_{\max} + 96\tau_{\mathrm{mix}}\phi_\infty^4 R_{\max}\sum_{t=0}^{\infty}\eta_t^2\right).
\end{aligned}$$

Thus, using (27), on $\mathcal{E}$, we have

$$f(\bar{\boldsymbol{\theta}}_T) - f(\boldsymbol{\theta}^*) \le \frac{C_{\mathrm{fast}}^2}{\omega(\sum_{t=1}^{T}\eta_{t-1})^2},$$

where

$$\begin{aligned}
C_{\mathrm{fast}} &:= 2R_{\max} + \left(2\sqrt{\sum_{t=0}^{\infty}\eta_t^2}\sqrt{2\log\frac{8}{\delta}} + 3\eta_0\right)48\tau_{\mathrm{mix}}\phi_\infty^2 R_{\max} \\
&\quad + 256\tau_{\mathrm{mix}}\phi_\infty^2 R_{\max}\sqrt{\sum_{t=0}^{\infty}\eta_t^2}\sqrt{2\log\frac{8}{\delta}} + 384\eta_0\tau_{\mathrm{mix}}\phi_\infty^2 R_{\max} \\
&\quad + 192\tau_{\mathrm{mix}}\phi_\infty^4 R_{\max}\sum_{t=0}^{\infty}\eta_t^2 .
\end{aligned}$$

On the other hand, using (20), we also have

$$\|\boldsymbol{e}_T\|^2 = \|\boldsymbol{e}_0\|^2 + \sum_{t=1}^{T}\eta_{t-1}^2\|\boldsymbol{g}(\boldsymbol{\theta}_{t-1}, Z_{t-1})\|^2 + 2\sum_{t=1}^{T}\eta_{t-1}\langle\bar{\boldsymbol{g}}(\boldsymbol{\theta}_{t-1}), \boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^*\rangle + B_T .$$

Rearranging terms yields

$$\begin{aligned}
2\sum_{t=1}^{T}\eta_{t-1}\langle\bar{\boldsymbol{g}}(\boldsymbol{\theta}_{t-1}), \boldsymbol{\theta}^* - \boldsymbol{\theta}_{t-1}\rangle &= \|\boldsymbol{e}_0\|^2 - \|\boldsymbol{e}_T\|^2 + \sum_{t=1}^{T}\eta_{t-1}^2\|\boldsymbol{g}(\boldsymbol{\theta}_{t-1}, Z_{t-1})\|^2 + B_T \\
&\le R_{\max}^2 + 768\tau_{\mathrm{mix}}\phi_\infty^2 R_{\max}^2\sqrt{\sum_{t=0}^{\infty}\eta_t^2}\sqrt{2\log\frac{8}{\delta}} \\
&\quad + 1152\eta_0\tau_{\mathrm{mix}}\phi_\infty^2 R_{\max}^2 \\
&\quad + 1344\tau_{\mathrm{mix}}\phi_\infty^4 R_{\max}^2\sum_{t=0}^{\infty}\eta_t^2 + 9\phi_\infty^4 R_{\max}^2\sum_{t=0}^{\infty}\eta_t^2 .
\end{aligned}$$

Using the convexity of f and (4), we also have

$$\begin{aligned}
f(\bar{\boldsymbol{\theta}}_T) - f(\boldsymbol{\theta}^*) &\le \left(\frac{1}{\sum_{t=1}^T \eta_{t-1}}\right) \sum_{t=1}^T \eta_{t-1} \langle \bar{\boldsymbol{g}}(\boldsymbol{\theta}_{t-1}), \boldsymbol{\theta}^* - \boldsymbol{\theta}_{t-1} \rangle \\
&\le \frac{R_{\max}^2}{\sum_{t=1}^T \eta_{t-1}} \left(0.5 + 384 \tau_{\text{mix}} \phi_\infty^2 \sqrt{\sum_{t=0}^\infty \eta_t^2} \sqrt{2 \log \frac{8}{\delta}} + 576 \eta_0 \tau_{\text{mix}} \phi_\infty^2 \right) \\
&\quad + \frac{R_{\max}^2}{\sum_{t=1}^T \eta_{t-1}} \left(672 \tau_{\text{mix}} \phi_\infty^4 \sum_{t=0}^\infty \eta_t^2 + 4.5 \phi_\infty^4 \sum_{t=0}^\infty \eta_t^2 \right) \\
&\le \frac{C_{\text{robust}}}{\sum_{t=1}^T \eta_{t-1}},
\end{aligned}$$

where

$$\begin{aligned}
C_{\text{robust}} \coloneqq R_{\max}^2 \Bigg(& 0.5 + 384 \tau_{\text{mix}} \phi_\infty^2 \sqrt{\sum_{t=0}^\infty \eta_t^2} \sqrt{2 \log \left(\frac{8}{\delta}\right)} \\
& + 576 \eta_0 \tau_{\text{mix}} \phi_\infty^2 + 672 \tau_{\text{mix}} \phi_\infty^4 \sum_{t=0}^\infty \eta_t^2 + 4.5 \phi_\infty^4 \sum_{t=0}^\infty \eta_t^2 \Bigg).
\end{aligned}$$

Combining the two upper bounds for $f(\bar{\boldsymbol{\theta}}_T) - f(\boldsymbol{\theta}^*)$, with probability at least $1-\delta$, we have

$$f(\bar{\boldsymbol{\theta}}_T) - f(\boldsymbol{\theta}^*) \le \min \left\{ \frac{C_{\text{fast}}^2}{\omega (\sum_{t=1}^T \eta_{t-1})^2}, \frac{C_{\text{robust}}}{\sum_{t=1}^T \eta_{t-1}} \right\}.$$

■

Appendix D. Explicit Derivation of Stepsize Schedules

In this section, we consider specific stepsize schedules (η_t) and derive corresponding corollaries of Theorem 6 and Theorem 7.

