Journal of Machine Learning Research 27 (2026) 1-49 Submitted 6/25; Revised 5/26; Published 11/26

Extensions of Robbins-Siegmund Theorem with Applications in Reinforcement Learning

Xinyu Liu XINYULIU@VIRGINIA.EDU
Department of Computer Science
University of Virginia
85 Engineer's Way, Charlottesville, VA, 22903, United States

Zixuan Xie XIE.ZIXUAN@EMAIL.VIRGINIA.EDU
Department of Computer Science
University of Virginia
85 Engineer's Way, Charlottesville, VA, 22903, United States

Shangtong Zhang SHANGTONG@VIRGINIA.EDU
Department of Computer Science
University of Virginia
85 Engineer's Way, Charlottesville, VA, 22903, United States

Editor: Nan Jiang

Abstract

The Robbins-Siegmund theorem establishes the convergence of stochastic processes that are almost supermartingales and is one of the most commonly used approaches for analyzing stochastic iterative algorithms in stochastic approximation and reinforcement learning (RL). However, its original form has a significant limitation as it requires the zero-order term to be summable. In many important RL applications, this summable condition, however, cannot be met. This limitation motivates us to extend the Robbins-Siegmund theorem for almost supermartingales where the zero-order term is not summable, but only square-summable. In particular, we introduce a novel and mild assumption on the increments of the stochastic processes. This together with the square-summable condition enables an almost sure convergence to a bounded set. Additionally, we further provide almost sure convergence rates, high probability concentration bounds, and L^p convergence rates. We then apply the new results to stochastic approximation and RL. Notably, we obtain the first almost sure convergence rate, the first high probability concentration bound, and the first L^p convergence rate for Q-learning with linear function approximation.

Keywords: Robbins-Siegmund theorem, stochastic approximation, reinforcement learning, linear Q-learning, convergence to a set

1. Introduction

The supermartingale is an important concept in stochastic processes and the celebrated Doob's supermartingale convergence theorem (Doob, 1953) establishes the almost sure convergence of supermartingales and is arguably one of the most important results in the field of martingales. However, many stochastic processes of interest are not supermartingales and, as a result, Doob's theorem does not apply. Instead, they are only very close to supermartingales and admit some noisy perturbations from supermartingales. Robbins

©2026 Xinyu Liu, Zixuan Xie, Shangtong Zhang.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v27/25-2392.html.

and Siegmund (1971) call them almost supermartingales. The Robbins-Siegmund theorem establishes the almost sure convergence of almost supermartingales under mild conditions and can be used to analyze a wide range of reinforcement learning (RL) algorithms whose associated ODE is globally asymptotically stable (GAS). Examples of RL algorithms with GAS ODEs include linear temporal difference learning (Sutton, 1988) and tabular Q-learning (Watkins, 1989). Numerous methods have been developed to analyze those algorithms and some are based on the Robbins-Siegmund theorem, see, e.g., Qian et al. (2026). However, there is a class of RL algorithms whose associated ODE is not GAS, e.g., linear SARSA (Rummery and Niranjan, 1994) and linear Q-learning (Baird, 1995). Linear SARSA is known to chatter inside a ball instead of converging to a single point, regardless of the choice of the learning rate (Gordon, 1996, 2000). Linear Q-learning in general can diverge (Baird, 1995). But with proper choices of the behavior policy and learning rate, Meyn (2024) shows that it can remain bounded almost surely. The follow-up work Liu et al. (2025) further shows that it can remain bounded in L^2. Such boundedness is usually the best one can hope for without imposing strong assumptions on the problem structure (Meyn, 2024). Although the Robbins-Siegmund theorem has enjoyed great success in RL algorithms with GAS ODEs, it falls short in analyzing RL algorithms with non-GAS ODEs. We argue that the root cause is that the standard Robbins-Siegmund theorem requires its zeroth-order to be summable. But for many RL algorithms with non-GAS ODEs such as linear SARSA and linear Q-learning, the resulting zeroth-order term is not summable and only square-summable. This calls for extensions of the Robbins-Siegmund theorem and this paper makes progress towards this goal with the following two contributions.

(i) We introduce a novel and mild assumption on the increments of almost supermartingales. This together with a square-summable condition on the zeroth-order term ensures an almost sure convergence to a bounded set for almost supermartingales, even if the zeroth-order term is not summable. Additionally, we further provide almost sure convergence rates, high probability concentration bounds, and L^p convergence rates, all to a bounded set.

(ii) Building on our results concerning almost supermartingales, we deliver new convergence results of general stochastic approximations with time-inhomogeneous Markovian noise, which is then used to analyze linear Q-learning, yielding the first almost sure convergence rate, the first high probability concentration bound, and the first L^p convergence rate, all to a bounded set.

To put (i) in the literature, we remark that recent notable extensions of the Robbins-Siegmund theorem include Liu and Yuan (2022); Karandikar and Vidyasagar (2024), which derive almost sure convergence rates for the Robbins-Siegmund theorem with a summable zeroth-order term. However, to our knowledge, no existing work is able to relax the summable condition on the zeroth-order term.

To put (ii) in the literature, we discuss RL algorithms with GAS and non-GAS ODEs separately. There are numerous works on RL algorithms with GAS ODEs (e.g., Tsitsiklis and Roy (1997); Abounadi et al. (2001); Sutton et al. (2008); Zhang et al. (2021a); Zhang and Whiteson (2022); Qian and Zhang (2025); Blaser et al. (2026)). Most related to this work is Lauand and Meyn (2024); Borkar et al. (2025). In particular, Borkar et al. (2025)

establish almost sure convergence and L^4 boundedness for RL algorithms with GAS ODEs (without a rate). Lauand and Meyn (2024) extend Borkar et al. (2025) by broadening the choice of learning rates and additionally establishing L^p convergence rates. For almost sure convergence, Lauand and Meyn (2024) still do not provide a rate. Borkar et al. (2025) also establish new results about central limit theorem which are not related to this work. So for the comparison axes relevant here, we focus on Lauand and Meyn (2024) and will discuss Borkar et al. (2025) only sparingly. Nevertheless, Lauand and Meyn (2024) is still not directly comparable to this work since it only applies to RL algorithms with GAS ODEs while this work focuses on RL algorithms with non-GAS ODEs. Numerous works also consider RL algorithms with non-GAS ODEs, e.g., policy gradient methods (Sutton et al., 1999), average-reward RL algorithms with global-clock learning rates (Antrobius and Zhang, 2026), and linear RL algorithms without full rank features (Xie et al., 2025; Wang and Zhang, 2026). Among those works, only a few are related to (ii) and the most notable ones are Meyn (2024) and Liu et al. (2025) that study the boundedness of linear Q-learning in an almost sure sense and L^2 sense respectively. Arguably, one may be able to combine the techniques of Lauand and Meyn (2024) and Meyn (2024). Such a hypothetical combination could potentially be used to analyze linear Q-learning to further provide an L^p convergence rate beyond the almost sure boundedness result of Meyn (2024). We, however, remark that such a combination has not been carried out. Furthermore, even if this combination were developed successfully, it would only affect our claim about "the first L^p convergence rate" in (ii). It is unclear whether such a combination would yield almost sure convergence rates and nonasymptotic high probability concentration bounds with sub-Weibull tails. Moreover, our results are applicable to not only linear Q-learning but also linear SARSA, similar to how Meyn (2024) applies to both linear Q-learning and linear SARSA. We focus only on linear Q-learning in our presentation to avoid verbosity.

2. Background

This section reviews the classical Robbins-Siegmund theorem and some latest extensions. We consider a probability space $(\Omega, \mathcal{F}, \Pr)$ equipped with a filtration $\{\mathcal{F}_n\}_{n\geq 0}$, where $\mathcal{F}_n$ represents the information available up to time n. For any random variable z, we denote the conditional expectation with respect to $\mathcal{F}_n$ by $\mathbb{E}_n[z] \doteq \mathbb{E}[z|\mathcal{F}_n]$. The Robbins-Siegmund theorem is concerned with the convergence of a non-negative real-valued random sequence $\{z_n\}_{n\geq 0}$ satisfying

$$\mathbb{E}_n[z_{n+1}] \leq (1+a_n)z_n + x_n - y_n \quad \text{a.s.}, \tag{RS}$$

where $\{z_n\}_{n\geq 0}$, $\{a_n\}_{n\geq 0}$, $\{x_n\}_{n\geq 0}$, and $\{y_n\}_{n\geq 0}$ are **non-negative** real-valued random sequences adapted to $\{\mathcal{F}_n\}_{n\geq 0}$. This non-negative condition is always imposed in the rest of the paper. Robbins and Siegmund (1971) call such $\{z_n\}$ almost supermartingales. Indeed, if $a_n = x_n = 0$, then $\{z_n\}$ immediately becomes a supermartingale.[1] We now present the Robbins-Siegmund theorem.

[1]This additionally requires verifying that $\{z_n\}$ is bounded in L_1, which is an easy exercise given (RS) and the non-negativity of $\{z_n\}$.

Lemma 1 (Theorem 1 of Robbins and Siegmund 1971) *Consider* $\{z_n\}$ *in* (RS). *Assume*

(A1) $\sum_{n=0}^{\infty} a_n < \infty$ *a.s.;*

(A2) $\sum_{n=0}^{\infty} x_n < \infty$ *a.s.*

Then

1. $\lim_{n\to\infty} z_n$ *exists and is finite a.s.;*
2. $\sum_{n=0}^{\infty} y_n < \infty$ *a.s.*

Here, the zero-order term in Section 1 refers to $\{x_n\}$ and the summable condition refers to (A2). Lemma 1 is an asymptotic result, which is recently extended to a non-asymptotic version by Liu and Yuan (2022); Karandikar and Vidyasagar (2024), under some additional assumptions.

Lemma 2 (Theorem 2 of Karandikar and Vidyasagar 2024) *Consider* $\{z_n\}$ *in* (RS). *Let the non-negative real-valued random sequence* $\{b_n\}_{n\geq 0}$ *be such that* $y_n = b_n z_n$. *For some* $\eta \in (0,1)$, *assume that*

(A1) $\sum_{n=0}^{\infty} a_n < \infty$, $\sum_{n=0}^{\infty} x_n < \infty$, $\sum_{n=0}^{\infty} b_n = \infty$ *a.s.;*

(A2) There exists some $n_1 \geq 1$ *such that for all* $n \geq n_1$, $b_n \geq \frac{\eta}{n}$ *a.s.;*

(A3) $\sum_{n=1}^{\infty} (n+1)^{\eta} x_n < \infty$, $\sum_{n=1}^{\infty} (b_n - \frac{\eta}{n}) = \infty$ *a.s.*

Then $\lim_{n\to\infty} n^{\eta} z_n = 0$ *a.s.*

Here (A3) is still a summable condition on the zero-order term.

Sometimes practitioners (Bertsekas and Tsitsiklis, 1996; Zhang et al., 2020; Barakat and Bianchi, 2021; Sebbouh et al., 2021; Xin et al., 2021; Qian et al., 2026) consider a special identification of $\{a_n, x_n, y_n\}$ in (RS). Namely, they consider a deterministic sequence $\{T_n \in (0,1)\}_{n\geq 0}$, some positive scalars α, ξ, and use $a_n = 0$, $x_n = \xi T_n^2$, and $y_n = \alpha T_n z_n$, yielding

$$\mathbb{E}_n[z_{n+1}] \leq (1 - \alpha T_n) z_n + \xi T_n^2.$$

The $\{T_n\}$ usually results from the learning rates of stochastic approximation and reinforcement learning algorithms so it satisfies the Robbins-Monro condition, i.e., $\sum_{n=0}^{\infty} T_n = \infty, \sum_{n=0}^{\infty} T_n^2 < \infty$. Then naturally, the zero-order term ξT_n^2 is summable. However, many important RL algorithms (see Section 6 for examples) do not reduce to the nice form above. For those algorithms, the corresponding zero-order term is T_n instead of T_n^2, i.e.,

$$\mathbb{E}_n[z_{n+1}] \leq (1 - \alpha T_n) z_n + \xi T_n. \tag{RS-Special}$$

Unfortunately, the zero-order term is no longer summable and is only square-summable under the Robbins-Monro condition. Since $T_n \geq T_n^2$, this situation essentially allows more noise, making $\{z_n\}$ further away from being a supermartingale. To our knowledge, no

existing extension of the Robbins-Siegmund theorem is able to handle this situation. This is the gap that this paper aims to bridge.

Notations. For a real number z, we denote its positive part by $z^+ \doteq \max(z, 0)$. Note that the operator $(\cdot)^+$ satisfies

$$\left|z_1^+ - z_2^+\right| \leq |z_1 - z_2| \quad \forall z_1, z_2 \in \mathbb{R}. \tag{1}$$

This inequality can be easily proved via the identity $z^+ = \frac{1}{2}z + \frac{1}{2}|z|$ and the triangle inequality. For an interval $Z \subset \mathbb{R}$, we denote $d(z, Z) \doteq \inf_{\tilde{z} \in Z} |z - \tilde{z}|$ as the distance from z to Z.

3. Asymptotic Extensions of Robbins-Siegmund Theorem

We now proceed to analyze (RS-Special) under the Robbins-Monro condition $\sum_{n=0}^{\infty} T_n = \infty$ and $\sum_{n=0}^{\infty} T_n^2 < \infty$. Consider a special case where z_n is deterministic. Then (RS-Special) degenerates to

$$z_{n+1} \leq (1 - \alpha T_n) z_n + \xi T_n.$$

Under the Robbins-Monro condition, it is trivial to show that $\lim_{n\to\infty} d(z_n, [0, \xi/\alpha]) = 0$, i.e., $\{z_n\}$ converge to a bounded interval. Building on this special case, one might then optimistically conjecture that under the Robbins-Monro condition, the $\{z_n\}$ in (RS-Special) also converge to a bounded interval almost surely. Unfortunately, this is impossible. Below we provide an example where the $\{z_n\}$ in (RS-Special) can be unbounded almost surely even when the Robbins-Monro condition holds.

Example 1 (Divergence with Robbins-Monro Condition) *Define* $T_n \doteq (n+1)^{-3/4}$. *Let* $\{X_{n+1}\}_{n\geq 0}$ *be independent with*

$$X_{n+1} \doteq \begin{cases} A_n & \text{with probability } p_n, \\ 0 & \text{with probability } 1 - p_n, \end{cases} \qquad n \geq 0,$$

where $A_n \doteq (n+1)^{1/4}, p_n \doteq \frac{T_n}{A_n} = \frac{1}{n+1}$. *Starting from* $z_0 = 0$ *and with filtration* $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$, *define recursively*

$$z_{n+1} = (1 - T_n)\, z_n + X_{n+1}, \qquad n \geq 0.$$

Then

1. $\{z_n\}$ *satisfies* $\mathbb{E}_n[z_{n+1}] = (1 - T_n)z_n + T_n$, *i.e.,* (RS-Special)*;*
2. $\limsup_{n\to\infty} z_n = \infty$ *a.s.*

Proof $\{T_n\}$ apparently fulfills the Robbins-Monro condition. Since X_{n+1} is independent of $\mathcal{F}_n$ and $\mathbb{E}_n[X_{n+1}] = p_n A_n = T_n$, we have $\mathbb{E}_n[z_{n+1}] = (1 - T_n)z_n + T_n$. That is, (RS-Special) holds with $\alpha = \xi = 1$.

Define the events $\Xi_n \doteq \{X_{n+1} = A_n\}$. They are independent with $\Pr(\Xi_n) = p_n = 1/(n+1)$ and $\sum_n \Pr(\Xi_n) = \infty$. By the second Borel-Cantelli lemma, $\Pr(\{\Xi_n \text{ infinitely often}\}) = 1$. On Ξ_n,

$$z_{n+1} = (1 - T_n)z_n + A_n \geq A_n = (n+1)^{1/4}.$$

Since Ξ_n occurs infinitely often almost surely, $z_{n+1} > (n+1)^{1/4}$ occurs infinitely often almost surely, forcing $\limsup_{n\to\infty} z_n = \infty$ almost surely. ■

Example 1 shows that the Robbins-Monro condition alone does not prevent the potential divergence of $\{z_n\}$ in (RS-Special). Intuitively, the divergence in Example 1 results from the spikes of X_{n+1}. X_{n+1} is mostly 0 but occasionally spikes to an increasingly large value A_n. Between the spikes, the factor $1 - T_n$ contracts z_n. But the spikes occur infinitely often, making z_n grow unbounded almost surely. This motivates us to introduce additional assumptions to control such spikes, or equivalently, control the growth of z_{n+1} from z_n.

Theorem 3 *Consider* $\{z_n\}$ *in* (RS-Special) *with* $\{T_n \in (0,1)\}_{n\geq 0}$ *being a sequence of deterministic real numbers. Suppose*

(A1) $\sum_n T_n = \infty$ *and* $\sum_n T_n^2 < \infty$*;*

(A2) There exists some constant $B_{3,1}$*, such that for all* $n \geq 0$*,*

$$|z_{n+1} - z_n| \leq B_{3,1} T_n (z_n + 1) \quad a.s.$$

Then there exists some constant B_3*, such that*

$$\lim_{n\to\infty} d(z_n, [0, B_3]) = 0 \quad a.s.$$

Proof We will identify $B_3 = \frac{\xi}{\alpha}$. Let $u_n \doteq d(z_n, [0, \frac{\xi}{\alpha}]) = \left(z_n - \frac{\xi}{\alpha}\right)^+$. Now we consider two cases.