Corollary 18 (High-probability bounded iterates) *In the setting of Theorem 6, consider the following stepsize schedule:*

$$\eta_t = \frac{1}{c \tau_{\text{mix}} \phi_\infty^2 \sqrt{t+1} \log(t+3)}, \quad \forall t \ge 0,$$

for some numerical constant $c > 0$. Fix any $\delta \in (0,1)$ and let $R_{\text{base}} \coloneqq \max\{\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|, \|\boldsymbol{\theta}^*\|, r_\infty/\phi_\infty\}$, $A_1(\delta) = 1536\sqrt{\frac{3}{\log 2}}\sqrt{2\log\frac{2}{\delta}} + 2304$, and $A_2 = \frac{8118}{\log 2}$. Then, provided that $c > c_{\min}(\delta) \coloneqq \frac{A_1(\delta)+\sqrt{A_1^2(\delta)+4A_2}}{2}$, with probability at least $1-\delta$, we have

$$\sup_{t \ge 0} \|\boldsymbol{\theta}_t\|_2 \le \rho R_{\text{base}}, \quad \text{where } \rho = \frac{2c}{\sqrt{c^2 - A_1(\delta) c - A_2}}.$$

Proof Fix $\delta \in (0,1)$. We substitute $a_t = \frac{1}{\sqrt{t+1}\log(t+3)}, \forall t \geq 0$ into Theorem 6. Lemma 24(i) gives $\sum_{t=0}^{\infty} a_t^2 \leq \frac{3}{\log 2}$. Thus, a sufficient requirement for c and ρ translates to

$$4 + 1536\frac{\rho^2}{c}\sqrt{\frac{3}{\log 2}}\sqrt{2\log\frac{2}{\delta}} + 2304\frac{\rho^2}{c} + 2706\frac{3}{\log 2}\frac{\rho^2}{c^2} \leq \rho^2\,.$$

We now derive an explicit relationship between c and ρ for the above inequality to hold. Denote $A_1(\delta) = 1536\sqrt{\frac{3}{\log 2}}\sqrt{2\log\frac{2}{\delta}} + 2304$ and $A_2 = 2706\frac{3}{\log 2}$. Then, the above inequality is equivalent to

$$4 + \frac{\rho^2}{c}A_1(\delta) + \frac{\rho^2}{c^2}A_2 \leq \rho^2\,. \tag{29}$$

For any fixed $c > 0$, rearranging (29) yields

$$c^2 - A_1(\delta)c - A_2 \geq \frac{4c^2}{\rho^2},$$

which requires $c^2 - A_1(\delta)c - A_2 > 0$ since the right-hand side is positive. In particular, for any $c > c_{\text{min}}(\delta) \coloneqq \frac{A_1(\delta)+\sqrt{A_1^2(\delta)+4A_2}}{2}$, choosing

$$\rho \geq \rho_{\text{min}}(c) \coloneqq \frac{2c}{\sqrt{c^2 - A_1(\delta)c - A_2}}$$

guarantees the condition. This completes the proof. ■

Corollary 19 (High-probability rate for PR averaging) *Under Assumption 1, with* $\phi_\infty, r_\infty, \tau_{\text{mix}}$ *and* ω *as in Lemmas 2–4, consider the following stepsize schedule:*

$$\eta_t = \frac{1}{c\tau_{\text{mix}}\phi_\infty^2\sqrt{t+1}\log(t+3)}, \qquad \forall t \geq 0,$$

for some numerical constant $c > 0$*. Fix any* $\delta \in (0,1)$ *and let* $R_{\text{base}} \coloneqq \max\{\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|, \|\boldsymbol{\theta}^*\|, r_\infty/\phi_\infty\}$, $R_{\text{max}} = \rho R_{\text{base}}$, $A_1(\delta) = 1536\sqrt{\frac{3}{\log 2}}\sqrt{2\log\frac{8}{\delta}} + 2304$*, and* $A_2 = \frac{8118}{\log 2}$*. Then, provided that* $c > c_{\text{min}}(\delta) \coloneqq \frac{A_1(\delta)+\sqrt{A_1^2(\delta)+4A_2}}{2}$*, with probability at least* $1-\delta$*, the following upper bound holds for* $f(\bar{\boldsymbol{\theta}}_T) - f(\boldsymbol{\theta}^*)$*:*

$$R_{\text{max}}^2 \min\left\{\underbrace{\frac{c^2\tau_{\text{mix}}^2\phi_\infty^4\log^2(T+2)}{\omega\,(\sqrt{T+1}-1)^2}\left(1 + \frac{264}{c} + \frac{176\sqrt{3}}{c\sqrt{\log 2}}\sqrt{2\log\frac{8}{\delta}} + \frac{288}{c^2\tau_{\text{mix}}\log 2}\right)^2}_{\textit{fast rate}},\right.$$

$$\left.\underbrace{\frac{c\tau_{\text{mix}}\phi_\infty^2\log(T+2)}{4(\sqrt{T+1}-1)}\left(1 + \frac{1152}{c} + \frac{768\sqrt{3}}{c\sqrt{\log 2}}\sqrt{2\log\frac{8}{\delta}} + \frac{4059}{c^2\tau_{\text{mix}}\log 2}\right)}_{\textit{robust rate}}\right\},$$

where $\rho = \frac{2c}{\sqrt{c^2 - A_1(\delta)c - A_2}}$*. That is,* $f(\bar{\boldsymbol{\theta}}_T) - f(\boldsymbol{\theta}^*) = \min\{\mathcal{O}(\log^2(T)/(\omega T)), \mathcal{O}(\log(T)/\sqrt{T})\}$ *up to constants depending on* c*,* τ_{mix}*,* ϕ_∞*, and* δ*.*

Proof The condition for (c, ρ) can be derived as in Corollary 18. We now focus on simplifying the constants C_{robust} and C_{fast}. Lemma 24(i) gives $\sum_{t=0}^{\infty} \eta_t^2 \leq \frac{3}{c^2 \tau_{\text{mix}}^2 \phi_\infty^4 \log 2}$. Thus, we have

$$\begin{aligned} C_{\text{robust}} &\leq R_{\max}^2 \left[\frac{1}{2} + 384 \tau_{\text{mix}} \phi_\infty^2 \sqrt{\frac{3}{c^2 \tau_{\text{mix}}^2 \phi_\infty^4 \log 2}} \sqrt{2 \log \frac{8}{\delta}} + 576 \eta_0 \tau_{\text{mix}} \phi_\infty^2 \right. \\ &\qquad \left. + (672 \tau_{\text{mix}} + 4.5)\, \phi_\infty^4 \frac{3}{c^2 \tau_{\text{mix}}^2 \phi_\infty^4 \log 2} \right] \\ &\leq R_{\max}^2 \left[\frac{1}{2} + \frac{384\sqrt{3}}{c\sqrt{\log 2}} \sqrt{2 \log \frac{8}{\delta}} + \frac{576}{c} + \frac{2029.5}{c^2 \tau_{\text{mix}} \log 2} \right], \end{aligned}$$