(1) When $u_n > 0$**.** Then $u_n = z_n - \frac{\xi}{\alpha}$, we have

$$u_{n+1}^2 = \left(\left(z_{n+1} - \tfrac{\xi}{\alpha}\right)^+\right)^2 \leq \left(z_{n+1} - \tfrac{\xi}{\alpha}\right)^2 = \left((z_{n+1} - z_n) + z_n - \tfrac{\xi}{\alpha}\right)^2 = ((z_{n+1} - z_n) + u_n)^2. \tag{2}$$

Thus

$$\begin{aligned}
\mathbb{E}_n[u_{n+1}^2] &\leq \mathbb{E}_n\Big[(u_n + (z_{n+1} - z_n))^2\Big] && \text{(By (2))} \\
&= u_n^2 + 2u_n\, \mathbb{E}_n[z_{n+1} - z_n] + \mathbb{E}_n\big[(z_{n+1} - z_n)^2\big] \\
&\leq u_n^2 + 2u_n(-\alpha T_n z_n + \xi T_n) + B_{3,1}^2 T_n^2 (z_n + 1)^2 \\
& && \text{(By (RS-Special) and Theorem 3 (A2))} \\
&= u_n^2 + 2u_n \cdot (-\alpha T_n u_n) + B_{3,1}^2 T_n^2 \Big(u_n + \tfrac{\xi}{\alpha} + 1\Big)^2 && \text{(Since } u_n = z_n - \tfrac{\xi}{\alpha}\text{)} \\
&\leq u_n^2 - 2\alpha T_n u_n^2 + 2B_{3,1}^2 T_n^2 u_n^2 + 2B_{3,1}^2 T_n^2 \Big(\tfrac{\xi}{\alpha} + 1\Big)^2 \\
&= \big(1 - 2\alpha T_n + 2B_{3,1}^2 T_n^2\big) u_n^2 + 2B_{3,1}^2 \Big(1 + \tfrac{\xi}{\alpha}\Big)^2 T_n^2.
\end{aligned}$$

(2) When $u_n = 0$**.** Then $z_n \leq \frac{\xi}{\alpha}$. By (1) we obtain

$$|u_{n+1}| = |u_{n+1} - u_n| \leq \Big|(z_{n+1} - \tfrac{\xi}{\alpha}) - (z_n - \tfrac{\xi}{\alpha})\Big| \leq |z_{n+1} - z_n|. \tag{3}$$

Thus

$$\begin{aligned}
\mathbb{E}_n\big[u_{n+1}^2\big] \leq & \mathbb{E}_n\Big[|z_{n+1}-z_n|^2\Big] && \text{(By (3))}\\
\leq & B_{3,1}^2(1+z_n)^2T_n^2 && \text{(By Theorem 3 (A2))}\\
\leq & B_{3,1}^2\Big(1+\tfrac{\xi}{\alpha}\Big)^2T_n^2 && \text{(Since } z_n\in\Big[0,\tfrac{\xi}{\alpha}\Big]\text{)}\\
\leq & \big(1-2\alpha T_n+2B_{3,1}^2T_n^2\big)u_n^2+2B_{3,1}^2\Big(1+\tfrac{\xi}{\alpha}\Big)^2T_n^2. && \text{(Since } u_n=0\text{)}
\end{aligned}$$

Combining two cases we have

$$\mathbb{E}_n\big[u_{n+1}^2\big] \leq \big(1-2\alpha T_n+2B_{3,1}^2T_n^2\big)u_n^2+2B_{3,1}^2\Big(1+\tfrac{\xi}{\alpha}\Big)^2T_n^2. \tag{4}$$

We now proceed to apply Lemma 1. We identify in (RS)

$$z_n\leftrightarrow u_n^2, a_n\leftrightarrow 2B_{3,1}^2T_n^2, x_n\leftrightarrow 2B_{3,1}^2\Big(1+\tfrac{\xi}{\alpha}\Big)^2T_n^2, y_n\leftrightarrow 2\alpha T_nu_n^2.$$

As $\sum_n T_n^2<\infty$, we have $\sum_n a_n<\infty$ and $\sum_n x_n<\infty$ a.s. Applying Lemma 1 then gives: 1. $\lim_{n\to\infty}u_n^2$ exists and is finite almost surely; 2. $\sum_{n=0}^{\infty}y_n<\infty$ a.s. That is, $\sum_{n=0}^{\infty}T_nu_n^2<\infty$ a.s.

We now illustrate that the only possibility is $\lim_{n\to\infty}u_n^2=0$ a.s. For a fixed sample path ω, consider $u_n(\omega)^2$ as a deterministic sequence (abbreviated as u_n^2). If the limit $\zeta_0\doteq\lim_{n\to\infty}u_n^2$ exists and $\zeta_0\in(0,\infty)$, then there exists N such that $u_n^2\geq\zeta_0/2$ for all $n\geq N$. Consequently,

$$\textstyle\sum_{n=N}^{\infty}T_nu_n^2\geq\frac{\zeta_0}{2}\sum_{n=N}^{\infty}T_n=\infty,$$

where the last equality holds because $\sum_{n=0}^{\infty}T_n=\infty$. This contradicts the convergence of $\sum_{n=0}^{\infty}T_nu_n^2$. By definition, we have $\zeta_0\geq 0$. Therefore, the only possibility is $\zeta_0=0$. That is, $0=\lim_{n\to\infty}u_n^2=\lim_{n\to\infty}\left(\left(z_n-\frac{\xi}{\alpha}\right)^+\right)^2=\lim_{n\to\infty}d\Big(z_n,\Big[0,\frac{\xi}{\alpha}\Big]\Big)^2$. Denoting $B_3\doteq\frac{\xi}{\alpha}$ then completes the proof. ■

As demonstrated by Example 1, the Robbins-Monro condition alone cannot prevent divergence when the zero-order term is only square-summable. To address this, Assumption (A2) of Theorem 3 explicitly controls the incremental growth by prescribing that z_{n+1} can grow at most linearly in z_n, gated by T_n. This additional structure successfully rules out the pathological spikes seen in Example 1. While exploring weaker growth conditions (e.g., sub-linear growth) to establish the minimal sufficient conditions remains an interesting open question, the affine form in (A2) is highly natural for our setting and is well suited to our proof technique. In particular, we will demonstrate in Section 6 that this additional assumption indeed trivially holds for important reinforcement learning applications, such as Q-learning with linear function approximation (cf. Lemma 13).

We now apply the same idea to study (RS). Similarly, we want to relax the summable assumption (A2) in Lemma 1. To this end, we additionally introduce a square-summable control on the growth between consecutive steps and an overall drift that is non-positive once the process is overly large.

Theorem 4 *Consider* $\{z_n\}$ *in* (RS). *Suppose that*

(A1) $\sum_n a_n < \infty$ *a.s.;*

(A2) $|z_{n+1} - z_n| \le b_n(z_n + 1)$ *a.s., where* $\{b_n\}_{n\ge 0}$ *is a non-negative real-valued random sequence adapted to* $\{\mathcal{F}_n\}_{n\ge 0}$ *and satisfies* $\sum_n b_n^2 < \infty$ *a.s.;*

(A3) There exists some constant B_4 *such that*

$$z_n > B_4 \implies x_n - y_n \le -c_n(z_n - B_4),$$

where $\{c_n\}_{n\ge 0}$ *is a non-negative real-valued random sequence adapted to* $\{\mathcal{F}_n\}_{n\ge 0}$ *and satisfies* $\sum_n c_n = \infty$ *a.s.*

Then

$$\lim_{n\to\infty} d(z_n, [0, B_4]) = 0 \quad a.s.$$

The proof follows a similar idea as in Theorem 3 and is provided in Appendix A.1. We now elaborate on the connection between Theorem 4 and Theorem 3. When reducing (RS) to (RS-Special), we set $a_n = 0$, $x_n = \xi T_n$, and $y_n = \alpha T_n z_n$, so that $x_n - y_n = -\alpha T_n(z_n - \xi/\alpha)$. This gives $B_4 = \xi/\alpha$ and $c_n = \alpha T_n$ in Theorem 4. With this reduction, Theorem 4 (A1) is absent in Theorem 3 since $a_n = 0$. Instead, Theorem 3 (A1) directly imposes the Robbins-Monro condition. The assumptions correspond as follows: The requirement $\sum_n T_n = \infty$ ensures that $1 - T_n$ generates enough contractive drift over time. Specifically, when $z_n > B_4 = \xi/\alpha$, we have $x_n - y_n = -\alpha T_n(z_n - \xi/\alpha) \le -c_n(z_n - B_4)$ with $c_n = \alpha T_n$, and $\sum_n T_n = \infty$ ensures $\sum_n c_n = \infty$, satisfying Theorem 4 (A3). The requirement $\sum_n T_n^2 < \infty$ ensures that the growth rate is square-summable, corresponding to $\sum_n b_n^2 < \infty$ in Theorem 4 (A2).

Notably, we are not claiming that Theorems 3 & 4 are stronger than Lemma 1. Indeed, we remove the summable assumption at the cost of introducing new assumptions and the convergence is to a bounded set instead of a point. Our key contribution and novelty lies in that we are the first to obtain convergence results for almost supermartingales without the summable assumption on the zero-order term. And according to Example 1, our results might be the best that one can hope for. Additionally, we do demonstrate in Section 6 that our results have significant applications in reinforcement learning.

4. Nonasymptotic Extensions of Robbins-Siegmund Theorem

Having established asymptotic extensions of the Robbins-Siegmund theorem, we in this section further establish nonasymptotic convergence rates. We start with (RS-Special) and an almost sure convergence rate.

Theorem 5 *Consider* $\{z_n\}_{n\ge 0}$ *in* (RS-Special) *and let all conditions in Theorem 3 hold. Additionally, assume there exists some constant* $\eta > 0$ *such that*

(A1) $\liminf_{n\to\infty} nT_n > \frac{\eta}{2\alpha}$;

(A2) $\sum_n (n+1)^\eta T_n^2 < \infty$.

Then

$$\lim_{n\to\infty} n^{\eta/2} d(z_n, [0, B_3]) = 0 \quad a.s.$$

Similar to the proof of Theorem 3 where we construct auxiliary sequences to invoke Lemma 1, we prove Theorem 5 by constructing auxiliary sequences to invoke Lemma 2, which is provided in Appendix B.1.

If T_n decays in an inverse linear manner, we can further obtain high probability concentration bound and L^p convergence rates.

Assumption 4.1 *$T_n = \frac{C_{4.1}}{n+n_0}$ for some positive constants n_0 and $C_{4.1} > 1/\alpha$.*

Theorem 6 *Consider $\{z_n\}$ in* (RS-Special)*, z_0 is deterministic, and let all conditions in Theorem 3 hold. Let Assumption 4.1 hold. Then there exist some integers B_6 and B_6' such that for any $\delta > 0$, it holds, with probability at least $1-\delta$, that for all $n \geq 0$,*

$$(d(z_n, [0, B_3]))^2 \leq B_6' \tfrac{1}{n+n_0} \left[\ln\left(\tfrac{B_6}{\delta}\right) + 1 + \ln(n+n_0)\right]^{B_6}.$$

We prove Theorem 6 by an inductive approach inspired by Chen et al. (2025), which is provided in Appendix B.2. We argue that Theorem 6 is a strong result as it is a maximal bound that simultaneously holds for all n and it achieves an exponential tail.

Corollary 7 *Consider $\{z_n\}$ in* (RS-Special)*, z_0 is deterministic, and let all conditions in Theorem 3 hold. Let Assumption 4.1 hold. Then for any integer $p \geq 2$ and any $n \geq 0$,*

$$\mathbb{E}\Big[(d(z_n, [0, B_3]))^{2p}\Big] \leq \big(a(n)b(n)^{B_6}\big)^p + B_6 p a(n)^p \exp(b(n))((B_6 p)!),$$

where $a(n) \doteq B_6'/(n+n_0)$, $b(n) \doteq \ln B_6 + 1 + \ln(n+n_0)$.

Corollary 7 can be obtained easily by integrating the sub-Weibull tails from Theorem 6. This integration procedure is identical to the proof of Corollary 1 in Qian et al. (2026) and is thus omitted.

Together, Theorems 5 & 6 and Corollary 7 provide a comprehensive set of nonasymptotic tools (including almost sure rates, high-probability concentration bounds, and L^p rates) for analyzing stochastic iterates that fit the (RS-Special) template. The following sections will demonstrate the power of this framework by applying it to stochastic approximation and linear Q-learning. We close this section by providing the counterpart of Theorem 5 for (RS). Notably, we do not provide the counterparts of Theorem 6 and Corollary 7 for (RS) because it is not clear how to embed the structure in Assumption 4.1 into (RS).

Theorem 8 *Consider $\{z_n\}$ in* (RS) *and let all the conditions of Theorem 4 hold. Additionally, assume there exists some constant $\eta > 0$ such that*

(A1) $\liminf_{n\to\infty} nc_n > \frac{\eta}{2}$*;*

(A2) $\sum_n (n+1)^\eta (a_n + b_n^2) < \infty$ *a.s.*

Then

$$\lim_{n\to\infty} n^{\eta/2} d(z_n, [0, B_4]) = 0 \quad a.s.$$

The proof follows a procedure similar to that of Theorem 5 and is provided in Appendix B.3.

5. Stochastic Approximation

We now use Theorems 5 & 6 and Corollary 7 to derive new convergence results for stochastic approximations. Specifically, we consider stochastic approximation algorithms in the form of

$$w_{t+1} = w_t + \alpha_t H(w_t, Y_{t+1}). \tag{SA}$$

Here $\{w_t \in \mathbb{R}^d\}$ is the stochastic iterates, w_0 is deterministic, $\{\alpha_t\}$ is a sequence of learning rates, $\{Y_t\}$ is a time-inhomogeneous Markov chain evolving in a finite space $\mathcal{Y}$, and $H : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}^d$ is the function that generates the incremental update. Importantly, we consider the setting where the transition kernel of $\{Y_t\}$ is controlled by $\{w_t\}$ itself. This setting admits important applications such as linear SARSA (Rummery and Niranjan, 1994; Zou et al., 2019; Zhang et al., 2023) and linear Q-learning (Baird, 1995; Meyn, 2024; Liu et al., 2025). In such settings, the convergence of $\{w_t\}$ to a point is usually not attainable (see, e.g., Zhang et al. 2023; Meyn 2024; Liu et al. 2025) unless strong assumptions are made (see, e.g., Zou et al. 2019). We, therefore, aim to establish the convergence to a bounded set instead. The high-level idea of our approach is to fit (SA) into the template (RS-Special). To this end, we use the skeleton iterates technique from Qian et al. (2026) to work with the Markovian nature of $\{Y_t\}$ and use a novel error decomposition from Liu et al. (2025) to work with the time-inhomogeneous nature of $\{Y_t\}$. One of our main contributions lies in a new analysis that successfully controls the complex bias terms arising from this integration, making these two powerful techniques compatible for the first time. After obtaining (RS-Special), our new results can then take over.

This contribution is crucial because Qian et al. (2026) have to assume that $\{Y_t\}$ is time-homogeneous so they can use the original Robbins-Siegmund theorem, while Liu et al. (2025) have to take total expectation on both sides of (RS-Special) to derive the rate at which the second moments of $\{w_t\}$ converge to a bounded interval. Liu et al. (2025) thus miss the opportunity to derive an almost sure convergence rate, a high probability concentration bound, and an L^p convergence rate. We now list our assumptions, many of which are adapted from Qian et al. (2026); Borkar et al. (2025); Liu et al. (2025).

Assumption LR1 (Assumption LR1 in Qian et al. 2026) *The learning rate has the form*

$$\alpha_t = \tfrac{C_\alpha}{(t+3)^\nu} \quad \textit{with} \quad \nu \in \left(\tfrac{2}{3}, 1\right].$$

Assumption LR2 (Assumption LR2 in Qian et al. 2026) *The learning rate has the form*

$$\alpha_t = \tfrac{C_\alpha}{(t+3)\ln^\nu(t+3)} \quad \textit{with} \quad \nu \in (0, 1).$$

The two forms of learning rates are from the skeleton iterates technique of Qian et al. (2026). Assumption LR1 is a standard choice of learning rate except that ν cannot be smaller than $2/3$, which is an inherent limitation of the skeleton iterates technique. This learning rate will be used for obtaining an almost sure convergence rate of $\{w_t\}$ by invoking Theorem 5. Assumption LR2 is specially designed such that the resulting T_n in (RS-Special) can verify Assumption 4.1. This learning rate will be used for obtaining high probability concentration bound and L^p convergence rate by invoking Theorem 6 and Corollary 7. We refer the reader to Qian et al. (2026) for more discussion about the two learning rates.

Assumption 5.1 (Assumption 3.1 & 3.2 in Zhang et al. 2022) *There exists a family of parameterized transition matrices* $\Lambda_P \doteq \left\{P_w \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{Y}|} \mid w \in \mathbb{R}^d\right\}$ *such that* $\Pr(Y_{t+1} \mid Y_t) = P_{w_t}(Y_t, Y_{t+1})$. *Furthermore, let* $\bar{\Lambda}_P$ *denote the closure of* Λ_P. *Then for any* $P \in \bar{\Lambda}_P$, *the time-homogeneous Markov chain induced by* P *is irreducible and aperiodic.*

Assumption 5.1 looks prohibitive but as can be seen soon in Section 6, this assumption can be trivially verified in applications. Under Assumption 5.1, for any w, the Markov chain induced by P_w has a unique stationary distribution, which we denote as $d_{\mathcal{Y},w}$. This allows us to define the expected update as $h(w) \doteq \mathbb{E}_{Y\sim d_{\mathcal{Y},w}}[H(w,Y)]$. Additionally, by Lemma 1 of Zhang et al. (2022), Assumption 5.1 implies that the Markov chains in Λ_P mix both geometrically and uniformly. In other words, there exist constants $C_{5.1} > 0$ and $\tau \in [0,1)$, independent of w, such that

$$\sup_{w,y} \sum_{y'} |P_w^n(y,y') - d_{\mathcal{Y},w}(y')| \le C_{5.1}\tau^n. \tag{5}$$

We then define

$$\tau_\alpha \doteq \min\{n \ge 0 \mid C_{5.1}\tau^n \le \alpha\} \tag{6}$$

to denote the number of steps that the Markov chain needs to mix to an accuracy α. We now introduce the norm we will work with in the rest of the paper. Let $\langle\cdot,\cdot\rangle$ be an arbitrary inner product for vectors in $\mathbb{R}^d$. Let $\|\cdot\|$ be the norm induced by this inner product (i.e., the inner product norm for vectors and the corresponding induced norm for matrices). For a non-negative real number r, we denote the closed ball in $\mathbb{R}^d$ centered at 0 with radius r by $\mathcal{B}(r) \doteq \left\{x \in \mathbb{R}^d \mid \|x\| \le r\right\}$.

Assumption 5.2 (Condition (30) in Borkar et al. 2025) *There exists a constant* $C_{5.2}$ *such that for* $\forall w_1, w_2$

$$\|P_{w_1} - P_{w_2}\| \le \frac{C_{5.2}}{1+\|w_1\|+\|w_2\|}\|w_1 - w_2\|.$$

Assumption 5.3 (Assumption (A2) in Borkar et al. 2025) *There exists some constant* L_h, *such that for* $\forall w_1, w_2, y$

$$\|H(w_1,y) - H(w_2,y)\| \le L_h\|w_1 - w_2\| \quad \textit{and} \quad \|H(0,y)\| \le L_h.$$

Assumption 5.2 is stronger than the standard Lipschitz continuity assumption (cf. Assumption 5.3) and is the key to handling the time-inhomogeneous nature of $\{Y_t\}$. The earliest work we are aware of using this stronger form of Lipschitz continuity to ensure the Lipschitz continuity of the following composition function is Konda (2002) (Assumption 5.1.3). More recent usages include Lauand and Meyn (2024); Borkar et al. (2025) and the follow-up work Liu et al. (2025).

Assumption 5.4 *There exist some deterministic constants* $C_{5.4,1} > 0$ *and* $C_{5.4,2}$ *such that*

$$\langle w, h(w)\rangle \le -C_{5.4,1}\|w\|^2 + C_{5.4,2}.$$

Assumption 5.4 ensures that the expected update direction $h(w)$ will decay $\|w\|$ at least when w is overly large. The factor $C_{5.4,2}$ is usually 0 when considering stochastic approximation algorithms with GAS ODEs. But for a stochastic approximation algorithm with a non-GAS ODE, see, e.g., Meyn (2024), it is usually strictly positive, constituting a main obstacle for analyzing the convergence of such algorithms.