where we used $\eta_0 \leq 1/(c \tau_{\text{mix}} \phi_\infty^2)$ and $\tau_{\text{mix}} \geq 1$, so that $\frac{2016}{c^2 \tau_{\text{mix}} \log 2} + \frac{13.5}{c^2 \tau_{\text{mix}}^2 \log 2} \leq \frac{2029.5}{c^2 \tau_{\text{mix}} \log 2}$. Similarly,

$$\begin{aligned} C_{\text{fast}} &\leq R_{\max} \left[2 + 528 \eta_0 \tau_{\text{mix}} \phi_\infty^2 + 352 \tau_{\text{mix}} \phi_\infty^2 \sqrt{\frac{3}{c^2 \tau_{\text{mix}}^2 \phi_\infty^4 \log 2}} \sqrt{2 \log \frac{8}{\delta}} \right. \\ &\qquad \left. + 192 \tau_{\text{mix}} \phi_\infty^4 \frac{3}{c^2 \tau_{\text{mix}}^2 \phi_\infty^4 \log 2} \right] \\ &\leq R_{\max} \left[2 + \frac{528}{c} + \frac{352\sqrt{3}}{c\sqrt{\log 2}} \sqrt{2 \log \frac{8}{\delta}} + \frac{576}{c^2 \tau_{\text{mix}} \log 2} \right]. \end{aligned}$$

Recalling $R_{\max} = \rho R_{\text{base}}$, we obtain

$$\begin{aligned} C_{\text{fast}} &\leq 2 \rho R_{\text{base}} \left(1 + \frac{264}{c} + \frac{176\sqrt{3}}{c\sqrt{\log 2}} \sqrt{2 \log \frac{8}{\delta}} + \frac{288}{c^2 \tau_{\text{mix}} \log 2} \right), \\ C_{\text{robust}} &\leq \frac{\rho^2 R_{\text{base}}^2}{2} \left(1 + \frac{1152}{c} + \frac{768\sqrt{3}}{c\sqrt{\log 2}} \sqrt{2 \log \frac{8}{\delta}} + \frac{4059}{c^2 \tau_{\text{mix}} \log 2} \right). \end{aligned}$$

Next, let $S_T := \sum_{t=1}^{T} \eta_{t-1} = \sum_{s=0}^{T-1} \eta_s$. By Lemma 24 with $p = 1$ and $\kappa = c \tau_{\text{mix}} \phi_\infty^2$, $S_T \geq \frac{2(\sqrt{T+1}-1)}{c \tau_{\text{mix}} \phi_\infty^2 \log(T+2)}$. Therefore,

$$\frac{1}{S_T} \leq \frac{c \tau_{\text{mix}} \phi_\infty^2 \log(T+2)}{2(\sqrt{T+1}-1)}, \qquad \frac{1}{S_T^2} \leq \frac{c^2 \tau_{\text{mix}}^2 \phi_\infty^4 \log^2(T+2)}{4(\sqrt{T+1}-1)^2}.$$

Substituting the above bounds into Theorem 7 yields the claimed result. ■

Appendix E. Removing the dependence on τ_{mix} in the stepsize

The choice in (15) depends on the unknown mixing parameter τ_{mix}, which may make the algorithm impractical. This dependence of the stepsize on τ_{mix} can be avoided by running TD(0) with the modified stepsize schedule

$$\eta_t = \frac{1}{c\, \phi_\infty^2 \sqrt{t+1} \log^2(t+3)}, \qquad \forall t \geq 0, \tag{30}$$

for some numerical constant $c > 0$. For $T \geq 1$, let

$$S_T := \sum_{t=1}^{T} \eta_{t-1}, \qquad \bar{\boldsymbol{\theta}}_T := \frac{1}{S_T} \sum_{t=1}^{T} \eta_{t-1} \boldsymbol{\theta}_{t-1} . \tag{31}$$

Lemma 20 (Deterministic bound on the finite prefix) *Assume the feature and reward bounds in Lemma 2. For any nonnegative stepsizes* $(\eta_t)_{t \geq 0}$ *and any integer* $m \geq 0$*, the TD(0) iterates satisfy*

$$\max_{0 \leq s \leq m} \|\boldsymbol{\theta}_s\| \leq \left(\|\boldsymbol{\theta}_0\| + \frac{r_\infty}{(1+\gamma)\phi_\infty} \right) \prod_{t=0}^{m-1} \left(1 + (1+\gamma)\phi_\infty^2 \eta_t\right) - \frac{r_\infty}{(1+\gamma)\phi_\infty} .$$

In particular, for the stepsize (30)*, define*

$$B_m := \left(\|\boldsymbol{\theta}_0\| + \frac{r_\infty}{(1+\gamma)\phi_\infty} \right) \prod_{t=0}^{m-1} \left(1 + \frac{1+\gamma}{c\sqrt{t+1}\log^2(t+3)} \right) - \frac{r_\infty}{(1+\gamma)\phi_\infty} .$$

Then $\max_{0 \leq s \leq m} \|\boldsymbol{\theta}_s\| \leq B_m$*. Furthermore,*

$$B_m \leq \left(\|\boldsymbol{\theta}_0\| + \frac{r_\infty}{(1+\gamma)\phi_\infty} \right) \exp\left(\frac{2(1+\gamma)\sqrt{m}}{c\log^2 3} \right) - \frac{r_\infty}{(1+\gamma)\phi_\infty} .$$

Proof The bounds in Lemma 2 imply

$$\begin{aligned} \|\boldsymbol{g}(\boldsymbol{\theta}_t, Z_t)\| &= |r_t + \gamma \langle \boldsymbol{\phi}(s_{t+1}), \boldsymbol{\theta}_t \rangle - \langle \boldsymbol{\phi}(s_t), \boldsymbol{\theta}_t \rangle| \, \|\boldsymbol{\phi}(s_t)\| \\ &\leq r_\infty \phi_\infty + (1+\gamma)\phi_\infty^2 \|\boldsymbol{\theta}_t\| . \end{aligned}$$

Therefore,

$$\|\boldsymbol{\theta}_{t+1}\| \leq \left(1 + (1+\gamma)\phi_\infty^2 \eta_t\right) \|\boldsymbol{\theta}_t\| + r_\infty \phi_\infty \eta_t .$$

Let $b := r_\infty / ((1+\gamma)\phi_\infty)$ and $u_t := \|\boldsymbol{\theta}_t\| + b$. Then $u_{t+1} \leq (1 + (1+\gamma)\phi_\infty^2 \eta_t) u_t$. Iterating this inequality yields, for every $s \leq m$,

$$u_s \leq u_0 \prod_{t=0}^{s-1} (1 + (1+\gamma)\phi_\infty^2 \eta_t) \leq u_0 \prod_{t=0}^{m-1} (1 + (1+\gamma)\phi_\infty^2 \eta_t) .$$

Taking the supremum over $0 \leq s \leq m$ and subtracting b proves the first claim. Substituting (30) gives the displayed formula for B_m. Finally, using $1 + x \leq e^x$ and

$$\sum_{t=0}^{m-1} \frac{1}{\sqrt{t+1}\log^2(t+3)} \leq \frac{2\sqrt{m}}{\log^2 3}$$

gives the closed-form bound. ■

Corollary 21 (Mixing-free stepsize) *Under Assumption 1, with* $\phi_\infty, r_\infty, \tau_{\text{mix}}$ *and* ω *as in Lemmas 2–4, run TD(0) with the stepsize schedule:*

$$\eta_t = \frac{1}{c\,\phi_\infty^2 \sqrt{t+1}\log^2(t+3)}, \qquad t \geq 0,$$

for some numerical constant $c > 0$*, and form the PR average* $\bar{\boldsymbol{\theta}}_T$ *as in* (31)*. Let*

$$m^* \coloneqq \min\{m \geq 0 : \log(m+3) \geq \tau_{\text{mix}}\}\,.$$

For $m \geq 0$*, define the deterministic prefix bound*

$$B_m \coloneqq \left(\|\boldsymbol{\theta}_0\| + \frac{r_\infty}{(1+\gamma)\phi_\infty}\right) \prod_{t=0}^{m-1} \left(1 + \frac{1+\gamma}{c\sqrt{t+1}\log^2(t+3)}\right) - \frac{r_\infty}{(1+\gamma)\phi_\infty}\,.$$

Fix any $\delta \in (0,1)$ *and set*

$$R_{\text{base}} \coloneqq \max\left\{B_{m^*} + \|\boldsymbol{\theta}^*\|\,,\ \|\boldsymbol{\theta}^*\|\,,\ \frac{r_\infty}{\phi_\infty}\right\}\,.$$

Define

$$A_1(\delta) \coloneqq 1536\sqrt{\frac{3}{\log 2}}\sqrt{2\log\frac{8}{\delta}} + 2304, \qquad A_2 \coloneqq \frac{8118}{\log 2},$$

$$c_{\min}(\delta) \coloneqq \frac{A_1(\delta) + \sqrt{A_1^2(\delta) + 4A_2}}{2}, \qquad \rho \coloneqq \frac{2c}{\sqrt{c^2 - A_1(\delta)c - A_2}}, \qquad R_{\max} \coloneqq \rho R_{\text{base}}\,.$$

Then, provided that $c > c_{\min}(\delta)$*, with probability at least* $1-\delta$*, we have*

$$\sup_{t\geq 0}\ \|\boldsymbol{\theta}_t\| \leq R_{\max},$$

and for every $T \geq 1$*,* $f(\bar{\boldsymbol{\theta}}_T) - f(\boldsymbol{\theta}^*)$ *is upper-bounded by*

$$R_{\max}^2 \min\left\{\underbrace{\frac{c^2\phi_\infty^4\log^4(T+2)}{\omega(\sqrt{T+1}-1)^2}\left(1 + \frac{264\tau_{\text{mix}}}{c\log^2 3} + \frac{176\sqrt{3}\tau_{\text{mix}}}{c\sqrt{\log 2}}\sqrt{2\log\frac{8}{\delta}} + \frac{288\tau_{\text{mix}}}{c^2\log 2}\right)^2}_{\textit{fast rate}},\right.$$
$$\left.\underbrace{\frac{c\phi_\infty^2\log^2(T+2)}{4(\sqrt{T+1}-1)}\left(1 + \frac{1152\tau_{\text{mix}}}{c\log^2 3} + \frac{768\sqrt{3}\tau_{\text{mix}}}{c\sqrt{\log 2}}\sqrt{2\log\frac{8}{\delta}} + \frac{4032\tau_{\text{mix}} + 27}{c^2\log 2}\right)}_{\textit{robust rate}}\right\}.$$

Equivalently, since $\tau_{\text{mix}} \geq 1$*, the horizon dependence is*

$$f(\bar{\boldsymbol{\theta}}_T) - f(\boldsymbol{\theta}^*) = \min\left\{\widetilde{\mathcal{O}}\left(\frac{R_{\max}^2\tau_{\text{mix}}^2\phi_\infty^4}{\omega T}\right), \widetilde{\mathcal{O}}\left(\frac{R_{\max}^2\tau_{\text{mix}}\phi_\infty^2}{\sqrt{T}}\right)\right\},$$

where the $\widetilde{\mathcal{O}}(\cdot)$ *notation hides logarithmic factors in* T *and* $1/\delta$*, as well as numerical constants depending on* c*. Moreover, the prefactor depends doubly exponentially on* τ_{mix} *through its dependence on* R_{base}*: using* $m^* \leq e^{\tau_{\text{mix}}}$ *and Lemma 20,*

$$B_{m^*} \leq \left(\|\boldsymbol{\theta}_0\| + \frac{r_\infty}{(1+\gamma)\phi_\infty}\right)\exp\left(\frac{2(1+\gamma)e^{\tau_{\text{mix}}/2}}{c\log^2 3}\right) - \frac{r_\infty}{(1+\gamma)\phi_\infty}\,.$$

Consequently, for fixed $c, \gamma, \phi_\infty, r_\infty, \boldsymbol{\theta}_0, \boldsymbol{\theta}^*$*, and* δ*, there is a constant* C_0 *independent of* τ_{mix} *such that*

$$R_{\max}^2 \le C_0 \exp\left(\frac{4(1+\gamma)e^{\tau_{\text{mix}}/2}}{c\log^2 3}\right) = \exp\left(\mathcal{O}\left(e^{\tau_{\text{mix}}/2}\right)\right).$$

Thus the total fast and robust prefactors are both of doubly exponential order $\exp(\mathcal{O}(e^{\tau_{\text{mix}}/2}))$ *up to polynomial factors in* τ_{mix}.