Theorem 9 (Almost Sure Convergence Rate for SA) *Let Assumptions 5.1 - 5.4 and LR1 hold. Let $\{w_t\}$ be the iterates generated by* (SA)*. Then there exists some B_9, such that the following statements hold. If $\nu < 1$, then for any $\zeta \in \left(0, \frac{3}{4}\nu - \frac{1}{2}\right)$,*

$$\lim_{t\to\infty} t^{\zeta} d(w_t, \mathcal{B}(B_9)) = 0 \quad a.s.$$

If $\nu = 1$, then for any $\zeta \in (0, \frac{1}{2})$ and $\nu_1 > 0$,

$$\lim_{t\to\infty} \exp\left(\zeta \ln^{1/(1+\nu_1)} t\right) d(w_t, \mathcal{B}(B_9)) = 0 \quad a.s.$$

Theorem 10 (Concentration for SA) *Let Assumptions 5.1 - 5.4 and LR2 hold. Then for any $\nu \in (0,1)$, there exist some deterministic constants $\bar{C}_\alpha, B_{10}, B'_{10}$, and integer C_{27} such that the iterates $\{w_t\}$ generated by* (SA) *satisfy the following concentration property whenever $C_\alpha \geq \bar{C}_\alpha$: for any $\delta \in (0,1)$, it holds, with probability at least $1-\delta$, that for all $t \geq 0$,*

$$d(w_t, \mathcal{B}(B_9)) \leq B_{10} \exp\left(\frac{-\ln^{1-\nu}(t+1)}{2(1-\nu)}\right)\left[\ln\left(\frac{1}{\delta}\right) + B'_{10} + \frac{\ln^{1-\nu}(t+1)}{1-\nu}\right]^{C_{27}/2}.$$

Corollary 11 (L^p Convergence Rates for SA) *Let conditions of Theorem 10 hold. Then the iterates $\{w_t\}$ generated by* (SA) *satisfy for any integer $p > 2$ and any $t \geq 0$*

$$\mathbb{E}[(d(w_t, \mathcal{B}(B_9)))^p] \leq \left(a(t) b(t)^{\lceil C_{27}/2 \rceil}\right)^p + \lceil C_{27}/2 \rceil p a(t)^p \exp(b(t))((\lceil C_{27}/2 \rceil p)!),$$

where $a(t) \doteq B_{10} \exp\left(\frac{-\ln^{1-\nu}(t+1)}{2(1-\nu)}\right)$, $b(t) \doteq B'_{10} + \frac{\ln^{1-\nu}(t+1)}{1-\nu}$.

We shall shortly present a sketch to highlight the key idea in proving Theorems 9 & 10 and Corollary 11 with the complete proof provided in Appendix C. Table 1 highlights our improvements over prior works. In particular, the focus of this table is stochastic approximation algorithms with non-GAS ODEs and time-inhomogeneous Markovian noise. To our knowledge, prior work of this kind is thin and we are only aware of Liu et al. (2025) and Theorem II.1 of Meyn (2024). We additionally include results about stochastic approximation with GAS ODEs only when they are highly relevant to our results. For instance, we additionally include Qian et al. (2026) because we rely on their skeleton iterates technique. We additionally include Borkar et al. (2025); Lauand and Meyn (2024) because their techniques can be potentially combined with Meyn (2024) to generate results comparable to ours. Furthermore, regarding specific metric comparisons in Table 1, Liu et al. (2025) establish the convergence rate of $d\Big(\mathbb{E}[\|w_t\|^2], [0, B^2]\Big)$, i.e., the rate at which the second moment decreases to below some threshold. By contrast, this work establishes L^2 convergence to a set,[2] i.e., the convergence rate of $\mathbb{E}[d(w_t, \mathcal{B})^2]$, which is stronger than Liu et al. (2025).

[2] An L^2 rate follows immediately from an L^p rate with any $p > 2$.

	Convergence Modes			Metric	P_{w_t}
	a.s.	$\Pr(\ldots) \geq 1-\delta$	L^p		
Qian et al. (2026)	✓	✓	✓	Distance to a point	
Borkar et al. (2025)	𝒻		𝒻	Distance to a point	✓
Lauand and Meyn (2024)	𝒻		✓	Distance to a point	✓
Theorem II.1 of Meyn (2024)	𝒻			Distance to a set	✓
Hypothetical Combination of Meyn (2024) Lauand and Meyn (2024)	𝒻		✓	Distance to a set	✓
Liu et al. (2025)			✓ ($p=2$)	$\mathbb{E}[\|w_t\|^2]$	✓
This work	✓	✓	✓	Distance to a set	✓

Table 1: Comparison with prior works about stochastic approximation with time-inhomogeneous Markovian noise. "a.s." is checked if an almost sure convergence is established. "$\Pr(\ldots) \geq 1-\delta$" is checked if high-probability convergence with sub-Weibull tails is established. "L^p" is checked if L^p convergence is established for all $p \geq 2$, unless otherwise noted. "✓" means the convergence rate is established for the corresponding mode of convergence, while " 𝒻 " means the convergence is established but the rate is not provided. "P_{w_t}" is checked if the Markov chain $\{Y_t\}$ is allowed to be time-inhomogeneous.

Remark 12 (Improvement over Liu et al. 2025) *Define*

$$D_t \doteq d(w_t, \mathcal{B}(B)) = \max(\|w_t\| - B, 0) \geq 0.$$

We have

$$\max\left(\|w_t\|^2 - B^2, 0\right) = D_t^2 + 2BD_t.$$

Indeed, when $\|w_t\| < B$, both sides are 0. When $\|w_t\| \geq B$, we have

$$\max\left(\|w_t\|^2 - B^2, 0\right) = \|w_t\|^2 - B^2 = (\|w_t\| - B)(\|w_t\| + B) = D_t(D_t + 2B) = D_t^2 + 2BD_t.$$

By the convexity of max *and Jensen's inequality, we have*

$$\max\left(\mathbb{E}[\|w_t\|^2] - B^2, 0\right) \leq \mathbb{E}\left[\left(\|w_t\|^2 - B^2\right)^+\right] = \mathbb{E}[D_t^2] + 2B\mathbb{E}[D_t] \leq \mathbb{E}[D_t^2] + 2B\sqrt{\mathbb{E}[D_t^2]},$$

which means that the convergence of $\mathbb{E}[d(w_t, \mathcal{B}(B))^2]$ implies the convergence of $d(\mathbb{E}[\|w_t\|^2], [0, B^2])$. Thus, even considering only the case $p = 2$, this work still constitutes a significant improvement over Liu et al. (2025).

Proof (Proof idea of Theorems 9 & 10 and Corollary 11) We rely on the skeleton iterates technique from Qian et al. (2026). This technique is essentially a significant refinement of the proof technique used in the proof of Proposition 4.8 of Bertsekas and Tsitsiklis (1996) and the foundation of this technique is recently formally verified in Lean by Zhang (2026). Instead of examining $\{w_t\}$ along the natural time scale $t = 0, 1, \ldots$, we examine a subsequence $\{w_{t_m}\}$ indexed by $\{t_m\}$. The subsequence $\{t_m\}$ is carefully designed such that the magnitude of updates between consecutive times $T_m = \sum_{t=t_m}^{t_{m+1}-1} \alpha_t$ satisfies the Robbins-Monro condition. We are then able to recover the template (RS-Special) by identifying $z_m = \|w_{t_m}\|^2$. In this process, we need to use an error decomposition technique from Liu et al. (2025) to work with the time-inhomogeneous nature of $\{Y_t\}$, thanks to the stronger notion of Lipschitz continuity in Assumption 5.2.

The main challenge here is to bound the bias term, denoted $s_{4,m}$ in Appendix C, which arises because the Markov chain $\{Y_t\}$ is time-inhomogeneous. Our analysis pivots on a key insight: we handle this term by dissecting the proof into two cases, depending on the relative positions of the main timescale t_m and the mixing time timescale $t - \tau_{\alpha_t}$. For the case where $t_m \geq t - \tau_{\alpha_t}$, we leverage the geometric mixing property of the Markov chain directly. For the case where $t_m < t - \tau_{\alpha_t}$, we employ the tower rule to shift the conditional expectation, allowing us to control the error. In both scenarios, this meticulous case analysis successfully demonstrates that the bias is bounded within $\mathcal{O}(T_m^2)$.[3] This crucial step ensures that the total error is properly controlled.

After recovering (RS-Special), all the rest is to verify assumptions of Theorems 5 & 6 and Corollary 7. In particular, we verify the assumption on the growth (Theorem 3 (A2)) below.

Lemma 13 *There exists some deterministic constants m_0 and C_{13}, such that for all $m \geq m_0$,*

$$|z_{m+1} - z_m| \leq C_{13} T_m (z_m + 1).$$

Proof

$$\begin{aligned}
\left| \left\|w_{t_{m+1}}\right\|^2 - \left\|w_{t_m}\right\|^2 \right| =& \left| \langle w_{t_{m+1}} - w_{t_m}, w_{t_{m+1}} + w_{t_m} \rangle \right| \\
\leq& \left\|w_{t_{m+1}} - w_{t_m}\right\| \times \left\|w_{t_{m+1}} + w_{t_m}\right\| \\
\leq& 2T_m C_{28}(\|w_{t_m}\| + 1) \times (2\|w_{t_m}\| + 2T_m C_{28}(\|w_{t_m}\| + 1)) \quad \text{(By (17))} \\
\leq& 4T_m C_{28}(1 + T_m C_{28})(\|w_{t_m}\|^2 + 1).
\end{aligned}$$

■

Let $C_{13} \doteq 4C_{28}(1 + T_0 C_{28})$. As promised before, this affine growth assumption is indeed easy to verify. We recall that a full proof is provided in Appendix C. ■

[3] For two non-negative sequences $\{a_m\}$ and $\{b_m\}$, we say $a_m = \mathcal{O}(b_m)$ if there exist some constant $C > 0$ and an integer m_0 such that $a_m \leq C b_m$ for all $m \geq m_0$.

6. Applications in Reinforcement Learning

In this section, we apply the results in Section 5 to derive new convergence results of reinforcement learning (RL, Sutton and Barto 2018) algorithms. We focus on linear Q-learning, though our results can also be applied to linear SARSA (Rummery and Niranjan, 1994) to advance the understanding of the long-standing "chattering" problem of linear SARSA (Gordon, 1996).

Q-learning (Watkins and Dayan, 1992) is one of the most celebrated RL algorithms. When combined with linear function approximation, the resulting algorithm, linear Q-learning, is largely believed to suffer from possible instability for decades, i.e., the iterates from linear Q-learning can possibly diverge to infinity. This issue exhibits the deadly triad in RL (Sutton and Barto, 2018; Zhang, 2022). Numerous efforts have been made in prior works to tweak the linear Q-learning algorithm to achieve stability. See Section 7 and Table 1 of Liu et al. (2025) for a more detailed survey. However, recent works (Meyn, 2024; Liu et al., 2025) confirm that linear Q-learning is guaranteed to be stable when paired with an ϵ-softmax behavior policy with an adaptive temperature, without any algorithm tweaks or strong assumptions. This work builds upon and improves over Meyn (2024); Liu et al. (2025). We now present our new convergence results for linear Q-learning.

Problem Setup. We consider an infinite horizon Markov Decision Process (MDP, Bellman 1957; Puterman 1994) with a finite state space $\mathcal{S}$, a finite action space $\mathcal{A}$, a reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, a transition function $p : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$, an initial distribution $p_0 : \mathcal{S} \to [0, 1]$, and a discount factor $\gamma \in [0, 1)$. At time 0, an initial state S_0 is sampled from p_0. At time t, an agent at a state S_t takes an action sampled from some policy $\pi : \mathcal{A} \times \mathcal{S} \to [0, 1]$.

The action-value function $q_\pi(s, a)$ represents the expected discounted reward starting from state s, taking action a, and following policy π thereafter

$$q_\pi(s, a) = \mathbb{E}_\pi\big[\textstyle\sum_{i=0}^{\infty} \gamma^i R_{t+i+1} \mid S_t = s, A_t = a\big].$$

The optimal action-value function is defined as $q_*(s, a) = \max_\pi q_\pi(s, a)$, which characterizes the maximum achievable value for each state-action pair. Theoretically, q^* is the unique solution to the Bellman optimality equation. The Q-learning algorithm is fundamentally an off-policy method designed to find this solution, meaning it can learn the target policy while using a different, exploratory behavior policy π to gather data. At each time step, the behavior policy π provides an action $A_t \sim \pi(\cdot|S_t)$, and the collected tuple $(S_t, A_t, R_{t+1}, S_{t+1})$ is used for learning.

In linear Q-learning, the optimal action-value function q^* is approximated using a linear parameterization $q^*(s, a) \approx x(s, a)^\top w$, where $x : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ is a feature mapping and $w \in \mathbb{R}^d$ is the parameter vector. The update rule for linear Q-learning is given by

$$\begin{aligned} &A_t \sim \mu_{w_t}(\cdot|S_t), && \text{(linear } Q\text{-learning)} \\ &w_{t+1} = w_t + \alpha_t(R_{t+1} + \gamma \max_a x(S_{t+1}, a)^\top w_t - x(S_t, A_t)^\top w_t)x(S_t, A_t). \end{aligned}$$

Here, $\{\alpha_t\}$ are learning rates and μ_w is the behavior policy

$$\mu_w(a|s) = \tfrac{\epsilon}{|\mathcal{A}|} + (1 - \epsilon)\tfrac{\exp\big(\kappa_w x(s,a)^\top w\big)}{\sum_b \exp\big(\kappa_w x(s,b)^\top w\big)},$$

where $\epsilon \in (0,1)$ sets a baseline level of uniform exploration. The adaptive temperature, $\kappa_w \doteq \frac{\kappa_0}{\max(1,\|w\|)}$ (for some constant $\kappa_0 > 0$), further influences the policy's exploration behavior by adjusting the greediness of the softmax component. This is exactly the tamed Gibbs policy introduced by Meyn (2024) and is also used later on by Liu et al. (2025).

The instability of linear Q-learning (the "deadly triad") is primarily driven by a positive feedback loop: as the parameter norm $\|w\|$ grows due to approximation errors, standard policies become overly greedy, exploiting these errors and pushing $\|w\|$ to infinity. To solve this, Meyn (2024) introduces the tamed Gibbs policy. By scaling the temperature inversely with $\|w\|$, it prevents the policy from collapsing into pure exploitation. This acts as a "restoring force" that safely maintains exploration at large parameter scales. Based on our theoretical analysis, the design of a stabilizing behavior policy should generally satisfy two core requirements: (1) it must adaptively align with w to induce a mean-field inward drift at large scales (as captured by our Assumption 5.4), and (2) it must maintain global Lipschitz continuity (Assumption 5.2) and sufficient exploration (Assumption 5.1) rather than degenerating into a discontinuous hard-max. We believe any search for new suitable behavior policies should follow this direction to ensure the stability of linear Q-learning. For example, we conjecture that replacing softmax with sparsemax or Gaussian-perturbed greedy selection, while retaining the same inverse-norm scaling and uniform exploration, would likewise stabilize linear Q-learning under suitable parameter choices.

Assumption 6.1 *The Markov chain $\{S_t\}$ induced by a uniformly random behavior policy is irreducible and aperiodic.*

Here, a uniformly random behavior policy refers to the policy that selects each action with equal probability, i.e., $\pi(a|s) = 1/|\mathcal{A}|$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$. This assumption ensures the Markov chain induced by the tamed Gibbs policy μ_w is also irreducible and aperiodic, since μ_w explicitly incorporates a baseline level of uniform exploration via its $\epsilon/|\mathcal{A}|$ term.

Assumption 6.2 *Let $X \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times d}$ be the feature matrix whose rows are given by $x(s,a)^\top$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. We assume that X has full column rank, i.e.,* $\mathrm{rank}(X) = d$.

Theorem 14 (Convergence of Linear Q-Learning) *Let Assumptions 6.1 and 6.2 hold. There exist some $\epsilon_0 > 0$ and κ_1, such that for any $\epsilon < \epsilon_0$ and $\kappa_0 > \kappa_1$, if Assumption LR1 holds, then the statements about $\{w_t\}$ in Theorem 9 hold for $\{w_t\}$ generated by* (linear Q-learning)*. If Assumption LR2 holds, then the statements about $\{w_t\}$ in Theorem 10 and Corollary 11 hold for $\{w_t\}$ generated by* (linear Q-learning)*.*

We remark that similar to prior works establishing the stability of linear Q-learning (Meyn, 2024; Liu et al., 2025), our analysis is restricted to settings with finite state and action spaces. Table 2 contextualizes the contribution of Theorem 14 over prior works, which is essentially inherited from Table 1. To make it more clear, using the notations in Theorem 9, Meyn (2024) shows that $\lim_{t\to\infty} d(w_t, \mathcal{B}(B_9)) = 0$ a.s. but with a broader choice of learning rates, including $\nu \in (1/2, 1]$. The hypothetical combination of Meyn (2024) and Lauand and Meyn (2024) can further broaden the learning rates to include $\nu \in (0, 1]$ but still does not have almost sure convergence rates. Neither Meyn (2024) nor the hypothetical combination provides high-probability concentration bounds with sub-Weibull tails (cf. Theorem 10). For

	a.s.	$\Pr(\ldots) \geq 1-\delta$	L^p	Metric
Meyn (2024)	✓			Distance to a set
Hypothetical Combination of Meyn (2024) Lauand and Meyn (2024)	✓		✓	Distance to a set
Liu et al. (2025)			✓ ($p=2$)	$\mathbb{E}[\|w_t\|^2]$
This work	✓	✓	✓	Distance to a set

Table 2: Comparison with prior works about the convergence of linear Q-learning without algorithm tweaks or strong assumptions. "a.s." is checked if an almost sure convergence is established. "$\Pr(\ldots) \geq 1-\delta$" is checked if convergence with a high probability with sub-Weibull tails is established. "L^p" is checked if L^p convergence is established for all $p \geq 2$, unless otherwise noted. "✓" means the convergence rate is established for the corresponding mode of convergence, while " ✓ " means the convergence is established but the rate is not provided.

L^p convergence rates, the hypothetical combination can provide better convergence rates and allow broader learning rates than Corollary 11. We close this section with the following proof of Theorem 14.

Proof We proceed by fitting (linear Q-learning) into (SA) and then verifying the assumptions of Theorems 9 & 10 and Corollary 11. Notably, the (linear Q-learning) we study is exactly the same as that of Meyn (2024); Liu et al. (2025) so we can reuse many of their results. First, by defining

$$\begin{aligned} Y_{t+1} \doteq& (S_t, A_t, S_{t+1}),\ y \doteq (s, a, s'), \\ \mathcal{Y} \doteq& \{(s \in \mathcal{S}, a \in \mathcal{A}, s' \in \mathcal{S}) \mid p(s'|s,a) > 0\}, \\ H(w, y) \doteq& (r(s,a) + \gamma \max_{a'} x(s', a')^\top w - x(s,a)^\top w) x(s,a), \end{aligned}$$

we are able to fit (linear Q-learning) into (SA). With the identification above, Assumptions 5.1 - 5.4 are already verified in Section 5.2 of Liu et al. (2025), which immediately completes the proof. In other words, we are able to achieve stronger results (thanks to our extensions of the Robbins-Siegmund theorem) with the same set of assumptions. In particular, Assumption 5.1 is immediately implied by Assumption 6.1 (see Section B.2 of Zhang et al. 2022 & Section 5.2 of Liu et al. 2025). Thus, as promised, Assumption 5.1 can indeed be trivially satisfied in RL applications. Assumption 5.4 is verified by Lemma 7 in Liu et al. (2025), which originates from Lemma A.9 of Meyn (2024). ■

7. Related Work

This section reviews related work on learning rates, convergence guarantees, and linear Q-learning.

Learning rates. In this work, we always assume that the learning rates are square-summable and diminishing. There is rich literature on stochastic approximation and RL where the learning rates are not square-summable but still diminishing. For instance, Moulines and Bach (2011) provide a non-asymptotic analysis showing that slower-decaying, non-square-summable learning rates of the form $\mathcal{O}(k^{-\alpha})$ ($\alpha \in (0,1)$), when coupled with Polyak-Ruppert averaging, robustly achieve optimal convergence rates for minimizing convex objective functions in both finite and infinite-dimensional settings. Toulis and Airoldi (2017) further investigate the asymptotic and finite-sample properties of stochastic gradients by introducing implicit stochastic gradient descent. They demonstrate that implicit updates combined with trajectory averaging maintain numerical stability and optimal statistical efficiency even when employing large learning rates with non-summable variances. Extending this to nonparametric least-squares regression, Dieuleveut and Bach (2016) prove that strictly non-square-summable step-sizes (such as constant or very slowly decaying rates) can still attain statistically optimal non-asymptotic convergence rates when trajectory averaging is applied. Building on these foundational insights, recent literature continues to expand the range of admissible step-size schedules. For instance, Chandak and Borkar (2026) specifically focus on contractive stochastic approximation and RL algorithms, deriving high-probability concentration bounds for strictly non-square-summable step sizes without invoking $\sum \alpha_n^2 < \infty$. Furthermore, Lauand and Meyn (2024) recently revisit standard step-size assumptions, demonstrating that the strict square-summability condition can be relaxed while still guaranteeing L^p convergence under time-inhomogeneous Markovian noise. Concurrently, a growing body of work has established fine-grained, non-asymptotic concentration bounds and central limit theorems for Polyak-Ruppert averaged stochastic approximation and Q-learning under slowly decaying, non-square-summable step-sizes (Mou et al., 2020; Li et al., 2023; Mou et al., 2024; Samsonov et al., 2024; Khodadadian and Zubeldia, 2025). One important question is whether we can analyze (RS-Special) with techniques from those works. Our evaluation is negative. Let $T_n = (n+1)^{-1/2}$ and $A_n \doteq (n+1)^{1/2}$ in Example 1. It can be similarly verified that $\{z_n\}$ in Example 1 can be unbounded almost surely. This means that at least we cannot directly invoke those works to analyze (RS-Special) since they typically conclude that the iterates converge to a single point, which conflicts with our new version of Example 1. Can we extend those works to allow convergence to a bounded set instead of a single point to analyze (RS-Special)? Our answer is unknown. We are not excluding this possibility but we are not pursuing this direction in this paper and we are not aware of any existing work that has done this extension.

There is also a rich literature on stochastic approximation and RL with constant learning rates. The foundational theories for analyzing the stability and ultimate boundedness of fixed step size algorithms are well-documented in classic textbooks (Kushner and Yin, 2003; Borkar, 2008). Building upon these foundations, a significant line of modern research has developed finite-sample guarantees, concentration inequalities, and geometric convergence rates to a neighborhood of the solution for constant step-size stochastic approximation and RL (Chen et al., 2020; Bhandari et al., 2021; Durmus et al., 2021). More recently, there has been substantial progress on the stability analysis of stochastic approximation via refined extensions of the ODE method. For example, Borkar et al. (2025) establish asymptotic statistics, central limit theorems, and L^p ultimate boundedness for stochastic approximation

under diminishing step-sizes. Although this work does not study the constant step-size regime directly, it provides a useful ODE-based perspective that has informed subsequent analyses beyond the classical setting. In the constant step-size regime, Lauand et al. (2025) then establish geometric ergodicity and explicit L^p bounds for fixed step-sizes. Furthermore, recent algorithm-specific analyses for linear TD and Q-learning have explicitly characterized the exact $\mathcal{O}(\alpha)$ stationary bias and distributional convergence induced by constant step-sizes (Allmeier and Gast, 2024; Zhang and Xie, 2024; Huo et al., 2026). While these are important works, they primarily characterize stationary fluctuations or ultimate boundedness around a fixed point, which is fundamentally distinct from our focus on establishing exact convergence to a bounded set under diminishing and square-summable learning rates.

Convergence rates. The study of different modes of convergence of stochastic approximations is an active research area and has a rich literature. However, most existing works study the convergence to a fixed point, which can be broadly categorized into three main lines of research. The first line focuses on the asymptotic properties and almost sure convergence rates of stochastic approximations under various noise conditions. These works establish precise limiting behaviors, such as the Law of the Iterated Logarithm and almost sure rates via martingale and Lyapunov methods, often accommodating unbounded variance or Markovian noise (Chong et al., 1999; Tadić, 2002; Koval and Schwabe, 2003; Tadić, 2004; Kouritzin and Sadeghi, 2015; Vidyasagar, 2023; Karandikar and Vidyasagar, 2024). The second line establishes non-asymptotic, finite-sample guarantees for general stochastic approximations with contractive operators. This literature derives tight high-probability concentration bounds, mean-square error rates, and variance-reduced optimal bounds for both constant and decaying step-sizes (Dalal et al., 2018b; Srikant and Ying, 2019; Thoppe and Borkar, 2019; Durmus et al., 2021; Chandak et al., 2022; Mou et al., 2022; Chen et al., 2025). The third line specifically investigates the finite-time analysis and optimal sample complexity of classical RL algorithms. These studies provide fine-grained statistical analyses for TD learning and Q-learning, tightening the dependencies on the state-action space and the effective horizon under both generative models and Markovian observation settings (Szepesvári, 1997; Even-Dar and Mansour, 2003; Korda and Prashanth, 2015; Dalal et al., 2018a; Qu and Wierman, 2020; Prashanth et al., 2021; Li et al., 2024; Chandak and Borkar, 2026). Crucially, while these works provide profound insights into the convergence rates to a unique optimal solution, they are not directly applicable to the regime studied in this paper, where the iterates inherently chatter and only converge to a bounded set.

Linear Q-learning. The potential divergence of linear Q-learning has driven numerous efforts to stabilize the algorithm, typically through explicit algorithmic modifications or restrictive assumptions. Common algorithmic interventions include utilizing target networks (Mnih et al., 2015; Zhang et al., 2021b) and replay buffers (Lin, 1992), employing projection operators (Melo et al., 2008; Carvalho et al., 2020), or adding ridge regularization (Lim and Lee, 2024). When analyzing the unmodified algorithm, prior works often impose strong assumptions on the features, data sampling, or the behavior policy. For example, some analyses require orthogonal and binary features (Lee and He, 2020), or assume transitions are sampled i.i.d. from a fixed distribution (Lee and He, 2020; Carvalho et al., 2020; Lim and Lee, 2024) or from the varying stationary distribution of the current ϵ-greedy policy (Gopalan and Thoppe, 2026). Another common approach is to impose restrictive matrix definiteness or expectation bounds on the behavior policy (Melo et al., 2008; Chen et al.,

2022), which essentially require the behavior policy to be arbitrarily close to the optimal policy. Similar stringent conditions, often coupled with the restrictive Bellman completeness assumption, are also prevalent in analyses of Q-learning with neural networks (Xu and Gu, 2020; Cai et al., 2024).

Despite these successful algorithmic interventions, it is crucial to recognize the inherent nature of the unmodified algorithm. As thoroughly discussed by Meyn (2024), standard linear Q-learning is fundamentally unstable due to the lack of a GAS ODE. This fundamental instability highlights why convergence to a single fixed point is generally unattainable without the aforementioned strong assumptions or algorithmic tweaks, making the convergence to a bounded set (as studied in this work and Meyn 2024; Liu et al. 2025) a more natural and fundamental characterization of its dynamics.

Interplay of noise and step-sizes. Recent advancements have also sought to relax standard noise and step-size assumptions from complementary perspectives. For instance, a concurrent work by Nguyen et al. (2026) (who explicitly note the connection to our results in their Remark 2) investigates the almost sure convergence of SA under heavy-tailed noise possessing only a finite p-th moment ($p > 1$). From the perspective of the Robbins-Siegmund theorem, their setting essentially modifies the standard almost supermartingale bound from $\mathbb{E}_n[z_{n+1}] \leq (1 - \alpha T_n)z_n + \xi T_n^2$ to $\mathbb{E}_n[z_{n+1}] \leq (1 - \alpha T_n)z_n + \xi T_n^p$. To guarantee almost sure convergence to a unique fixed point, they naturally enforce the p-th power summability of the step-sizes (i.e., $\sum T_n^p < \infty$). This provides a fascinating orthogonal complement to our work. While Nguyen et al. (2026) adjust the step-size summability condition to counteract the $\mathcal{O}(T_n^p)$ noise term and preserve point convergence, our framework tackles the regime where the zero-order term inherently scales with T_n, yielding $\mathbb{E}_n[z_{n+1}] \leq (1 - \alpha T_n)z_n + \xi T_n$ as shown in (RS-Special). Since the Robbins-Monro condition strictly requires $\sum T_n = \infty$, the summability of the zero-order term is fundamentally broken. However, by maintaining the standard square-summability of the step-sizes ($\sum T_n^2 < \infty$), our extension uniquely establishes exact convergence to a bounded set, which is critical for inherently oscillatory algorithms like linear Q-learning.

8. Conclusion

We extend the classical Robbins-Siegmund theorem to handle the critical regime where the zero-order term has a divergent sum but is square-summable, a setting that arises naturally in many modern stochastic approximation and RL algorithms. Our extended framework provides precise almost sure convergence rates, high-probability concentration bounds, and L^p convergence rates, enabling fine-grained analysis of algorithms previously beyond theoretical reach.

As a key application, we establish the first complete convergence characterization of linear Q-learning, including almost sure convergence rates, maximal concentration bound with sub-Weibull tails, and L^p convergence rates, demonstrating that linear Q-learning converges with explicit, quantifiable rates despite having been deemed unstable for decades.

Our framework opens several promising directions: extending the analysis to nonlinear function approximation, handling non-Markovian noise models, and developing adaptive algorithms that exploit the precise convergence characterization to optimize learning rates.

More broadly, our results suggest that many algorithms previously thought to be theoretically intractable may yield to careful analysis using our extended framework.

Acknowledgments

This work is supported in part by the US National Science Foundation under the awards III-2128019, SLES-2331904, and CAREER-2442098, the Commonwealth Cyber Initiative's Central Virginia Node under the award VV-1Q26-001, VV-1Q27-007, and VV-1Q27005, 4VA at UVA, a Cisco Faculty Research Award, and a Google Research Award.

Appendix A. Proof of Section 3

This appendix proves Theorem 4.

A.1 Proof of Theorem 4

Proof Let $u_n \doteq (z_n - B_4)^+$. Again we consider two cases. **(1) When $u_n > 0$.** Then $u_n = z_n - B_4$, Theorem 4 (A3) gives $x_n - y_n \le -c_n u_n$. Similar to (2), we have $u_{n+1}^2 \le (z_{n+1} - z_n + u_n)^2$. Thus

$$\begin{aligned}
&\mathbb{E}_n\big[u_{n+1}^2\big] \\
\le&\mathbb{E}_n\big[u_n^2 + 2u_n(z_{n+1} - z_n) + (z_{n+1} - z_n)^2\big] \\
\le&u_n^2 + 2u_n\mathbb{E}_n[z_{n+1} - z_n] + \mathbb{E}_n\big[b_n^2(z_n + 1)^2\big] && \text{(By Theorem 4 (A2))} \\
\le&u_n^2 + 2u_n(a_n z_n + x_n - y_n) + b_n^2(u_n + 1 + B_4)^2 && \text{(By (RS))} \\
\le&u_n^2 + 2u_n(a_n(u_n + B_4) - c_n u_n) + 2b_n^2 u_n^2 + 2(1 + B_4)^2 b_n^2 \\
& && \text{(Since } u_n > 0 \text{ and } x_n - y_n \le -c_n u_n\text{)} \\
=&u_n^2 + 2a_n u_n^2 + 2a_n B_4 u_n - 2c_n u_n^2 + 2b_n^2 u_n^2 + 2(1 + B_4)^2 b_n^2 \\
\le&u_n^2 + 2a_n u_n^2 + a_n B_4(u_n^2 + 1) - 2c_n u_n^2 + 2b_n^2 u_n^2 + 2(1 + B_4)^2 b_n^2 \\
=&u_n^2 + (2 + B_4)a_n u_n^2 + a_n B_4 - 2c_n u_n^2 + 2b_n^2 u_n^2 + 2(1 + B_4)^2 b_n^2 \\
=&(1 + 2b_n^2 + (2 + B_4)a_n)u_n^2 + 2(1 + B_4)^2 b_n^2 + a_n B_4 - 2c_n u_n^2.
\end{aligned}$$

(2) When $u_n = 0$. We have $z_n \le B_4$. By (1), we obtain

$$|u_{n+1}| = |u_{n+1} - u_n| \le |(z_{n+1} - B_4) - (z_n - B_4)| = |z_{n+1} - z_n|.$$

Thus we have

$$\begin{aligned}
\mathbb{E}_n\big[u_{n+1}^2\big] \le&\mathbb{E}_n\big[(z_{n+1} - z_n)^2\big] \\
\le&b_n^2(z_n + 1)^2 && \text{(By Theorem 4 (A2))} \\
\le&b_n^2(B_4 + 1)^2 && \text{(Since } z_n \in [0, B_4]\text{)} \\
\le&(1 + 2b_n^2 + (2 + B_4)a_n)u_n^2 + 2(1 + B_4)^2 b_n^2 + a_n B_4 - 2c_n u_n^2. && \text{(Since } u_n = 0\text{)}
\end{aligned}$$

Combining the two cases, there exist non-negative sequences

$$d_n \doteq 2b_n^2 + (2 + B_4)a_n, e_n \doteq 2(1 + B_4)^2 b_n^2 + B_4 a_n,$$

such that

$$\mathbb{E}_n\big[u_{n+1}^2\big] \le (1+d_n)u_n^2 + e_n - 2c_n u_n^2.$$

Since $\sum_n d_n < \infty$ and $\sum_n e_n < \infty$ a.s., the Robbins–Siegmund theorem (Lemma 1) yields that u_n^2 converges almost surely to a finite limit. Hence $(z_n - B_4)^+$ has an a.s. finite limit. From Lemma 1, we also derive $\sum_n 2c_n u_n^2 < \infty$. Given $\sum_n c_n = \infty$ a.s., following the same argument in Theorem 3 then completes the proof. ■

Appendix B. Proof of Section 4

This appendix provides the proofs of the nonasymptotic results in Section 4.

B.1 Proof of Theorem 5

Proof Recall we derived the following bound in (4)

$$\mathbb{E}_n\big[u_{n+1}^2\big] \le \big(1 - 2\alpha T_n + 2B_{3,1}^2 T_n^2\big)u_n^2 + 2B_{3,1}^2\Big(1+\tfrac{\xi}{\alpha}\Big)^2 T_n^2,$$

where $u_n = \Big(z_n - \frac{\xi}{\alpha}\Big)^+$. We apply Lemma 2 with the identification

$$z_n \leftrightarrow u_n^2, a_n \leftrightarrow 2B_{3,1}^2 T_n^2, x_n \leftrightarrow 2B_{3,1}^2\Big(1+\tfrac{\xi}{\alpha}\Big)^2 T_n^2, b_n \leftrightarrow 2\alpha T_n.$$

We verify the assumptions (A1)–(A3) of Lemma 2.

(A1) of Lemma 2. Theorem 5 (A1) gives $\sum_n T_n = \infty$, thus $\sum_n b_n = \infty$; Theorem 5 (A2) gives $\sum_n T_n^2 < \infty$, thus $\sum_n a_n < \infty$ and $\sum_n x_n < \infty$. These trivially confirm that Lemma 2 (A1) holds.

(A2) of Lemma 2. From Theorem 5 (A1), there exists some n_1 such that for all $n \ge n_1$, we then have $nT_n > \frac{\eta}{2\alpha}$. That is, $b_n = 2\alpha T_n \ge \frac{\eta}{n}$ for $n \ge n_1$.

(A3) of Lemma 2. The first part is satisfied by Theorem 5 (A2), we now validate the second part. Theorem 5 (A1) implies there exist some constant $\eta_0 > 0$, such that $\frac{\eta_0}{2\alpha} \le \liminf_n nT_n - \frac{\eta}{2\alpha}$. Thus there exists some n_1 such that for all $n \ge n_1$, we have $\eta_0 \le 2\alpha n T_n - \eta$. That is,

$$\begin{aligned}\textstyle\sum_{n=1}^{\infty}\big(b_n - \tfrac{\eta}{n}\big) &= \textstyle\sum_{n=1}^{\infty}\Big(\frac{n\cdot 2\alpha T_n - \eta}{n}\Big)\\ &= \textstyle\sum_{n=1}^{n_1}\big(2\alpha T_n - \frac{\eta}{n}\big) + \sum_{n=n_1+1}^{\infty}\Big(\frac{n\cdot 2\alpha T_n - \eta}{n}\Big)\\ &\ge \textstyle\sum_{n=1}^{n_1}\big(2\alpha T_n - \frac{\eta}{n}\big) + \sum_{n=n_1+1}^{\infty}\frac{\eta_0}{n}.\end{aligned}$$

Here the first summation is finite, while the second summation converges to ∞. Thus (A3) of Lemma 2 holds.

By Lemma 2, we obtain $\lim_{n\to\infty} n^\eta u_n^2 = 0$ a.s., which implies

$$\lim_{n\to\infty} n^\eta u_n^2 = \lim_{n\to\infty} n^\eta((z_n - B_3)^+)^2 = \lim_{n\to\infty} n^\eta d(z_n,[0,B_3])^2 = 0 \quad \text{a.s.}$$

Taking square roots then completes the proof. ■

B.2 Proof of Theorem 6

Proof To establish the concentration bound, we adopt the inductive (bootstrapping / recursive) framework from Chen et al. (2025). Our argument begins with the establishment of the inductive basis.

Lemma 15 *There exist some deterministic integers C_{15} and C'_{15} such that for any $n \geq 0$,*

$$\left((z_n - B_3)^+\right)^2 \leq C'_{15}(n+1)^{C_{15}} \quad a.s.$$

The proof is in Appendix B.2.1.

Lemma 16 *There exists some deterministic constant $\bar{n}$ such that the following implication relation holds. If for any $\delta \in (0,1)$, there exists some non-decreasing sequence $\{B_n(\delta)\}_{n=\bar{n}}^{\infty}$ such that*

$$\Pr\left(\forall n \geq \bar{n}, \left((z_n - B_3)^+\right)^2 \leq B_n(\delta)\right) \geq 1-\delta, \tag{7}$$

then for any $\delta' \in (0, 1-\delta)$, there exists a new sequence $\{B_n(\delta,\delta')\}_{n=\bar{n}}^{\infty}$ such that

$$\Pr\left(\forall n \geq \bar{n}, \left((z_n - B_3)^+\right)^2 \leq B_n(\delta,\delta')\right) \geq 1-\delta-\delta',$$

where $B_n(\delta,\delta') = C_{16}T_n \max\{B_n(\delta), 1\}\left(1 + \ln(1/\delta') + \ln(n+n_0)\right)$ with C_{16} being some constant depending on $\bar{n}$ and independent of n, δ, δ'.

Similar to our previous analysis, we need to consider two scenarios: $u_n > 0$ and $u_n = 0$. The proof is in Appendix B.2.2. In essence, each application of Lemma 16 enhances our bound by a factor of T_n. Given that Lemma 15 establishes an almost sure bound of order $n^{C_{15}}$, we can achieve the desired bound by applying Lemma 16 exactly $C_{15}+1$ times. Further applications would be counterproductive, as they would violate the non-decreasing property of the bound B_n.