Proof Let

$$L_\delta := \sqrt{2\log\frac{8}{\delta}}, \qquad H := \sum_{t=0}^{\infty} \eta_t^2 .$$

We first prove the bounded-iterates event. Consider the shifted process

$$\boldsymbol{\theta}'_k := \boldsymbol{\theta}_{m^*+k}, \qquad Z'_k := Z_{m^*+k}, \qquad \eta'_k := \eta_{m^*+k} .$$

The shifted chain has the same transition kernel, hence the same mixing constant τ_{mix}. Then

$$\eta'_k = \frac{1}{c\tau_{\text{mix}}\phi_\infty^2} a'_k, \qquad a'_k := \frac{\tau_{\text{mix}}}{\sqrt{m^*+k+1}\log^2(m^*+k+3)} .$$

By the definition of m^*, for every $k \ge 0$,

$$0 < a'_k \le \frac{1}{\sqrt{m^*+k+1}\log(m^*+k+3)}, \qquad a'_0 \le 1,$$

and $(a'_k)_{k\ge 0}$ is non-increasing. Moreover, Lemma 24(i) gives

$$\begin{aligned}\sum_{k=0}^{\infty}(a'_k)^2 &= \tau_{\text{mix}}^2 \sum_{k=0}^{\infty} \frac{1}{(m^*+k+1)\log^4(m^*+k+3)} \\ &\le \sum_{k=0}^{\infty} \frac{1}{(m^*+k+1)\log^2(m^*+k+3)} \le \frac{3}{\log 2} .\end{aligned}$$

Conditioning on $\mathcal{F}_{m^*} = \sigma(Z_0, \ldots, Z_{m^*})$ and applying Theorem 6 to the shifted recursion with confidence parameter $\delta/4$, we obtain an event $\mathcal{E}_R^+$ such that

$$\mathbb{P}\{\mathcal{E}_R^+ \mid \mathcal{F}_{m^*}\} \ge 1 - \frac{\delta}{4} \quad a.s.,$$

and on $\mathcal{E}_R^+$,

$$\sup_{k\ge 0} \|\boldsymbol{\theta}_{m^*+k}\| \le \rho \max\left\{\|\boldsymbol{\theta}_{m^*} - \boldsymbol{\theta}^*\|, \|\boldsymbol{\theta}^*\|, \frac{r_\infty}{\phi_\infty}\right\} \le \rho R_{\text{base}} = R_{\max} .$$

The constants $A_1(\delta)$, A_2, and $c_{\min}(\delta)$ are exactly those obtained from the last square-summability bound and the confidence parameter $\delta/4$. Taking expectations in the preceding conditional probability bound gives $\mathbb{P}\{\mathcal{E}_R^+\} \ge 1 - \delta/4$. On the finite prefix, Lemma 20 gives $\sup_{0\le s\le m^*} \|\boldsymbol{\theta}_s\| \le B_{m^*} \le R_{\text{base}} \le R_{\max}$. Therefore the event

$$\mathcal{E}_R := \left\{\sup_{t\ge 0} \|\boldsymbol{\theta}_t\| \le R_{\max}\right\}$$

satisfies

$$\mathbb{P}\{\mathcal{E}_R\} \geq 1 - \frac{\delta}{4}\,.$$

We next follow the proof strategy of Theorem 7 in Section C, spelling out the high-probability events needed for the present mixing-free stepsize. For every $T \geq 1$, the energy decomposition (20), which will be used for the robust-rate bound, gives

$$\|\boldsymbol{e}_T\|^2 = \|\boldsymbol{e}_0\|^2 + \sum_{t=1}^{T} \eta_{t-1}^2 \|\boldsymbol{g}(\boldsymbol{\theta}_{t-1}, Z_{t-1})\|^2 + 2\sum_{t=1}^{T} \eta_{t-1} \langle \bar{\boldsymbol{g}}(\boldsymbol{\theta}_{t-1}), \boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^* \rangle + B_T, \tag{32}$$

where $\boldsymbol{e}_t = \boldsymbol{\theta}_t - \boldsymbol{\theta}^*$ and

$$B_T = 2\sum_{t=1}^{T} \eta_{t-1} \left\langle \boldsymbol{g}(\boldsymbol{\theta}_{t-1}, Z_{t-1}) - \bar{\boldsymbol{g}}(\boldsymbol{\theta}_{t-1}),\, \boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^* \right\rangle .$$

Similarly, the PR decomposition (26), which will be used for the fast-rate bound, gives

$$\boldsymbol{A}\bar{\boldsymbol{e}}_T = I_{1,T} + I_{2,T} + I_{3,T}, \tag{33}$$

where

$$\bar{\boldsymbol{e}}_T := \frac{1}{S_T} \sum_{t=1}^{T} \eta_{t-1} \boldsymbol{e}_{t-1},$$

and

$$I_{1,T} = \frac{1}{S_T} \sum_{t=1}^{T} (\boldsymbol{e}_{t-1} - \boldsymbol{e}_t), \qquad I_{2,T} = \frac{1}{S_T} \sum_{t=1}^{T} \eta_{t-1} \boldsymbol{\xi}_{t-1}, \qquad I_{3,T} = \frac{1}{S_T} \sum_{t=1}^{T} \eta_{t-1} \boldsymbol{\delta}_{t-1} \boldsymbol{e}_{t-1}\,.$$

We now introduce three additional high-probability events, $\mathcal{E}_2$, $\mathcal{E}_3$, and $\mathcal{E}_4$, to control the terms appearing in the two decompositions above. First, applying Lemma 16 with confidence parameter $\delta/4$ and using $r_\infty \phi_\infty + 2\phi_\infty^2 \|\boldsymbol{\theta}^*\| \leq 3\phi_\infty^2 R_{\max}$, we obtain an event $\mathcal{E}_2$ such that

$$\mathbb{P}\{\mathcal{E}_2\} \geq 1 - \frac{\delta}{4},$$

and, on $\mathcal{E}_2$, for all $T \geq 1$,

$$\|I_{2,T}\| \leq \frac{48\tau_{\text{mix}} \phi_\infty^2 R_{\max}}{S_T} \left(2\sqrt{H} L_\delta + 3\eta_0\right) .$$