Lemma 17 *Under Assumption 4.1, there exists some constants $C_{17} = C_{15}+1$ and C'_{17} such that for any $\delta > 0$, it holds, with probability at least $1-\delta$, that for all $n \geq 0$,*

$$\left((z_n - B_3)^+\right)^2 \leq C'_{17}\tfrac{1}{n+n_0}\left[\ln\left(\tfrac{C_{17}}{\delta}\right) + 1 + \ln(n+n_0)\right]^{C_{17}}.$$

The proof follows the same methodology as Qian et al. (2026). The details are in Appendix B.2.3.

■

B.2.1 Proof of Lemma 15

Proof For any $n \geq 0$, we have

$$|u_{n+1} - u_n| \leq \left|\left(z_{n+1} - \tfrac{\xi}{\alpha}\right) - \left(z_n - \tfrac{\xi}{\alpha}\right)\right| \qquad \text{(By (1))}$$

$$
\begin{aligned}
&= |z_{n+1} - z_n| \\
&\le B_{3,1} T_n (z_n + 1) && \text{(By Theorem 3 (A2))} \\
&\le B_{3,1} T_n \Big(u_n + \tfrac{\xi}{\alpha} + 1 \Big). && \text{(By definition of } u_n)
\end{aligned}
$$

Hence

$$u_{n+1} \le u_n + B_{3,1} T_n \Big(u_n + \tfrac{\xi}{\alpha} + 1 \Big) \le u_0 + B_{3,1} \textstyle\sum_{i=0}^n T_i u_i + 2 B_{3,1} \Big(1 + \tfrac{\xi}{\alpha} \Big) \sum_{i=0}^n T_i.$$

We first recall a discrete Gronwall inequality.

Lemma 18 (Discrete Gronwall Inequality) *(Lemma 8 in Section 11.2 of Borkar 2008) For non-negative real sequences $\{x_n\}_{n\ge 0}$ and $\{a_n\}_{n\ge 0}$ and scalars $C, L \ge 0$, it holds*

$$x_{n+1} \le C + L \textstyle\sum_{i=0}^n a_i x_i \quad \forall n \implies x_{n+1} \le (C + x_0) \exp(L \sum_{i=0}^n a_i) \quad \forall n.$$

Applying the above result, we obtain that for all $n \ge 0$,

$$u_n \le \Big(u_0 + 2 B_{3,1} \Big(1 + \tfrac{\xi}{\alpha} \Big) \textstyle\sum_{i=0}^n T_i \Big) \exp \Big(1 + B_{3,1} \sum_{i=0}^{n-1} T_i \Big).$$

Noting $T_n = \frac{C_{4.1}}{n+n_0}$ and $\sum_{i=0}^n T_i \le C_{4.1} \ln(n + n_0 + 1)$ then completes the proof. ■

B.2.2 Proof of Lemma 16

Proof For any $\delta \in (0,1)$, consider the non-decreasing sequence $\{B_n(\delta)\}$ defined in (7). For each $n \ge \bar{n}$, where $\bar{n}$ is a sufficiently large constant, define an event

$$E_n(\delta) = \big\{ u_k^2 \le B_k(\delta),\, \forall k = \bar{n}, \bar{n}+1, \ldots, n \big\}.$$

Notably, $\{E_n(\delta)\}_{n \ge \bar{n}}$ is by definition a sequence of decreasing events, i.e., $E_{n+1}(\delta) \subset E_n(\delta)$. In addition, according to (7), we have $\Pr(E_n(\delta)) \ge 1 - \delta$ for any $n \ge \bar{n}$. Let $\lambda_n = \theta T_n^{-1} (\max\{B_n(\delta), 1\})^{-1}$ where θ is a tunable constant yet to be chosen.

We now proceed to show that $\big\{ \exp\big(\lambda_n \mathbb{1}_{E_n(\delta)} u_n^2 \big) \big\}$ is "almost" a supermartingale.

(1) When $u_n > 0$. Then $u_n = z_n - \frac{\xi}{\alpha}$. By (2), we have $u_{n+1}^2 \le u_n^2 + (z_{n+1} - z_n)^2 + 2 u_n (z_{n+1} - z_n)$. By first multiplying $\lambda_{n+1} \mathbb{1}_{E_{n+1}(\delta)}$ and then taking exponential on both sides of the inequality,

$$
\begin{aligned}
&\exp\big(\lambda_{n+1} \mathbb{1}_{\mathbb{E}_{n+1}(\delta)} u_{n+1}^2 \big) \\
\le\, &\exp\big(\lambda_{n+1} \mathbb{1}_{\mathbb{E}_{n+1}(\delta)} u_n^2 \big) \times \exp\big(\lambda_{n+1} \mathbb{1}_{\mathbb{E}_{n+1}(\delta)} (z_{n+1} - z_n)^2 \big) \\
&\times \exp\big(\lambda_{n+1} \mathbb{1}_{\mathbb{E}_{n+1}(\delta)} 2 u_n (z_{n+1} - z_n) \big) \\
\le\, &\exp\big(\lambda_{n+1} \mathbb{1}_{\mathbb{E}_{n}(\delta)} u_n^2 \big) \times \exp\big(\lambda_{n+1} \mathbb{1}_{\mathbb{E}_{n}(\delta)} (z_{n+1} - z_n)^2 \big) \\
&\times \exp\big(\lambda_{n+1} \mathbb{1}_{\mathbb{E}_{n}(\delta)} 2 u_n (z_{n+1} - z_n) \big). && \text{(Since } E_{n+1}(\delta) \subset E_n(\delta))
\end{aligned}
$$

Taking expectation conditioned on $\mathcal{F}_n$ on both sides of the previous inequality, we obtain

$$\mathbb{E}_n \big[\exp\big(\lambda_{n+1} \mathbb{1}_{E_{n+1}(\delta)} u_{n+1}^2 \big) \big]$$

$$\leq \exp\big(\lambda_{n+1}\mathbb{1}_{E_n(\delta)}u_n^2\big) \times \exp\bigg(\lambda_{n+1}\mathbb{1}_{E_n(\delta)}2B_{3,1}^2T_n^2\bigg(u_n^2+\Big(1+\tfrac{\xi}{\alpha}\Big)^2\bigg)\bigg)$$
$$\times \mathbb{E}_n\big[\exp\big(\lambda_{n+1}\mathbb{1}_{E_n(\delta)}2u_n(z_{n+1}-z_n)\big)\big], \tag{8}$$

where the second term is obtained by

$$\begin{aligned}(z_{n+1}-z_n)^2 \leq& B_{3,1}^2T_n^2(z_n+1)^2 && \text{(By Theorem 3 (A2))}\\ =& B_{3,1}^2T_n^2\Big(\Big(u_n+\tfrac{\xi}{\alpha}\Big)+1\Big)^2 && \text{(Since } u_n = z_n - \tfrac{\xi}{\alpha}\text{)}\\ \leq& 2B_{3,1}^2T_n^2\bigg(u_n^2+\Big(1+\tfrac{\xi}{\alpha}\Big)^2\bigg).\end{aligned}$$

To bound the last term, we will use Hoeffding's lemma. We have

$$\begin{aligned}\mathbb{E}_n\big[\lambda_{n+1}\mathbb{1}_{E_n(\delta)}2u_n(z_{n+1}-z_n)\big] =& 2\lambda_{n+1}\mathbb{1}_{E_n(\delta)}u_n\mathbb{E}_n[(z_{n+1}-z_n)]\\ \leq& 2\lambda_{n+1}\mathbb{1}_{E_n(\delta)}u_n(-\alpha T_n z_n+\xi T_n). && \text{(By (RS-Special))}\\ =& -2\lambda_{n+1}\mathbb{1}_{E_n(\delta)}\alpha T_n u_n^2. && \text{(Since } u_n = z_n - \tfrac{\xi}{\alpha}\text{)}\end{aligned}$$

Additionally, by (A2) of Theorem 3, we obtain

$$\begin{aligned}&\big|\lambda_{n+1}\mathbb{1}_{E_n(\delta)}2u_n(z_{n+1}-z_n)\big|\\ \leq& 2\lambda_{n+1}\mathbb{1}_{E_n(\delta)}B_{3,1}T_nu_n\Big(u_n+\tfrac{\xi}{\alpha}+1\Big)\\ \leq& 2\lambda_{n+1}\mathbb{1}_{E_n(\delta)}B_{3,1}T_nu_n\Big(B_n(\delta)^{1/2}+1+\tfrac{\xi}{\alpha}\Big). && \text{(Definition of } E_n(\delta)\text{)}\end{aligned}$$

Now we can apply Hoeffding's lemma to the last term in (8), which gives us

$$\begin{aligned}&\mathbb{E}_n\left[\exp\big(\lambda_{n+1}\mathbb{1}_{E_n(\delta)}2u_n(z_{n+1}-z_n)\big)\right]\\ \leq& \exp\bigg(-2\lambda_{n+1}\mathbb{1}_{E_n(\delta)}\alpha T_nu_n^2+32\lambda_{n+1}^2\mathbb{1}_{E_n(\delta)}u_n^2B_{3,1}^2T_n^2\Big(B_n(\delta)^{1/2}+1+\tfrac{\xi}{\alpha}\Big)^2/8\bigg)\\ \leq& \exp\bigg(-2\lambda_{n+1}\mathbb{1}_{E_n(\delta)}\alpha T_nu_n^2+8\lambda_{n+1}^2\mathbb{1}_{E_n(\delta)}u_n^2B_{3,1}^2T_n^2\Big(B_n(\delta)+\Big(1+\tfrac{\xi}{\alpha}\Big)^2\Big)\bigg).\end{aligned}$$

Applying this back to (8), we have

$$\begin{aligned}&\mathbb{E}_n\big[\exp\big(\lambda_{n+1}\mathbb{1}_{E_{n+1}(\delta)}u_{n+1}^2\big)\big]\\ \leq& \exp\bigg(\lambda_{n+1}\mathbb{1}_{E_n(\delta)}u_n^2\Big(1-2\alpha T_n+2B_{3,1}^2T_n^2+8\lambda_{n+1}B_{3,1}^2T_n^2\Big(B_n(\delta)+\Big(1+\tfrac{\xi}{\alpha}\Big)^2\Big)\Big)\bigg)\\ &\times\exp\bigg(\lambda_{n+1}\mathbb{1}_{E_n(\delta)}2B_{3,1}^2\Big(1+\tfrac{\xi}{\alpha}\Big)^2T_n^2\bigg).\end{aligned}$$

(2) When $u_n = 0$. We obtain

$$\begin{aligned}&\mathbb{E}_n\big[\exp\big(\lambda_{n+1}\mathbb{1}_{E_{n+1}(\delta)}u_{n+1}^2\big)\big]\\ \leq& \mathbb{E}_n\big[\exp\big(\lambda_{n+1}\mathbb{1}_{E_n(\delta)}u_{n+1}^2\big)\big] && \text{(Since } E_{n+1}(\delta)\subset E_n(\delta)\text{)}\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}_n\big[\exp\big(\lambda_{n+1}\mathbb{1}_{E_n(\delta)}(z_{n+1}-z_n)^2\big)\big] && \text{(By (3))}\\
&\le \mathbb{E}_n\left[\exp\left(\lambda_{n+1}\mathbb{1}_{E_n(\delta)}B_{3,1}^2\Big(u_n+1+\tfrac{\xi}{\alpha}\Big)^2T_n^2\right)\right]\\
&\qquad\qquad \text{(By Theorem 3 (A2) and since } z_n\in[0,\tfrac{\xi}{\alpha}]\text{)}\\
&= \exp\left(\lambda_{n+1}\mathbb{1}_{E_n(\delta)}B_{3,1}^2\Big(1+\tfrac{\xi}{\alpha}\Big)^2T_n^2\right). && \text{(Since } u_n=0\text{)}
\end{aligned}
$$

To combine two cases and simplify notation, define

$$
\begin{aligned}
D_{1,n} &\doteq \tfrac{\lambda_{n+1}}{\lambda_n}\left(1-2\alpha T_n+2B_{3,1}^2T_n^2+8\lambda_{n+1}B_{3,1}^2T_n^2\left(B_n(\delta)+\Big(1+\tfrac{\xi}{\alpha}\Big)^2\right)\right),\\
D_{2,n} &\doteq \lambda_{n+1}2B_{3,1}^2\Big(1+\tfrac{\xi}{\alpha}\Big)^2T_n^2.
\end{aligned}
$$

Then, the previous inequalities (for both cases where $u_n>0$ or $u_n=0$) can be written in a unified general form as follows

$$
\mathbb{E}_n\big[\exp\big(\lambda_{n+1}\mathbb{1}_{E_{n+1}(\delta)}u_{n+1}^2\big)\big]\le\exp\big((D_{1,n})^+\lambda_n\mathbb{1}_{E_n(\delta)}u_n^2\big)\exp\big(D_{2,n}\mathbb{1}_{E_n(\delta)}\big). \tag{9}
$$

Next, we bound the terms $D_{1,n}$ and $D_{2,n}$. Since $\lambda_n=\theta T_n^{-1}\max\{B_n(\delta),1\}^{-1}$, denote $C_{16,1}\doteq 1+\frac{\xi}{\alpha}$, then for n sufficiently large we have

$$
\begin{aligned}
D_{1,n}\mathbb{1}_{E_n(\delta)} &= \tfrac{\lambda_{n+1}}{\lambda_n}(1-2\alpha T_n+2B_{3,1}^2C_{16,1}^2T_n^2+8\lambda_{n+1}B_{3,1}^2T_n^2(B_n(\delta)+C_{16,1}^2))\mathbb{1}_{E_n(\delta)}\\
&\le \tfrac{\lambda_{n+1}}{\lambda_n}(1-\alpha T_n+8\lambda_{n+1}B_{3,1}^2T_n^2(B_n(\delta)+C_{16,1}^2))\mathbb{1}_{E_n(\delta)}\\
&= \tfrac{T_n\max\{B_n(\delta),1\}}{T_{n+1}\max\{B_{n+1}(\delta),1\}}\left(1-\alpha T_n+\tfrac{8\theta B_{3,1}^2T_n^2(B_n(\delta)+C_{16,1}^2)}{T_{n+1}\max\{B_{n+1}(\delta),1\}}\right)\mathbb{1}_{E_n(\delta)}\\
&\le \tfrac{T_n}{T_{n+1}}\left(1-\alpha T_n+\tfrac{8B_{3,1}^2\theta T_n^2}{T_{n+1}}\tfrac{C_{16,1}^2+B_n(\delta)+1}{\max\{B_{n+1}(\delta),1\}}\right)\mathbb{1}_{E_n(\delta)}\\
&\le \tfrac{T_n}{T_{n+1}}\left(1-\alpha T_n+\tfrac{8B_{3,1}^2\theta T_n^2}{T_{n+1}}(C_{16,1}^2+2)\right)\\
&\le \tfrac{T_n}{T_{n+1}}\left(1-\alpha T_n+16B_{3,1}^2\theta T_n(C_{16,1}^2+2)\right).
\end{aligned}
$$

Recall $C_{4.1}>1/\alpha$ (Assumption 4.1). Select any $\kappa\in(1/C_{4.1},\alpha)$ and define $\theta\doteq\frac{1}{C_{16,1}^2+2}\min\{1,\frac{\alpha-\kappa}{16B_{3,1}^2}\}$. Then we have

$$
D_{1,n}\mathbb{1}_{E_n(\delta)}\le\tfrac{T_n}{T_{n+1}}(1-T_n\kappa)=\tfrac{n+n_0+1}{n+n_0}\left(1-\tfrac{C_{4.1}}{n+n_0}\kappa\right).
$$

Since $C_{4.1}\kappa>1$, it is easy to compute that $D_{1,n}\mathbb{1}_{E_n(\delta)}<1$ holds for all n. Now, consider $D_{2,n}$. We have by definition of λ_n that

$$
D_{2,n}\mathbb{1}_{E_n(\delta)}=\tfrac{2\theta B_{3,1}^2C_{16,1}^2T_n^2}{T_{n+1}\max\{B_{n+1}(\delta),1\}}\mathbb{1}_{E_n(\delta)}\le\tfrac{2\theta B_{3,1}^2C_{16,1}^2T_n^2}{T_{n+1}}\le 4B_{3,1}^2C_{16,1}^2T_n.
$$

Using the upper bounds we obtained for the terms $D_{1,n}$ and $D_{2,n}$ in (9), we have for any $n\ge\bar{n}$,

$$
\mathbb{E}_n\big[\exp\big(\lambda_{n+1}\mathbb{1}_{E_{n+1}(\delta)}u_{n+1}^2\big)\big]
$$

$$\begin{aligned}&\leq \exp\left((D_{1,n})^+\lambda_n \mathbb{1}_{E_n(\delta)} u_n^2\right) \exp\left(D_{2,n}\mathbb{1}_{E_n(\delta)}\right)\\ &\leq \exp\left(\lambda_n \mathbb{1}_{E_n(\delta)} u_n^2\right) \exp\left(4B_{3,1}^2 C_{16,1}^2 T_n\right).\end{aligned}$$

For $n \geq \bar{n}$, define $Z_n \doteq \exp\Big(\lambda_n \mathbb{1}_{E_n(\delta)} u_n^2 - 4B_{3,1}^2 C_{16,1}^2 \sum_{i=\bar{n}}^{n-1} T_i\Big)$. We next show that $\{Z_n\}$ is a supermartingale with respect to the filtration $\{\mathcal{F}_n\}$. Indeed, it holds that

$$\begin{aligned}\mathbb{E}_n[Z_{n+1}] =&\mathbb{E}_n\left[\exp\left(\lambda_{n+1}\mathbb{1}_{E_{n+1}(\delta)} u_{n+1}^2\right)\right] \exp\left(-4B_{3,1}^2 C_{16,1}^2 \textstyle\sum_{i=\bar{n}}^{n} T_i\right)\\ &\leq \exp\left(\lambda_n \mathbb{1}_{E_n(\delta)} u_n^2\right) \exp\left(4B_{3,1}^2 C_{16,1}^2 T_n\right) \exp\left(-4B_{3,1}^2 C_{16,1}^2 \textstyle\sum_{i=\bar{n}}^{n} T_i\right)\\ &= Z_n.\end{aligned}$$

For any $\delta' \in (0, 1-\delta)$, we now construct $B(\delta, \delta')$. By Ville's maximal inequality, we get for any ϵ,

$$\begin{aligned}&\Pr\Big(\sup_{n\geq\bar{n}} \exp\Big(\lambda_n \mathbb{1}_{E_n(\delta)} u_n^2 - 4B_{3,1}^2 C_{16,1}^2 \textstyle\sum_{i=\bar{n}}^{n-1} T_i\Big) \geq \exp(\epsilon)\Big)\\ \leq&\mathbb{E}\left[\exp\left(\lambda_{\bar{n}} \mathbb{1}_{E_{\bar{n}}(\delta)} u_{\bar{n}}^2 - \epsilon\right)\right]\\ \leq&\mathbb{E}\left[\exp\left(\theta T_{\bar{n}}^{-1} B_{\bar{n}}(\delta)^{-1} \mathbb{1}_{E_{\bar{n}}(\delta)} u_{\bar{n}}^2 - \epsilon\right)\right]\\ \leq&\mathbb{E}\left[\exp\left(\theta T_{\bar{n}}^{-1} - \epsilon\right)\right] = \exp\left(\theta T_{\bar{n}}^{-1} - \epsilon\right).\end{aligned}$$

Select ϵ such that $\exp\left(\theta T_{\bar{n}}^{-1} - \epsilon\right) = \delta'$, then it holds, with probability at least $1-\delta'$, that for all $n \geq \bar{n}$,

$$\lambda_n \mathbb{1}_{E_n(\delta)} u_n^2 - 4B_{3,1}^2 C_{16,1}^2 \textstyle\sum_{i=\bar{n}}^{n-1} T_i \leq \theta T_{\bar{n}}^{-1} - \ln \delta',$$

implying

$$\lambda_n \mathbb{1}_{E_n(\delta)} u_n^2 \leq \theta T_{\bar{n}}^{-1} + \ln(1/\delta') + 4B_{3,1}^2 C_{16,1}^2 C_{4.1} \ln(n + n_0).$$

Next, we complete Lemma 16 by removing $\mathbb{1}_{E_n(\delta)}$ in the LHS of the above inequality. For simplicity of notation, we denote the RHS of the above inequality as $\epsilon(n, \bar{n}, \delta')$. We then have

$$\begin{aligned}&\Pr\left(\lambda_n u_n^2 \leq \epsilon(n, \bar{n}, \delta'), \forall n \geq \bar{n}\right)\\ =&\Pr\left(\textstyle\bigcap_{n=\bar{n}}^{\infty} \{\lambda_n u_n^2 \leq \epsilon(n, \bar{n}, \delta')\}\right)\\ \geq&\Pr\left(\textstyle\bigcap_{n=\bar{n}}^{\infty} \{\lambda_n \mathbb{1}_{E_n(\delta)} u_n^2 \leq \epsilon(n, \bar{n}, \delta')\} \cap E_n(\delta)\right).\end{aligned}$$

To proceed, note that for any two events, A and B, we have

$$\Pr(A \cap B) = 1 - \Pr(A^c \cup B^c) \geq 1 - \Pr(A^c) - \Pr(B^c) = \Pr(A) + \Pr(B) - 1.$$

Therefore, we have

$$\begin{aligned}&\Pr\left(\lambda_n u_n^2 \leq \epsilon(n, \bar{n}, \delta'), \forall n \geq \bar{n}\right)\\ \geq&\Pr\left(\textstyle\bigcap_{n=\bar{n}}^{\infty} \{\lambda_n \mathbb{1}_{E_n(\delta)} u_n^2 \leq \epsilon(n, \bar{n}, \delta')\} \cap E_n(\delta)\right)\\ \geq&\Pr\left(\textstyle\bigcap_{n=\bar{n}}^{\infty} \{\lambda_n \mathbb{1}_{E_n(\delta)} u_n^2 \leq \epsilon(n, \bar{n}, \delta')\}\right) + \Pr\left(\textstyle\bigcap_{n=0}^{\infty} E_n(\delta)\right) - 1\end{aligned}$$

$$
\begin{aligned}
&= \Pr\big(\lambda_n \mathbb{1}_{E_n(\delta)} u_n^2 \le \epsilon(n, \bar{n}, \delta'), \forall n \ge \bar{n}\big) + \lim_{n\to\infty} \Pr(E_n(\delta)) - 1 \\
&\ge (1-\delta') + (1-\delta) - 1 \\
&= 1 - \delta - \delta'.
\end{aligned}
$$

Using the definitions of $\epsilon(n, \bar{n}, \delta')$ and λ_n, we arrive at the following result. For any $\delta' \in (0, 1-\delta)$, it holds, with probability at least $1-\delta-\delta'$, that for any $n \ge \bar{n}$

$$
u_n^2 \le \frac{T_n \max\{B_n(\delta), 1\}}{\theta} \big(\theta T_{\bar{n}}^{-1} + \ln(1/\delta') + 4B_{3,1}^2 C_{16,1} C_{4.1} \ln(n+n_0)\big).
$$

Defining C_{16} accordingly then completes the proof. ■

B.2.3 Proof of Lemma 17

Proof For any $\delta \in (0,1)$, we define $B_n(\delta) = C'_{15}(n+1)^{C_{15}} \ge 1$. From Lemma 15, we conclude that

$$
\Pr\big(u_n^2 \le B_n(\delta), \forall n \ge \bar{n}\big) = 1 \ge 1 - \delta.
$$

This lays out the antecedent term in the implication relationship specified in Lemma 16. We define $C_{17} \doteq C_{15} + 1$. Let $\delta_1, \delta_2, \cdots, \delta_{C_{17}} > 0$ be such that $\sum_{i=1}^{C_{17}} \delta_i \le 1$. Repeatedly applying Lemma 16 for C_{17} times, we have that it holds, with probability at least $1 - \sum_{i=1}^{C_{17}} \delta_i$, that for any $n \ge \bar{n}$,

$$
u_n^2 \le T_n^{C_{17}} B_n(\delta) \textstyle\prod_{i=1}^{C_{17}} C_{16} \left[\ln(1/\delta_i) + 1 + \ln(n+n_0)\right].
$$

By choosing $\delta_i = \delta / C_{17}$ for all $i \in \{1, 2, \cdots, C_{17}\}$, the previous inequality reads

$$
\begin{aligned}
u_n^2 &\le C_{16}^{C_{17}} T_n^{C_{17}} B_n(\delta) \left[\ln(C_{17}/\delta) + 1 + \ln(n+n_0)\right]^{C_{17}} \\
&\le C_{16}^{C_{17}} C_{4.1}^{C_{17}} C'_{15} (n+n_0)^{-1} \left[\ln(C_{17}/\delta) + 1 + \ln(n+n_0)\right]^{C_{17}}.
\end{aligned}
$$

Notably, the above inequality holds only for $n \ge \bar{n}$. For $n < \bar{n}$, Lemma 15 gives a trivial almost sure bound that $u_n^2 \le C'_{15}(\bar{n}+1)^{C_{15}}$. To get a bound for all n, we select a constant $C_{17,1}$ such that

$$
C_{17,1} \min_{n=0,\ldots,\bar{n}} C_{16}^{C_{17}} C_{4.1}^{C_{17}} C'_{15} (n+n_0)^{-1} \left[\ln C_{17} + 1 + \ln(n+n_0)\right]^{C_{17}} > 1.
$$

Since $\ln(C_{17}/\delta) \ge \ln C_{17}$, we have

$$
C_{17,1} \min_{n=0,\ldots,\bar{n}} C_{16}^{C_{17}} C_{4.1}^{C_{17}} C'_{15} (n+n_0)^{-1} \left[\ln(C_{17}/\delta) + 1 + \ln(n+n_0)\right]^{C_{17}} > 1.
$$

It then holds that for any $n \ge 0$,

$$
u_n^2 \le C'_{15}(\bar{n}+1)^{C_{15}} C_{17,1} C_{16}^{C_{17}} C_{4.1}^{C_{17}} C'_{15} (n+n_0)^{-1} \left[\ln(C_{17}/\delta) + 1 + \ln(n+n_0)\right]^{C_{17}}.
$$

Defining C'_{17} accordingly then completes the proof. ■

B.3 Proof of Theorem 8

Proof Recall we have derived the following bound in Theorem 4

$$\mathbb{E}_n\left[u_{n+1}^2\right] \le (1+d_n)u_n^2 + e_n - 2c_n u_n^2,$$

where $d_n = 2b_n^2 + (2+B_4)a_n, e_n = 2(1+B_4)^2 b_n^2 + B_4 a_n$. We proceed via applying Lemma 2. In particular, we identify

$$z_n \leftrightarrow u_n^2, a_n^* \leftrightarrow 2b_n^2 + (2+B_4)a_n, x_n \leftrightarrow 2(1+B_4)^2 b_n^2 + B_4 a_n, b_n^* \leftrightarrow 2c_n.$$

(A1) of Lemma 2. Theorem 8 (A2) gives $\sum_n (a_n + b_n^2) < \infty$ a.s., thus $\sum_n a_n < \infty$ and $\sum_n b_n^2 < \infty$ a.s.; Theorem 8 (A1) gives $\liminf_{n\to\infty} nc_n > \frac{\eta}{2}$, thus $\sum_n c_n = \infty$ a.s., these trivially confirm that Lemma 2 (A1) holds.

(A2) of Lemma 2. We have $\liminf_{n\to\infty} nc_n > \frac{\eta}{2}$. So when n is sufficiently large, it holds that $2c_n n \ge \eta$.

(A3) of Lemma 2. The first part is satisfied by Theorem 8 (A2), we now validate the second part. By Theorem 8 (A1), there exist some constants n_1 and $\eta_0 > 0$, such that for all $n \ge n_1$, we have $\frac{\eta_0}{2} \le nc_n - \frac{\eta}{2}$. That is,

$$\begin{aligned}
&\textstyle\sum_{n=1}^{\infty}\left(b_n^* - \frac{\eta}{n}\right)\\
=&\textstyle\sum_{n=1}^{\infty}\left(\frac{n\cdot 2c_n - \eta}{n}\right)\\
=&\textstyle\sum_{n=1}^{n_1}\left(2c_n - \frac{\eta}{n}\right) + \sum_{n=n_1+1}^{\infty}\left(\frac{n\cdot 2c_n - \eta}{n}\right)\\
\ge&\textstyle\sum_{n=1}^{n_1}\left(2c_n - \frac{\eta}{n}\right) + \sum_{n=n_1+1}^{\infty}\frac{\eta_0}{n}.
\end{aligned}$$

Here the first summation is finite, while the second summation converges to ∞. Thus Lemma 2 (A3) holds.

Having verified all conditions of Lemma 2 for this setup, invoking Lemma 2 yields

$$\lim_{n\to\infty} n^{\eta} u_n^2 = \lim_{n\to\infty} n^{\eta}((z_n - B_4)^+)^2 = 0 \quad \text{a.s.}$$

Taking square roots then completes the proof. ■

Appendix C. Proof of Section 5

We begin with a shared part that provides the common groundwork. The proofs of Theorem 9 and Theorem 10 are then completed separately in Sections C.1 and C.2, respectively.

To proceed, we construct a secondary time scale adapted to $\{\alpha_t\}$. Define

$$T_m = \frac{C_\alpha \ln^{\nu_1}(m+3)}{(m+3)^{\nu_2}}, \tag{10}$$

where (ν_1, ν_2) are chosen according to the learning-rate regime

$$\begin{cases} 0 < \nu_1 < 1, \nu_2 = 1 & \text{for Assumption LR1 with } \nu = 1,\\ \nu_1 = 0, \frac{1}{2} < \nu_2 < \frac{\nu}{2-\nu} & \text{for Assumption LR1 with } \nu \in (2/3, 1),\\ \nu_1 = 0, \nu_2 = 1 & \text{for Assumption LR2.} \end{cases}$$

Initialize $t_0 = 0$ and define, for $m = 0, 1, \ldots,$

$$t_{m+1} \doteq \min\left\{k \,\middle|\, \textstyle\sum_{t=t_m}^{k-1} \alpha_t \geq T_m\right\}.$$

Under this construction, the segment length satisfies

$$\bar{\alpha}_m \;\doteq\; \textstyle\sum_{t=t_m}^{t_{m+1}-1} \alpha_t \;\geq\; T_m, \tag{11}$$

so the time axis is partitioned into segments of length $\bar{\alpha}_m$, anchored at $\{t_m\}$. The relationship between the scales t and m is quantified below (cf. Lemma 1 and 2 of Qian et al. 2026):

Lemma 19 *There exist some C_{19} and m_0 such that for all $m \geq m_0$ and $t \geq t_m$, we have $\alpha_t \leq C_{19} T_m^2$.*

Lemma 20 *There exists some m_0 such that for all $m \geq m_0$, we have $\bar{\alpha}_m \leq 2T_m$.*

We now investigate the iterates w_t segment by segment. Telescoping (SA) yields

$$\begin{aligned} w_{t_{m+1}} =& w_{t_m} + \textstyle\sum_{t=t_m}^{t_{m+1}-1} \alpha_t H(w_t, Y_{t+1}) \\ =& w_{t_m} + \textstyle\sum_{t=t_m}^{t_{m+1}-1} \alpha_t (h(w_{t_m}) + H(w_t, Y_{t+1}) - h(w_{t_m})) \\ =& w_{t_m} + \bar{\alpha}_m h(w_{t_m}) + \underbrace{\textstyle\sum_{t=t_m}^{t_{m+1}-1} \alpha_t (H(w_t, Y_{t+1}) - h(w_{t_m}))}_{s_m}. \end{aligned}$$

We now examine the iterates w_t in (SA) with respect to the timescale t_m. Specifically, we can rewrite (SA) as (Skeleton SA):

$$w_{t_{m+1}} = w_{t_m} + \bar{\alpha}_m h(w_{t_m}) + s_m, \tag{Skeleton SA}$$

where s_m is the new noise and $\{\bar{\alpha}_m\}$ is the new learning rates.

We now construct a Lyapunov function $L(w) = \|w\|^2$ to study the behavior of $\{w_{t_m}\}$. Since $\|\cdot\|$ is an inner product norm, we obtain

$$\left\|w_{t_{m+1}}\right\|^2 = \|w_{t_m}\|^2 + 2\left\langle w_{t_m}, \bar{\alpha}_m h(w_{t_m}) + s_m \right\rangle + \|\bar{\alpha}_m h(w_{t_m}) + s_m\|^2. \tag{12}$$

To bound the last two terms, we follow Zou et al. (2019) to introduce an auxiliary Markov chain. For each time step t, we define a dedicated auxiliary chain, denoted by $\left\{\tilde{Y}_k^{(t)}\right\}_{k\geq 0}$. The chain $\left\{\tilde{Y}_k^{(t)}\right\}$ is constructed to be identical to $\{Y_k\}$ up to time step $k = t - \tau_{\alpha_t}$, after which it evolves independently according to the fixed transition matrix $P_{w_{t-\tau_{\alpha_t}}}$. This step is critical for addressing the time-inhomogeneity of $\{Y_k\}$, as its transition dynamics depend on the iterates $\{w_k\}$. By freezing the transition dynamics at $P_{w_{t-\tau_{\alpha_t}}}$ after $k = t - \tau_{\alpha_t}$, the auxiliary chain $\left\{\tilde{Y}_k^{(t)}\right\}$ isolates the effect of time-dependent noise, enabling rigorous control of the coupling between $\{w_k\}$ and $\{Y_k\}$. This approach is essential for deriving almost-sure convergence results and concentration bounds, which are more challenging in the time-inhomogeneous setting compared to the time-homogeneous case considered in Qian et al. (2026).

Let $\mathcal{F}_t \doteq \sigma(w_0, Y_1, \ldots, Y_t)$ be the filtration until time t and recall we use $\mathbb{E}_t[\cdot] \doteq \mathbb{E}[\cdot \mid \mathcal{F}_t]$ to denote the conditional expectation given $\mathcal{F}_t$. We then perform decomposition

$$s_m = \textstyle\sum_{t=t_m}^{t_{m+1}-1} \alpha_t (H(w_t, Y_{t+1}) - h(w_{t_m})) = s_{1,m} + s_{2,m} + s_{3,m} + s_{4,m},$$

where

$$\begin{aligned}
s_{1,m} &\doteq \textstyle\sum_{t=t_m}^{t_{m+1}-1} \alpha_t (H(w_t, Y_{t+1}) - H(w_{t_m}, Y_{t+1})), \\
s_{2,m} &\doteq \textstyle\sum_{t=t_m}^{t_{m+1}-1} \alpha_t \Big(H(w_{t_m}, Y_{t+1}) - H\Big(w_{t_m}, \tilde{Y}_{t+1}^{(t)}\Big)\Big), \\
s_{3,m} &\doteq \textstyle\sum_{t=t_m}^{t_{m+1}-1} \alpha_t \Big(H\Big(w_{t_m}, \tilde{Y}_{t+1}^{(t)}\Big) - \mathbb{E}_{t_m}\Big[H\Big(w_{t_m}, \tilde{Y}_{t+1}^{(t)}\Big)\Big]\Big), \\
s_{4,m} &\doteq \textstyle\sum_{t=t_m}^{t_{m+1}-1} \alpha_t \Big(\mathbb{E}_{t_m}\Big[H\Big(w_{t_m}, \tilde{Y}_{t+1}^{(t)}\Big)\Big] - h(w_{t_m})\Big).
\end{aligned}$$

Notably, $\mathbb{E}_{t_m}[s_{3,m}] = 0$. Taking $\mathbb{E}_{t_m}[\cdot]$ on both sides of (12) then yields

$$\begin{aligned}
\mathbb{E}_{t_m}\Big[\big\|w_{t_{m+1}}\big\|^2\Big] =& \|w_{t_m}\|^2 + 2\bar{\alpha}_m \langle w_{t_m}, h(w_{t_m})\rangle + 2\langle w_{t_m}, \mathbb{E}_{t_m}[s_{1,m} + s_{2,m} + s_{4,m}]\rangle \\
&+ \mathbb{E}_{t_m}\Big[\|\bar{\alpha}_m h(w_{t_m}) + s_{1,m} + s_{2,m} + s_{3,m} + s_{4,m}\|^2\Big].
\end{aligned} \tag{13}$$

We now bound the terms of the RHS one by one.

Lemma 21 *There exists some deterministic C_{21} and m_0 such that for all $m \geq m_0$,*

$$\|s_{1,m}\| \leq T_m^2 C_{21}(\|w_{t_m}\| + 1).$$

The proof is in Appendix C.4.

Lemma 22 *There exist some deterministic $C_{22,1}$, $C_{22,2}$ and m_0 such that for all $m \geq m_0$,*

$$\|\mathbb{E}_{t_m}[s_{2,m}]\| \leq T_m^2 C_{22,1}(\|w_{t_m}\| + 1), \tag{14}$$

and

$$\|s_{2,m}\| \leq T_m C_{22,2}(\|w_{t_m}\| + 1). \tag{15}$$

The proof is in Appendix C.5.

Lemma 23 *There exists some deterministic C_{23} and m_0 such that for all $m \geq m_0$,*

$$\|s_{3,m}\| \leq T_m C_{23}(\|w_{t_m}\| + 1).$$

The proof is identical to Lemma 6 in Qian et al. (2026) and thus is omitted.

Lemma 24 *There exists some deterministic C_{24} and m_0 such that for all $m \geq m_0$,*

$$\|s_{4,m}\| \leq T_m^2 C_{24}(\|w_{t_m}\| + 1).$$

The proof of this lemma is non-trivial and is the key to handling the time-inhomogeneous noise within the skeleton framework. The core of the proof is a case analysis based on the relationship between the secondary time scale's anchor point, t_m, and the starting point of the auxiliary chain, $t - \tau_{\alpha_t}$. In particular, for the more challenging case where $t_m < t - \tau_{\alpha_t}$, we employ the tower rule for conditional expectations to bound the bias term. The full proof is in Appendix C.6. Assembling the above bounds we obtain

Lemma 25 *There exist some deterministic $C_{25,1} > 0, C_{25,2}$ and m_0 such that for all $m \geq m_0$,*

$$\mathbb{E}_{t_m}\left[\left\|w_{t_{m+1}}\right\|^2\right] \leq (1 - C_{25,1}T_m)\|w_{t_m}\|^2 + C_{25,2}T_m.$$

The proof is in Appendix C.7.

C.1 Proof of Theorem 9

Proof We first establish the following lemma as an intermediate result.

Lemma 26 *For any $\nu \in (2/3, 1]$ and $\eta \in (0, 2\nu_2 - 1)$, it holds that*

$$\lim_{m\to\infty} m^{\eta/2}\Big(\|w_{t_m}\|^2 - \tfrac{\xi}{\alpha}\Big)^+ = 0 \quad a.s.$$

Proof Denote $z_m \doteq \|w_{t_m}\|^2$, Lemma 25 has verified (RS-Special). Recall that we proceed under Assumption LR1. Theorem 3 (A1) holds trivially, and Theorem 3 (A2) is verified by Lemma 13. We proceed to verify (A1) & (A2) of Theorem 5. Let the η in Theorem 5 be any number in $(0, 2\nu_2 - 1)$.