Next, to control $I_{3,T}$ in (33) and the Markov-bias term B_T in (32), we use the stopped process at radius $R_{\max}$:

$$T_{\text{exit}} := \inf\{t \geq 0 : \|\boldsymbol{\theta}_t\| > R_{\max}\}, \quad \tilde{\boldsymbol{\theta}}_t := \boldsymbol{\theta}_{t \wedge T_{\text{exit}}}, \quad \tilde{\boldsymbol{e}}_t := \tilde{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*, \quad \tilde{\eta}_t := \eta_t \mathbf{1}\{t < T_{\text{exit}}\}\,.$$

Define the corresponding localized quantities:

$$\tilde{I}_{3,T} := \frac{1}{S_T} \sum_{t=1}^{T} \tilde{\eta}_{t-1} \boldsymbol{\delta}_{t-1} \tilde{\boldsymbol{e}}_{t-1}, \qquad \tilde{B}_T := 2\sum_{t=1}^{T} \tilde{\eta}_{t-1} \left\langle \boldsymbol{g}(\tilde{\boldsymbol{\theta}}_{t-1}, Z_{t-1}) - \bar{\boldsymbol{g}}(\tilde{\boldsymbol{\theta}}_{t-1}),\, \tilde{\boldsymbol{\theta}}_{t-1} - \boldsymbol{\theta}^* \right\rangle .$$

Applying Lemma 17 with confidence parameter $\delta/4$ gives an event $\mathcal{E}_3$ such that

$$\mathbb{P}\{\mathcal{E}_3\} \geq 1 - \frac{\delta}{4},$$

and, on $\mathcal{E}_3$, for all $T \geq 1$,

$$\left\|\tilde{I}_{3,T}\right\| \leq \frac{1}{S_T}\left(256\tau_{\mathrm{mix}}\phi_\infty^2 R_{\max}\sqrt{H}L_\delta + 384\eta_0\tau_{\mathrm{mix}}\phi_\infty^2 R_{\max} + 192\tau_{\mathrm{mix}}\phi_\infty^4 R_{\max}H\right) .$$

Finally, Lemma 13 with confidence parameter $\delta/4$, together with the deterministic quadratic-term bound used in the proof of Lemma 14, gives an event $\mathcal{E}_4$ such that

$$\mathbb{P}\{\mathcal{E}_4\} \geq 1 - \frac{\delta}{4},$$

and, on $\mathcal{E}_4$, for all $T \geq 1$,

$$\begin{aligned}\sum_{t=1}^{T}\tilde{\eta}_{t-1}^2\left\|\boldsymbol{g}(\tilde{\boldsymbol{\theta}}_{t-1}, Z_{t-1})\right\|^2 + \tilde{B}_T \leq {} & 768\tau_{\mathrm{mix}}\phi_\infty^2 R_{\max}^2\sqrt{H}L_\delta + 1152\eta_0\tau_{\mathrm{mix}}\phi_\infty^2 R_{\max}^2 \\ & + 1344\tau_{\mathrm{mix}}\phi_\infty^4 R_{\max}^2 H + 9\phi_\infty^4 R_{\max}^2 H .\end{aligned} \tag{34}$$

Define

$$\mathcal{E} := \mathcal{E}_R \cap \mathcal{E}_2 \cap \mathcal{E}_3 \cap \mathcal{E}_4 .$$

Since each of $\mathcal{E}_R$, $\mathcal{E}_2$, $\mathcal{E}_3$, and $\mathcal{E}_4$ has probability at least $1 - \delta/4$, the union bound gives $\mathbb{P}\{\mathcal{E}\} \geq 1 - \delta$. On $\mathcal{E}_R$, we have $T_{\mathrm{exit}} = \infty$, so the stopped and original processes coincide. Hence, on $\mathcal{E}$, $\tilde{I}_{3,T} = I_{3,T}$ and $\tilde{B}_T = B_T$ for every $T \geq 1$.

We now derive the fast-rate control on $\mathcal{E}$. By Lemma 15,

$$\|I_{1,T}\| \leq \frac{2R_{\max}}{S_T} .$$

Combining (33) with the preceding bounds on $I_{1,T}$, $I_{2,T}$, and $\tilde{I}_{3,T} = I_{3,T}$, for all $T \geq 1$,

$$\|\boldsymbol{A}\bar{\boldsymbol{e}}_T\| \leq \frac{R_{\max}}{S_T}\left[2 + 528\eta_0\tau_{\mathrm{mix}}\phi_\infty^2 + 352\tau_{\mathrm{mix}}\phi_\infty^2\sqrt{H}L_\delta + 192\tau_{\mathrm{mix}}\phi_\infty^4 H\right] . \tag{35}$$

For the present stepsize,

$$\eta_0 = \frac{1}{c\phi_\infty^2\log^2 3}, \qquad H = \frac{1}{c^2\phi_\infty^4}\sum_{t=0}^{\infty}\frac{1}{(t+1)\log^4(t+3)} \leq \frac{3}{c^2\phi_\infty^4\log 2}, \tag{36}$$

where the last inequality follows from $\log(t+3) \geq \log 3 > 1$ and Lemma 24(i). Hence

$$\begin{aligned}&2 + 528\eta_0\tau_{\mathrm{mix}}\phi_\infty^2 + 352\tau_{\mathrm{mix}}\phi_\infty^2\sqrt{H}L_\delta + 192\tau_{\mathrm{mix}}\phi_\infty^4 H \\ &\quad \leq 2\left(1 + \frac{264\tau_{\mathrm{mix}}}{c\log^2 3} + \frac{176\sqrt{3}\tau_{\mathrm{mix}}}{c\sqrt{\log 2}}L_\delta + \frac{288\tau_{\mathrm{mix}}}{c^2\log 2}\right) .\end{aligned}$$

Using (27), (35), and the preceding simplification,

$$f(\bar{\boldsymbol{\theta}}_T) - f(\boldsymbol{\theta}^*) \le \frac{4R_{\max}^2}{\omega S_T^2}\left(1 + \frac{264\tau_{\text{mix}}}{c\log^2 3} + \frac{176\sqrt{3}\tau_{\text{mix}}}{c\sqrt{\log 2}}L_\delta + \frac{288\tau_{\text{mix}}}{c^2\log 2}\right)^2 . \tag{37}$$