When $\nu < 1$. By the definition of T_m in (10), we have $T_m = \frac{C_\alpha}{(m+3)^{\nu_2}}$ with $\nu_2 < 1$. So when m is sufficiently large, it holds that $mT_m \to \infty$ as $m \to \infty$, thus $\liminf_{m\to\infty} mT_m > \frac{\eta}{C_{25,1}}$. Additionally, from $\eta < 2\nu_2 - 1$ we derive

$$\textstyle\sum_{m=1}^{\infty}(m+1)^\eta T_m^2 = C_\alpha^2 \sum_{m=1}^{\infty} \frac{(m+1)^\eta}{(m+3)^{2\nu_2}} < \infty.$$

When $\nu = 1$. We have $\nu_2 = 1$ and $\eta \in (0, 1)$, hence $T_m = \frac{C_\alpha \ln^{\nu_1}(m+3)}{m+3}$. Again we have $\liminf_{m\to\infty} mT_m = \infty > \frac{\eta}{C_{25,1}}$. Additionally, from $\eta < 1$ we obtain

$$\textstyle\sum_{m=1}^{\infty}(m+1)^\eta T_m^2 = C_\alpha^2 \sum_{m=1}^{\infty} \frac{(m+1)^\eta \ln^{2\nu_1}(m+3)}{(m+3)^2} < \infty.$$

Therefore, under Assumption LR1, identifying $\alpha = C_{25,1}$, $\xi = C_{25,2}$ and $B_{3,1} = C_{13}$, invoking Theorem 5 then completes the proof. ■

Having established the convergence rate of (Skeleton SA) in Lemma 26, we now complete the proof by mapping the convergence rate back to (SA). Recall the m_0 in Lemma 20. Then for all $m \geq m_0$, from Lemma 28, it is easy to see for any $t \in [t_m, t_{m+1}]$,

$$\|w_t\|^2 \leq 2C_{28}^2(\bar{\alpha}_m^2 + \|w_{t_m}\|^2) \leq 8C_{28}^2(T_m^2 + \|w_{t_m}\|^2).$$

Therefore, we have

$$\begin{aligned}\left(\|w_t\|^2 - 8C_{28}^2\tfrac{\xi}{\alpha}\right)^+ \leq& \max\left(0, 8C_{28}^2(T_m^2 + \|w_{t_m}\|^2) - 8C_{28}^2\tfrac{\xi}{\alpha}\right) \\ \leq& \max\left(0, 8C_{28}^2 T_m^2 + 8C_{28}^2\left(\|w_{t_m}\|^2 - \tfrac{\xi}{\alpha}\right)^+\right) \\ =& 8C_{28}^2 T_m^2 + 8C_{28}^2\left(\|w_{t_m}\|^2 - \tfrac{\xi}{\alpha}\right)^+.\end{aligned} \tag{16}$$

For any $\eta \in (0, 2\nu_2 - 1)$, we have

$$m^{\eta/2}\sup_{t\in[t_m,t_{m+1}]}\left(\|w_t\|^2 - 8C_{28}^2\tfrac{\xi}{\alpha}\right)^+ \leq 8C_{28}^2\left(m^{\eta/2}T_m^2 + m^{\eta/2}\left(\|w_{t_m}\|^2 - \tfrac{\xi}{\alpha}\right)^+\right).$$

The definition of T_m in (10) and Lemma 26 then confirms that

$$\lim_{m\to\infty} m^{\eta/2}\sup_{t\in[t_m,t_{m+1}]}\left(\|w_t\|^2 - 8C_{28}^2\tfrac{\xi}{\alpha}\right)^+ = 0.$$

Denote $B_9 \doteq \sqrt{8C_{28}^2\frac{\xi}{\alpha}}$. Notice

$$\begin{aligned}\left(\|w_t\|^2 - B_9^2\right)^+ =& \max\left(0, \|w_t\|^2 - B_9^2\right) \\ =& \max\left(0, (\|w_t\| - B_9)(\|w_t\| + B_9)\right) \\ =& (\|w_t\| + B_9)\max\left(0, \|w_t\| - B_9\right) \\ \geq& B_9(\|w_t\| - B_9)^+,\end{aligned}$$

we then obtain

$$\lim_{m\to\infty} m^{\eta/2}\sup_{t\in[t_m,t_{m+1}]}\left(\|w_t\| - B_9\right)^+ = 0.$$

To finalize the proof, we now convert the rate coefficient from m to t. **When $\nu < 1$.** Following Section A.10 of Qian et al. (2026), we can get $m^{1-\nu_2} = \Theta(t_m^{1-\nu})$, where $a_n = \Theta(b_n)$ means for two non-negative sequences $\{a_n\}$ and $\{b_n\}$, we have $a_n = \mathcal{O}(b_n)$ and $b_n = \mathcal{O}(a_n)$. Thus

$$\begin{aligned}0 &= \lim_{m\to\infty}(m-1)^{\eta/2}\sup_{t\in[t_{m-1},t_m]}\left(\|w_t\| - B_9\right)^+ \\ &= \lim_{m\to\infty} m^{\eta/2}\sup_{t\in[t_{m-1},t_m]}\left(\|w_t\| - B_9\right)^+ \\ &= \lim_{m\to\infty} t_m^{\frac{1-\nu}{2(1-\nu_2)}\eta}\sup_{t\in[t_{m-1},t_m]}\left(\|w_t\| - B_9\right)^+ \\ &\geq \lim_{m\to\infty}\sup_{t\in[t_{m-1},t_m]} t^{\frac{1-\nu}{2(1-\nu_2)}\eta}(\|w_t\| - B_9)^+.\end{aligned}$$

We now optimize the choice of $\nu_2 \in (1/2, \nu/(2-\nu))$ and $\eta \in (0, 2\nu_2 - 1)$. Notice that $\sup_{\nu_2\in(1/2,\nu/(2-\nu)),\eta\in(0,2\nu_2-1)}\frac{1-\nu}{2(1-\nu_2)}\eta = \frac{3}{4}\nu - \frac{1}{2}$. We then conclude that for any $\zeta < \frac{3}{4}\nu - \frac{1}{2}$,

$$\lim_{t\to\infty} t^{\zeta}(\|w_t\| - B_9)^+ = 0.$$

When $\nu = 1$. Again following Section A.10 of Qian et al. (2026), we have

$$\ln(m+3) \geq (\ln(t_{m+1}+2))^{\frac{1}{1+\nu_1}} - B_{9,1},$$

where $B_{9,1} > 0$ is some constant. Define $B_{9,2} \doteq \exp(-\zeta B_{9,1})$, we then have

$$\begin{aligned} 0 &= \lim_{m\to\infty}(m+2)^{\zeta} \sup_{t\in[t_{m-1},t_m]} (\|w_t\| - B_9)^+ \\ &\geq \lim_{m\to\infty} B_{9,2} \exp\left(\zeta \ln^{1/(1+\nu_1)} t_m\right) \sup_{t\in[t_{m-1},t_m]} (\|w_t\| - B_9)^+ \\ &\geq B_{9,2} \lim_{t\to\infty} \exp\left(\zeta \ln^{1/(1+\nu_1)} t\right) (\|w_t\| - B_9)^+. \end{aligned}$$

Notice that $(\|w_t\| - B_9)^+$ is actually the distance from w_t to the ball $\mathcal{B}(B_9)$, i.e. $d(w_t, \mathcal{B}(B_9))$. This completes the proof. ■

C.2 Proof of Theorem 10

Proof The proof is analogous to the previous section. Under Assumption LR2, Assumption 4.1 is satisfied by (10) with $C_\alpha > \bar{C}_\alpha$. Recall that all other assumptions are already verified in Appendix C.1, we can apply Theorem 6 to obtain the concentration bound of $\{w_{t_m}\}$.

Lemma 27 *Under Assumption LR2, there exist some constants C_{27} and C'_{27} such that for any $\delta > 0$, it holds, with probability at least $1-\delta$, that for all $m \geq 0$,*

$$\left(\left(\|w_{t_m}\|^2 - \tfrac{\xi}{\alpha}\right)^+\right)^2 \leq C'_{27}\tfrac{1}{m+3}\left[\ln\left(\tfrac{C_{27}}{\delta}\right) + 1 + \ln(m+3)\right]^{C_{27}}.$$

Now we map this result from m back to t. First recall that we proceed under Assumption LR2 and have

$$\alpha_t = \tfrac{C_\alpha}{(t+3)\ln^\nu(t+3)}, \quad T_m = \tfrac{C_\alpha}{m+3}.$$

Applying the bound in Lemma 27 to (16), similar to Appendix C.1, for any $t \in [t_m, t_{m+1}]$

$$\begin{aligned} (\|w_t\| - B_9)^+ \leq& \tfrac{1}{B_9}\left(\|w_t\|^2 - B_9^2\right)^+ \\ \leq& \tfrac{1}{B_9} 8C_{28}(T_m^2 + \sqrt{C'_{27}}\tfrac{1}{\sqrt{m+3}}[\ln(C_{27}/\delta) + 1 + \ln(m+3)]^{C_{27}/2}). \end{aligned}$$

Following Section A.10 of Qian et al. (2026), we obtain $m+3 \leq (m_0+3)\exp\left(\frac{\ln^{1-\nu}(t_{m+1}+1)}{1-\nu}\right)$. Additionally, there exists some constant $B_{10,1}$ such that $m+2 \geq B_{10,1}\exp\left(\frac{\ln^{1-\nu} t_{m+1}}{1-\nu}\right)$. Then for all $m \geq m_0$ and any $t \in [t_m, t_{m+1}]$, we have, for some proper constants,

$$\begin{aligned} &(\|w_t\| - B_9)^+ \\ \leq& \tfrac{1}{B_9} 8C_{28}\left(\tfrac{C_\alpha^2}{(m+3)^2} + \sqrt{C'_{27}}\tfrac{1}{\sqrt{m+3}}[\ln(C_{27}/\delta) + 1 + \ln(m+3)]^{C_{27}/2}\right) \end{aligned}$$

$$\leq B_{10,2}\frac{1}{\sqrt{m+3}}\left[\ln(1/\delta)+B_{10,3}+\ln(m+3)\right]^{C_{27}/2}$$
$$\leq B_{10,2}\left(B_{10,1}\exp\left(\frac{\ln^{1-\nu}(t_{m+1})}{2(1-\nu)}\right)\right)^{-1}\left[\ln(1/\delta)+B_{10,3}+\ln\left((m_0+3)\exp\left(\frac{\ln^{1-\nu}(t_{m+1})}{1-\nu}\right)\right)\right]^{C_{27}/2}$$

$$\leq B_{10,4}\exp\left(\frac{-\ln^{1-\nu}(t+1)}{2(1-\nu)}\right)\left[\ln(1/\delta)+B_{10,5}+\frac{\ln^{1-\nu}(t+1)}{1-\nu}\right]^{C_{27}/2}.$$

For all $t \leq t_{m_0}$, we can again use a trivial almost sure bound from (18), following a similar argument as in the case of $n < \bar{n}$ in Lemma 17 in Section B.2.3. Combining the two bounds then completes the proof. ■

C.3 Auxiliary Lemma

Lemma 28 *There exists some constant C_{28} such that for any m and any $t \in [t_m, t_{m+1}]$,*

$$\begin{aligned}\|w_t\| &\leq C_{28}(\bar{\alpha}_m + \|w_{t_m}\|),\\ \|w_t - w_{t_m}\| &\leq 2T_m C_{28}(\|w_{t_m}\|+1).\end{aligned} \tag{17}$$

Furthermore, for any $t \geq 0$, it holds that

$$\|w_t\| \leq \left(\|w_0\| + C_{28}\textstyle\sum_{\tau=0}^{t-1}\alpha_\tau\right)\exp\left(C_{28}\textstyle\sum_{\tau=0}^{t-1}\alpha_\tau\right). \tag{18}$$

Proof Given that we allow time-inhomogeneous transition kernel, we cannot directly apply Lemma 16 from Qian et al. (2026). Nevertheless, under Assumption 5.3, we can establish that

$$\|H(w,y)\| \leq \|H(w,y)-H(0,y)\| + \|H(0,y)\| \leq L_h\|w-0\| + L_h = L_h\|w\| + L_h.$$

Crucially, this bound is independent of the state y. Because the bound holds uniformly, the specific dynamics of the Markov chain (whether time-homogeneous or inhomogeneous) do not influence the subsequent analysis for this lemma. The remainder of the proof therefore proceeds identically to the approach in Qian et al. (2026), which relies primarily on an application of the discrete Gronwall's inequality to establish the bounds on the iterates. ■

Lemma 29 *Let $\alpha_{i,j} \doteq \sum_{k=i}^{j}\alpha_k$. For any $t_2 > t_1$, if $\alpha_{t_1,t_2-1}L_h < 1$, then the following bound holds*

$$\|w_{t_1} - w_{t_2}\| \leq \frac{\alpha_{t_1,t_2-1}L_h}{1-\alpha_{t_1,t_2-1}L_h}(\|w_{t_2}\|+1).$$

Proof We first establish a uniform bound on the norm of any intermediate iterate w_i for $i \in [t_1, t_2-1]$ in terms of $\|w_{t_2}\|$. For any $i \in [t_1, t_2-1]$, we can express w_i in terms of the future iterate w_{t_2} as

$$w_i = w_{t_2} - \textstyle\sum_{j=i}^{t_2-1}\alpha_j H(w_j, Y_{j+1}).$$

By taking the norm and applying the triangle inequality, along with Assumption 5.3, we have

$$\|w_i\| \leq \|w_{t_2}\| + \sum_{j=i}^{t_2-1} \alpha_j \|H(w_j, Y_{j+1})\| \leq \|w_{t_2}\| + \sum_{j=i}^{t_2-1} \alpha_j L_h(\|w_j\| + 1).$$

Denote $M \doteq \max_{k\in[t_1,t_2-1]} \|w_k\|$. Then for any j in this interval, $\|w_j\| \leq M$. The inequality above thus becomes

$$\|w_i\| \leq \|w_{t_2}\| + \left(\sum_{j=t_1}^{t_2-1} \alpha_j\right) L_h(M+1) = \|w_{t_2}\| + \alpha_{t_1,t_2-1} L_h(M+1).$$

Since this holds for all $i \in [t_1, t_2 - 1]$, it must also hold for the iterate that achieves the maximum norm M. Therefore

$$M \leq \|w_{t_2}\| + \alpha_{t_1,t_2-1} L_h(M+1).$$

Given that $\alpha_{t_1,t_2-1} L_h < 1$, we can solve for M

$$M = \max_{i\in[t_1,t_2-1]} \|w_i\| \leq \frac{\|w_{t_2}\| + \alpha_{t_1,t_2-1} L_h}{1 - \alpha_{t_1,t_2-1} L_h}. \tag{19}$$

We now bound the norm of the difference between w_{t_1} and w_{t_2}

$$\begin{aligned}
\|w_{t_2} - w_{t_1}\| =& \left\| \sum_{i=t_1}^{t_2-1} \alpha_i H(w_i, Y_{i+1}) \right\| \\
\leq& \sum_{i=t_1}^{t_2-1} \alpha_i L_h(\|w_i\| + 1) \\
\leq& \sum_{i=t_1}^{t_2-1} \alpha_i L_h \left(\frac{\|w_{t_2}\| + \alpha_{t_1,t_2-1} L_h}{1 - \alpha_{t_1,t_2-1} L_h} + 1 \right). && \text{(By (19))} \\
\leq& \sum_{i=t_1}^{t_2-1} \alpha_i L_h \left(\frac{\|w_{t_2}\| + 1}{1 - \alpha_{t_1,t_2-1} L_h} \right) \\
=& \frac{\alpha_{t_1,t_2-1} L_h}{1 - \alpha_{t_1,t_2-1} L_h} (\|w_{t_2}\| + 1).
\end{aligned}$$

■

C.4 Proof of Lemma 21

Proof

$$\begin{aligned}
\|s_{1,m}\| \leq& \sum_{t=t_m}^{t_{m+1}-1} \alpha_t L_h \|w_t - w_{t_m}\| \\
\leq& \sum_{t=t_m}^{t_{m+1}-1} \alpha_t L_h 2T_m C_{28}(\|w_{t_m}\| + 1) && \text{(Lemma 28)} \\
=& 2T_m \bar{\alpha}_m L_h C_{28}(\|w_{t_m}\| + 1) \\
\leq& 4T_m^2 L_h C_{28}(\|w_{t_m}\| + 1). && \text{(Lemma 20)}
\end{aligned}$$

■

C.5 Proof of Lemma 22

Proof Denote $\alpha_{i,j} \doteq \sum_{k=i}^{j} \alpha_k$. Using Lemma 19, we have

$$\alpha_{t-\tau_{\alpha_t},t-1} = \mathcal{O}(\alpha_t \tau_{\alpha_t}) = \mathcal{O}(\alpha_t \ln t).$$

Indeed, by (5),

$$\tau_{\alpha_t} = \left\lceil \frac{\ln \alpha_t - \ln C_{5.1}}{\ln \tau} \right\rceil = \begin{cases} \mathcal{O}(\nu \ln t), & \text{for Assumption LR1,} \\ \mathcal{O}(\ln t), & \text{for Assumption LR2.} \end{cases}$$

For sufficiently large m_0, we then have $t - \tau_{\alpha_t} \geq 0$ for all $t \in [t_m, t_{m+1} - 1]$ and all $m \geq m_0$. We now record the conditional version of the auxiliary-chain comparison. Applying the same telescoping argument as in Section B.4 of Liu et al. (2025) after conditioning on $\mathcal{F}_{t-\tau_{\alpha_t}}$, there exists some deterministic constant $C_{22,3}$ such that for all sufficiently large t,

$$\sum_{y'} \left| \Pr\big(Y_{t+1} = y' \mid \mathcal{F}_{t-\tau_{\alpha_t}}\big) - \Pr\Big(\tilde{Y}_{t+1}^{(t)} = y' \mid \mathcal{F}_{t-\tau_{\alpha_t}}\Big) \right| \leq C_{22,3} |\mathcal{Y}| \alpha_{t-\tau_{\alpha_t},t-1} \ln(t + t_0) \quad \text{a.s.} \tag{20}$$

This is the conditional form that will be used below.