We next derive the robust-rate control on $\mathcal{E}$. Using the robust rate decomposition (32), the identity $\tilde{B}_T = B_T$ on $\mathcal{E}$, and (34),

$$2\sum_{t=1}^{T}\eta_{t-1}\langle\bar{\boldsymbol{g}}(\boldsymbol{\theta}_{t-1}), \boldsymbol{\theta}^* - \boldsymbol{\theta}_{t-1}\rangle \le R_{\max}^2 + 768\tau_{\text{mix}}\phi_\infty^2 R_{\max}^2\sqrt{H}L_\delta + 1152\eta_0\tau_{\text{mix}}\phi_\infty^2 R_{\max}^2 + 1344\tau_{\text{mix}}\phi_\infty^4 R_{\max}^2 H + 9\phi_\infty^4 R_{\max}^2 H\,.$$

By convexity of f and (4),

$$f(\bar{\boldsymbol{\theta}}_T) - f(\boldsymbol{\theta}^*) \le \frac{R_{\max}^2}{S_T}\left[\frac{1}{2} + 384\tau_{\text{mix}}\phi_\infty^2\sqrt{H}L_\delta + 576\eta_0\tau_{\text{mix}}\phi_\infty^2 + 672\tau_{\text{mix}}\phi_\infty^4 H + \frac{9}{2}\phi_\infty^4 H\right] .$$

Using (36),

$$\frac{1}{2} + 384\tau_{\text{mix}}\phi_\infty^2\sqrt{H}L_\delta + 576\eta_0\tau_{\text{mix}}\phi_\infty^2 + 672\tau_{\text{mix}}\phi_\infty^4 H + \frac{9}{2}\phi_\infty^4 H \le \frac{1}{2}\left(1 + \frac{1152\tau_{\text{mix}}}{c\log^2 3} + \frac{768\sqrt{3}\tau_{\text{mix}}}{c\sqrt{\log 2}}L_\delta + \frac{4032\tau_{\text{mix}} + 27}{c^2\log 2}\right) .$$

Therefore,

$$f(\bar{\boldsymbol{\theta}}_T) - f(\boldsymbol{\theta}^*) \le \frac{R_{\max}^2}{2S_T}\left(1 + \frac{1152\tau_{\text{mix}}}{c\log^2 3} + \frac{768\sqrt{3}\tau_{\text{mix}}}{c\sqrt{\log 2}}L_\delta + \frac{4032\tau_{\text{mix}} + 27}{c^2\log 2}\right) . \tag{38}$$

Finally, Lemma 24 with $p = 2$ and $\kappa = c\phi_\infty^2$ gives

$$S_T \ge \frac{2(\sqrt{T+1} - 1)}{c\phi_\infty^2\log^2(T+2)},$$

and hence

$$\frac{1}{S_T} \le \frac{c\phi_\infty^2\log^2(T+2)}{2(\sqrt{T+1} - 1)}, \qquad \frac{1}{S_T^2} \le \frac{c^2\phi_\infty^4\log^4(T+2)}{4(\sqrt{T+1} - 1)^2} . \tag{39}$$

Substituting (39) into (37) and (38) gives the displayed explicit rates. The big-$\widetilde{\mathcal{O}}$ statement follows directly from the displayed formula, and the doubly-exponential dependence on τ_{mix} follows from Lemma 20 and $m^* \le e^{\tau_{\text{mix}}}$. ■

Appendix F. Technical Lemmas

F.1. Pinelis' Inequality

Lemma 22 (Pinelis 1994, Thm. 3.5) *Let $(f_j)_{j\ge 0}$ be a martingale in a $(2, D)$-smooth Banach space $(X, \|\cdot\|)$ with $f_0 = 0$. Let $d_j := f_j - f_{j-1}$, $f^* := \sup_{j\ge 0} \|f_j\|$, and $\|d_j\|_\infty := \operatorname{ess\,sup} \|d_j\|$. If $\sum_{j\ge 1} \|d_j\|_\infty^2 \le b_*^2$, then for all $r \ge 0$,*

$$\mathbb{P}(f^* \ge r) \le 2\exp\left(-\frac{r^2}{2D^2 b_*^2}\right).$$

F.2. Miscellaneous lemmas

Lemma 23 (Dirichlet matrix) *The Dirichlet matrix*

$$\boldsymbol{L}_{\text{Dir}} := \boldsymbol{D} - \frac{1}{2}\big(\boldsymbol{D}\boldsymbol{P}^\mu + (\boldsymbol{P}^\mu)^\top \boldsymbol{D}\big)$$

is symmetric and positive semidefinite. More precisely, if

$$\boldsymbol{P}_{\text{sym}} := \frac{1}{2}\Big(\boldsymbol{P}^\mu + \boldsymbol{D}^{-1}(\boldsymbol{P}^\mu)^\top \boldsymbol{D}\Big),$$

then, for every $\boldsymbol{x} \in \mathbb{R}^n$,

$$\boldsymbol{x}^\top \boldsymbol{L}_{\text{Dir}} \boldsymbol{x} = \frac{1}{2}\sum_{i,j} \pi(i)\boldsymbol{P}_{\text{sym}}(i,j)(x_i - x_j)^2 = \frac{1}{2}\sum_{i,j} \pi(i)P^\mu(i,j)(x_i - x_j)^2 \ge 0\,.$$

Proof First, $\boldsymbol{L}_{\text{Dir}}$ is symmetric because $\boldsymbol{D}^\top = \boldsymbol{D}$:

$$\boldsymbol{L}_{\text{Dir}}^\top = \left(\boldsymbol{D} - \frac{1}{2}\big(\boldsymbol{D}\boldsymbol{P}^\mu + (\boldsymbol{P}^\mu)^\top \boldsymbol{D}\big)\right)^\top = \boldsymbol{D} - \frac{1}{2}\big((\boldsymbol{P}^\mu)^\top \boldsymbol{D} + \boldsymbol{D}\boldsymbol{P}^\mu\big) = \boldsymbol{L}_{\text{Dir}}\,.$$

Define the time-reversal kernel

$$\boldsymbol{P}^\mu_{\text{rev}} := \boldsymbol{D}^{-1}(\boldsymbol{P}^\mu)^\top \boldsymbol{D}, \qquad \boldsymbol{P}^\mu_{\text{rev}}(i,j) = \frac{\pi(j)P^\mu(j,i)}{\pi(i)}\,.$$

Since $\pi^\top \boldsymbol{P}^\mu = \pi^\top$, $\boldsymbol{P}^\mu_{\text{rev}}$ is row-stochastic:

$$\sum_j \boldsymbol{P}^\mu_{\text{rev}}(i,j) = \frac{1}{\pi(i)}\sum_j \pi(j)P^\mu(j,i) = 1\,.$$

Hence $\boldsymbol{P}_{\text{sym}} = \frac{1}{2}(\boldsymbol{P}^\mu + \boldsymbol{P}^\mu_{\text{rev}})$ is also row-stochastic and satisfies detailed balance:

$$\pi(i)\boldsymbol{P}_{\text{sym}}(i,j) = \frac{1}{2}\big(\pi(i)P^\mu(i,j) + \pi(j)P^\mu(j,i)\big) = \pi(j)\boldsymbol{P}_{\text{sym}}(j,i)\,.$$

Moreover,

$$\boldsymbol{D}\boldsymbol{P}_{\text{sym}} = \frac{1}{2}\big(\boldsymbol{D}\boldsymbol{P}^\mu + (\boldsymbol{P}^\mu)^\top \boldsymbol{D}\big),$$

so $\boldsymbol{L}_{\text{Dir}} = \boldsymbol{D} - \boldsymbol{D}\boldsymbol{P}_{\text{sym}}$. Therefore, for any $\boldsymbol{x} \in \mathbb{R}^n$,

$$\begin{aligned} \boldsymbol{x}^\top \boldsymbol{L}_{\text{Dir}} \boldsymbol{x} &= \sum_i \pi(i) x_i^2 - \sum_{i,j} \pi(i) \boldsymbol{P}_{\text{sym}}(i,j) x_i x_j \\ &= \frac{1}{2} \sum_{i,j} \pi(i) \boldsymbol{P}_{\text{sym}}(i,j)(x_i - x_j)^2 \geq 0, \end{aligned}$$

where the second equality uses that $\boldsymbol{P}_{\text{sym}}$ is row-stochastic and satisfies detailed balance. Finally, using the displayed identity for $\pi(i)\boldsymbol{P}_{\text{sym}}(i,j)$ and swapping indices in the second term gives

$$\frac{1}{2} \sum_{i,j} \pi(i) \boldsymbol{P}_{\text{sym}}(i,j)(x_i - x_j)^2 = \frac{1}{2} \sum_{i,j} \pi(i) P^{\mu}(i,j)(x_i - x_j)^2 \,.$$

This proves the claimed identity and $\boldsymbol{L}_{\text{Dir}} \succeq 0$. ■

Lemma 24 (Logarithmic stepsize estimates) *The following estimates hold, where* $\log(\cdot)$ *denotes the natural logarithm.*

(i) Square summability. *Let* $(a_t)_{t \geq 0}$ *be defined by*

$$a_t = \frac{1}{\sqrt{t+1}\,\log(t+3)} \qquad \forall t \geq 0 \,.$$

Define the tail sums

$$Q_T \;:=\; \sum_{t=T}^{\infty} a_t^2, \qquad T \in \mathbb{N} \cup \{0\} \,.$$

Then, for all integers $T \geq 0$*,*

$$Q_T \;\leq\; \frac{3}{\log(T+2)} \,. \tag{40}$$

In particular,

$$\sum_{t=0}^{\infty} a_t^2 \;\leq\; \frac{3}{\log 2} \;<\; \infty \,.$$

(ii) Partial-sum lower bound. *Let* $p \geq 0$*,* $\kappa > 0$*, and let* $(\eta_t)_{t \geq 0}$ *be defined by*

$$\eta_t = \frac{1}{\kappa\sqrt{t+1}\,\log^p(t+3)}, \qquad t \geq 0 \,.$$

For every integer $T \geq 1$*, define*

$$S_T := \sum_{t=1}^{T} \eta_{t-1} \,.$$

Then

$$S_T \geq \frac{2(\sqrt{T+1} - 1)}{\kappa \log^p(T+2)} \,.$$

Proof For (i), by definition,

$$a_t^2 = \frac{1}{(t+1)\ \log^2(t+3)}.$$

For every $t \geq 0$, we have $t+1 \geq \frac{t+3}{3}$, hence

$$\frac{1}{(t+1)\log^2(t+3)} \leq \frac{3}{(t+3)\log^2(t+3)}.$$

Therefore,

$$Q_T \leq \sum_{t=T}^{\infty} \frac{3}{(t+3)\log^2(t+3)}.$$

Let $n = t+3$. Then

$$\sum_{t=T}^{\infty} \frac{1}{(t+3)\log^2(t+3)} = \sum_{n=T+3}^{\infty} \frac{1}{n\log^2 n}.$$

Now define $f(x) = \frac{1}{x\log^2 x}$ for $x > 1$. Since f is decreasing on $(1, \infty)$, the integral test yields, for all $N \geq 3$,

$$\sum_{n=N}^{\infty} f(n) \leq \int_{N-1}^{\infty} f(x)\, dx = \int_{N-1}^{\infty} \frac{1}{x\log^2 x}\, dx = \frac{1}{\log(N-1)}.$$

Applying this with $N = T+3$ gives

$$\sum_{n=T+3}^{\infty} \frac{1}{n\log^2 n} \leq \frac{1}{\log(T+2)}.$$

Substituting back proves (40). Taking $T = 0$ gives the claimed finite upper bound on $\sum_{t=0}^{\infty} a_t^2$.

For (ii), let $s = t$. Then

$$S_T = \frac{1}{\kappa}\sum_{s=1}^{T} \frac{1}{\sqrt{s}\ \log^p(s+2)}.$$

Since $p \geq 0$, $\log(\cdot)$ is increasing, and $s+2 \leq T+2$ for all $1 \leq s \leq T$,

$$\frac{1}{\sqrt{s}\ \log^p(s+2)} \geq \frac{1}{\sqrt{s}\ \log^p(T+2)}.$$

Therefore,

$$S_T \geq \frac{1}{\kappa\log^p(T+2)} \sum_{s=1}^{T} \frac{1}{\sqrt{s}}.$$

The function $x \mapsto x^{-1/2}$ is decreasing on $[1, \infty)$, hence

$$\sum_{s=1}^{T} \frac{1}{\sqrt{s}} \geq \int_{1}^{T+1} \frac{1}{\sqrt{x}}\, dx = 2(\sqrt{T+1} - 1).$$

Combining the last two displays proves the claim. ■