We next establish the scale estimate needed for the main part of the proof. According to Section A.10 and A.14 of Qian et al. (2026), we have

$$\ln t_{m+1} = \begin{cases} \mathcal{O}\big(\ln^{1+\nu_1} m\big), & \text{for Assumption LR1 with } \nu = 1, \\ \frac{1-\nu_2}{1-\nu} \mathcal{O}(\ln m), & \text{for Assumption LR1 with } \nu \in (2/3, 1), \\ (1-\nu)^{\frac{1}{1-\nu}} \mathcal{O}\Big(\ln^{\frac{1}{1-\nu}} m\Big), & \text{for Assumption LR2.} \end{cases}$$

Combining these bounds with Lemma 19 gives

$$\max_{t_m \leq t < t_{m+1}} \big\{ \alpha_{t-\tau_{\alpha_t},t-1} \ln(t + t_0) \big\} = \mathcal{O}(T_m). \tag{21}$$

Define

$$I_m^- \doteq \{t \in \{t_m, \ldots, t_{m+1} - 1\} : t - \tau_{\alpha_t} < t_m\},$$
$$I_m^+ \doteq \{t \in \{t_m, \ldots, t_{m+1} - 1\} : t - \tau_{\alpha_t} \geq t_m\}.$$

Let $\tau_m^\star \doteq \max_{t_m \leq t < t_{m+1}} \tau_{\alpha_t}$. If $t \in I_m^-$, then $t - t_m < \tau_{\alpha_t} \leq \tau_m^\star$. Hence, using the monotonicity of α_t,

$$\sum_{t \in I_m^-} \alpha_t \leq \sum_{j=0}^{\tau_m^\star} \alpha_{t_m+j} \leq (\tau_m^\star + 1) \alpha_{t_m} \leq C \alpha_{t_m} \ln(t_{m+1} + t_0) = \mathcal{O}(T_m^2). \tag{22}$$

For Assumption LR1 with $\nu \in (2/3, 1)$, the last equality uses $m^{1-\nu_2} = \Theta(t_m^{1-\nu})$ and the strict choice $\nu_2 < \nu/(2-\nu)$; for Assumption LR1 with $\nu = 1$ and for Assumption LR2, it follows from the same scale relations in Section A.10 and A.14 of Qian et al. (2026).

We now prove (14). For the boundary indices $t \in I_m^-$, we use only the crude bound from Assumption 5.3:

$$\begin{aligned}
&\left\|\mathbb{E}_{t_m}\left[H(w_{t_m}, Y_{t+1}) - H(w_{t_m}, \tilde{Y}_{t+1}^{(t)})\right]\right\| \\
\leq &\mathbb{E}_{t_m}\left[\|H(w_{t_m}, Y_{t+1})\| + \left\|H(w_{t_m}, \tilde{Y}_{t+1}^{(t)})\right\|\right] \\
\leq &2L_h(\|w_{t_m}\| + 1).
\end{aligned}$$

Therefore, by (22),

$$\textstyle\sum_{t\in I_m^-} \alpha_t \left\|\mathbb{E}_{t_m}\left[H(w_{t_m}, Y_{t+1}) - H(w_{t_m}, \tilde{Y}_{t+1}^{(t)})\right]\right\| \leq \mathcal{O}(T_m^2)(\|w_{t_m}\| + 1). \tag{23}$$

For the main indices $t \in I_m^+$, we have $t - \tau_{\alpha_t} \geq t_m$, so w_{t_m} is $\mathcal{F}_{t-\tau_{\alpha_t}}$-measurable. By the tower rule and (20),

$$\begin{aligned}
&\left\|\mathbb{E}_{t_m}\left[H(w_{t_m}, Y_{t+1}) - H(w_{t_m}, \tilde{Y}_{t+1}^{(t)})\right]\right\| \\
= &\left\|\mathbb{E}_{t_m}\left[\mathbb{E}_{t-\tau_{\alpha_t}}\left[H(w_{t_m}, Y_{t+1}) - H(w_{t_m}, \tilde{Y}_{t+1}^{(t)})\right]\right]\right\| \\
\leq &\mathbb{E}_{t_m}\left[\left\|\mathbb{E}_{t-\tau_{\alpha_t}}\left[H(w_{t_m}, Y_{t+1}) - H(w_{t_m}, \tilde{Y}_{t+1}^{(t)})\right]\right\|\right] \\
= &\mathbb{E}_{t_m}\left[\left\|\textstyle\sum_{y'} H(w_{t_m}, y')\Big(\Pr\big(Y_{t+1} = y' \mid \mathcal{F}_{t-\tau_{\alpha_t}}\big) - \Pr\Big(\tilde{Y}_{t+1}^{(t)} = y' \mid \mathcal{F}_{t-\tau_{\alpha_t}}\Big)\Big)\right\|\right] \\
\leq &L_h(\|w_{t_m}\| + 1)\mathbb{E}_{t_m}\left[\textstyle\sum_{y'}\left|\Pr\big(Y_{t+1} = y' \mid \mathcal{F}_{t-\tau_{\alpha_t}}\big) - \Pr\Big(\tilde{Y}_{t+1}^{(t)} = y' \mid \mathcal{F}_{t-\tau_{\alpha_t}}\Big)\right|\right] \\
\leq &L_h(\|w_{t_m}\| + 1)C_{22,3}|\mathcal{Y}|\alpha_{t-\tau_{\alpha_t},t-1}\ln(t + t_0).
\end{aligned}$$

Consequently,

$$\begin{aligned}
&\textstyle\sum_{t\in I_m^+} \alpha_t \left\|\mathbb{E}_{t_m}\left[H(w_{t_m}, Y_{t+1}) - H(w_{t_m}, \tilde{Y}_{t+1}^{(t)})\right]\right\| && (24)\\
\leq &L_h(\|w_{t_m}\| + 1)C_{22,3}|\mathcal{Y}|\textstyle\sum_{t\in I_m^+}\alpha_t\alpha_{t-\tau_{\alpha_t},t-1}\ln(t + t_0) \\
\leq &L_h(\|w_{t_m}\| + 1)C_{22,3}|\mathcal{Y}|\alpha_{t_m,t_{m+1}-1}\max_{t_m\leq t<t_{m+1}}\left\{\alpha_{t-\tau_{\alpha_t},t-1}\ln(t + t_0)\right\} \\
\leq &2T_mL_h(\|w_{t_m}\| + 1)C_{22,3}|\mathcal{Y}|\max_{t_m\leq t<t_{m+1}}\left\{\alpha_{t-\tau_{\alpha_t},t-1}\ln(t + t_0)\right\} && \text{(By Lemma 20)} \\
= &\mathcal{O}(T_m^2)(\|w_{t_m}\| + 1). && \text{(By (21))}
\end{aligned}$$

Combining (23) and (24), we obtain

$$\|\mathbb{E}_{t_m}[s_{2,m}]\| \leq T_m^2 C_{22,1}(\|w_{t_m}\| + 1)$$

for a sufficiently large deterministic constant $C_{22,1}$. This proves (14).

We now prove (15). By Assumption 5.3,

$$\begin{aligned}
\|s_{2,m}\| \leq &\textstyle\sum_{t=t_m}^{t_{m+1}-1}\alpha_t\Big(\|H(w_{t_m}, Y_{t+1})\| + \left\|H(w_{t_m}, \tilde{Y}_{t+1}^{(t)})\right\|\Big) \\
\leq &2L_h(\|w_{t_m}\| + 1)\bar{\alpha}_m \\
\leq &4L_hT_m(\|w_{t_m}\| + 1). && \text{(By Lemma 20)}
\end{aligned}$$

Selecting $C_{22,2}$ accordingly completes the proof. ■

C.6 Proof of Lemma 24

Proof We first decompose $s_{4,m}$ as

$$\begin{aligned}\|s_{4,m}\| \leq& \textstyle\sum_{t=t_m}^{t_{m+1}-1} \alpha_t \Big\| \mathbb{E}_{t_m}\Big[H\Big(w_{t_m}, \tilde{Y}_{t+1}^{(t)}\Big) - h\big(w_{t-\tau_{\alpha_t}}\big)\Big]\Big\| \\ \leq& \textstyle\sum_{t=t_m}^{t_{m+1}-1} \alpha_t \Big\| \mathbb{E}_{t_m}\Big[H\Big(w_{t_m}, \tilde{Y}_{t+1}^{(t)}\Big) - h\big(w_{t-\tau_{\alpha_t}}\big)\Big]\Big\| \\ &+ \sum_{t=t_m}^{t_{m+1}-1} \alpha_t \big\| \mathbb{E}_{t_m}\big[h\big(w_{t-\tau_{\alpha_t}}\big)\big] - h(w_{t_m})\big\|.\end{aligned}$$

We start from the first term by bounding each summand indexed by $t \in [t_m, t_{m+1}-1]$. **When $t_m \geq t - \tau_{\alpha_t}$.** We decompose the summand as

$$\begin{aligned}&\Big\| \mathbb{E}_{t_m}\Big[H\Big(w_{t_m}, \tilde{Y}_{t+1}^{(t)}\Big) - h\big(w_{t-\tau_{\alpha_t}}\big)\Big]\Big\| \\ =&\Big\| \textstyle\sum_y P_{w_{t-\tau_{\alpha_t}}}^{\tau_{\alpha_t}+1}\big(Y_{t-\tau_{\alpha_t}}, y\big) H(w_{t_m}, y) - \sum_y d_{\mathcal{Y}, w_{t-\tau_{\alpha_t}}}(y) H\big(w_{t-\tau_{\alpha_t}}, y\big)\Big\| \\ \leq&\Big\| \textstyle\sum_y \Big(P_{w_{t-\tau_{\alpha_t}}}^{\tau_{\alpha_t}+1}\big(Y_{t-\tau_{\alpha_t}}, y\big) - d_{\mathcal{Y}, w_{t-\tau_{\alpha_t}}}(y)\Big) H(w_{t_m}, y)\Big\| \\ &+ \Big\| \sum_y d_{\mathcal{Y}, w_{t-\tau_{\alpha_t}}}(y)\big(H(w_{t_m}, y) - H\big(w_{t-\tau_{\alpha_t}}, y\big)\big)\Big\|.\end{aligned}$$

To see the first equality, we recall that $\mathbb{E}_{t_m}$ is the conditional expectation conditioned on $Y_1, Y_2, \ldots, Y_{t_m}$, so $\left\{\tilde{Y}_k^{(t)}\right\}$ is not adapted to $\mathcal{F}_{t_m}$ for $k = t - \tau_{\alpha_t} + 1, \ldots, t+1$. Then

$$\begin{aligned}&\Big\| \textstyle\sum_y \Big(P_{w_{t-\tau_{\alpha_t}}}^{\tau_{\alpha_t}+1}\big(Y_{t-\tau_{\alpha_t}}, y\big) - d_{\mathcal{Y}, w_{t-\tau_{\alpha_t}}}(y)\Big) H(w_{t_m}, y)\Big\| \\ \leq& L_h(\|w_{t_m}\|+1) \cdot \textstyle\sum_y \Big| P_{w_{t-\tau_{\alpha_t}}}^{\tau_{\alpha_t}+1}\big(Y_{t-\tau_{\alpha_t}}, y\big) - d_{\mathcal{Y}, w_{t-\tau_{\alpha_t}}}(y)\Big| && \text{(Assumption 5.3)} \\ \leq& L_h(\|w_{t_m}\|+1) \cdot C_{5.1}\tau^{\tau_{\alpha_t}+1} && \text{(By (5))} \\ \leq& L_h(\|w_{t_m}\|+1) \cdot \alpha_t, && \text{(By (6))}\end{aligned}$$

and

$$\begin{aligned}&\Big\| \textstyle\sum_y d_{\mathcal{Y}, w_{t-\tau_{\alpha_t}}}(y)\big(H(w_{t_m}, y) - H\big(w_{t-\tau_{\alpha_t}}, y\big)\big)\Big\| \\ \leq& \textstyle\sum_y d_{\mathcal{Y}, w_{t-\tau_{\alpha_t}}}(y) \big\| H(w_{t_m}, y) - H\big(w_{t-\tau_{\alpha_t}}, y\big)\big\| \\ \leq& L_h \big\| w_{t_m} - w_{t-\tau_{\alpha_t}}\big\| && \text{(Assumption 5.3)} \\ \leq& L_h \tfrac{\alpha_{t-\tau_{\alpha_t}, t_m-1} L_h}{1-\alpha_{t-\tau_{\alpha_t}, t_m-1} L_h}(\|w_{t_m}\|+1) && \text{(Lemma 29)} \\ \leq& 2L_h^2 \alpha_{t-\tau_{\alpha_t}, t-1}(\|w_{t_m}\|+1).\end{aligned}$$

To see the last inequality (we only need it to hold for sufficiently large t), we first notice that $\alpha_{t-\tau_{\alpha_t}, t_m-1} \leq \alpha_{t-\tau_{\alpha_t}, t-1}$ since we are bounding terms for $t \in [t_m, t_{m+1}-1]$. We also have $\lim_{t\to\infty} \alpha_{t-\tau_{\alpha_t}, t-1} = 0$. So for sufficiently large t, we have $1 - \alpha_{t-\tau_{\alpha_t}, t_m-1} L_h > 1/2$, which yields the last inequality. Putting everything together, we obtain

$$\textstyle\sum_{t=t_m}^{t_{m+1}-1} \alpha_t \Big\| \mathbb{E}_{t_m}\Big[H\Big(w_{t_m}, \tilde{Y}_{t+1}^{(t)}\Big) - h\big(w_{t-\tau_{\alpha_t}}\big)\Big]\Big\|$$

$$
\begin{aligned}
&\le \textstyle\sum_{t=t_m}^{t_{m+1}-1} \alpha_t \cdot L_h(\|w_{t_m}\|+1)\cdot \alpha_t + \sum_{t=t_m}^{t_{m+1}-1} \alpha_t \cdot L_h^2 \alpha_{t-\tau_{\alpha_t},t-1}(\|w_{t_m}\|+1)\\
&= L_h(\|w_{t_m}\|+1)\Big(\textstyle\sum_{t=t_m}^{t_{m+1}-1} \alpha_t^2 + \sum_{t=t_m}^{t_{m+1}-1} \alpha_t \cdot L_h \alpha_{t-\tau_{\alpha_t},t-1}\Big)\\
&\le L_h(\|w_{t_m}\|+1)\big(\alpha_{t_m}\alpha_{t_m,t_{m+1}-1} + \alpha_{t_m,t_{m+1}-1} L_h \max_{t\in[t_m,t_{m+1}-1]} \alpha_{t-\tau_{\alpha_t},t-1}\big)\\
&\le L_h(\|w_{t_m}\|+1)\big(C_{19}T_m^2\cdot 2T_m + 2T_m L_h \max_{t\in[t_m,t_{m+1}-1]} \alpha_{t-\tau_{\alpha_t},t-1}\big)\\
&\qquad\qquad \text{(By (11), Lemma 19 and Lemma 20)}\\
&= \mathcal{O}(T_m^2)(\|w_{t_m}\|+1). \qquad \text{(By (21))}
\end{aligned}
$$

When $t_m < t-\tau_{\alpha_t}$. Using the tower rule we have

$$
\mathbb{E}_{t_m}\Big[H\Big(w_{t_m},\tilde{Y}_{t+1}^{(t)}\Big) - h\big(w_{t-\tau_{\alpha_t}}\big)\Big] = \mathbb{E}_{t_m}\Big[\mathbb{E}_{t-\tau_{\alpha_t}}\Big[H\Big(w_{t_m},\tilde{Y}_{t+1}^{(t)}\Big)\Big] - h\big(w_{t-\tau_{\alpha_t}}\big)\Big].
$$

Consider the inner expectation term

$$
\begin{aligned}
&\Big\|\mathbb{E}_{t-\tau_{\alpha_t}}\Big[H\Big(w_{t_m},\tilde{Y}_{t+1}^{(t)}\Big)\Big] - h\big(w_{t-\tau_{\alpha_t}}\big)\Big\|\\
=&\Big\|\textstyle\sum_y P_{w_{t-\tau_{\alpha_t}}}^{\tau_{\alpha_t}+1}\big(Y_{t-\tau_{\alpha_t}},y\big)H(w_{t_m},y) - h\big(w_{t-\tau_{\alpha_t}}\big)\Big\|\\
=&\Big\|\textstyle\sum_y \Big(P_{w_{t-\tau_{\alpha_t}}}^{\tau_{\alpha_t}+1}\big(Y_{t-\tau_{\alpha_t}},y\big) - d_{\mathcal{Y},w_{t-\tau_{\alpha_t}}}(y)\Big)H(w_{t_m},y)\Big\|\\
&+\Big\|\textstyle\sum_y d_{\mathcal{Y},w_{t-\tau_{\alpha_t}}}(y)\big(H(w_{t_m},y) - H\big(w_{t-\tau_{\alpha_t}},y\big)\big)\Big\|.
\end{aligned}
$$

Then

$$
\begin{aligned}
&\Big\|\textstyle\sum_y \Big(P_{w_{t-\tau_{\alpha_t}}}^{\tau_{\alpha_t}+1}\big(Y_{t-\tau_{\alpha_t}},y\big) - d_{\mathcal{Y},w_{t-\tau_{\alpha_t}}}(y)\Big)H(w_{t_m},y)\Big\|\\
\le& L_h(\|w_{t_m}\|+1)\textstyle\sum_y \Big|P_{w_{t-\tau_{\alpha_t}}}^{\tau_{\alpha_t}+1}\big(Y_{t-\tau_{\alpha_t}},y\big) - d_{\mathcal{Y},w_{t-\tau_{\alpha_t}}}(y)\Big|\\
\le& L_h(\|w_{t_m}\|+1)C_{5.1}\alpha_t, \qquad \text{(By (5) and (6))}
\end{aligned}
$$

and

$$
\begin{aligned}
&\Big\|\textstyle\sum_y d_{\mathcal{Y},w_{t-\tau_{\alpha_t}}}(y)\big(H(w_{t_m},y) - H\big(w_{t-\tau_{\alpha_t}},y\big)\big)\Big\|\\
\le& \textstyle\sum_y d_{\mathcal{Y},w_{t-\tau_{\alpha_t}}}(y)\big\|H(w_{t_m},y) - H\big(w_{t-\tau_{\alpha_t}},y\big)\big\|\\
\le& L_h\big\|w_{t_m} - w_{t-\tau_{\alpha_t}}\big\| \qquad \text{(Assumption 5.3)}\\
\le& L_h C_{28}\cdot 2T_m(\|w_{t_m}\|+1) \qquad \text{(Lemma 28)}\\
\le& T_m L_h C_{24,1}(\|w_{t_m}\|+1).
\end{aligned}
$$

That is, we obtain

$$
\begin{aligned}
&\textstyle\sum_{t=t_m}^{t_{m+1}-1}\alpha_t\Big\|\mathbb{E}_{t_m}\Big[H\Big(w_{t_m},\tilde{Y}_{t+1}^{(t)}\Big) - h\big(w_{t-\tau_{\alpha_t}}\big)\Big]\Big\|\\
\le& \textstyle\sum_{t=t_m}^{t_{m+1}-1}\alpha_t\big(L_h(\|w_{t_m}\|+1)\cdot C_{5.1}\alpha_t + T_m L_h C_{24,1}(\|w_{t_m}\|+1)\big)\\
=&(C_{5.1}+C_{24,1})L_h(\|w_{t_m}\|+1)\textstyle\sum_{t=t_m}^{t_{m+1}-1}\alpha_t(\alpha_t+T_m)\\
\le&(C_{5.1}+C_{24,1})L_h(\|w_{t_m}\|+1)2T_m\big(\max_{t\in[t_m,t_{m+1}-1]}\alpha_t + T_m\big) \quad \text{(By (11) and Lemma 20)}\\
\le& C_{24,2}T_m^2(\|w_{t_m}\|+1). \qquad \text{(Lemma 20)}
\end{aligned}
$$

We now consider the second term. We first establish the Lipschitz continuity of $h(w)$.

Lemma 30 *Under Assumptions 5.1-5.3, there exists some deterministic L_h' such that for any w_1, w_2*

$$\|h(w_1) - h(w_2)\| \leq L_h' \|w_1 - w_2\|.$$

Proof Recall $h(w)$ is defined as $h(w) = \sum_{y\in\mathcal{Y}} d_{\mathcal{Y},w}(y) H(w,y)$, we have

$$\begin{aligned}
\|h(w_1) - h(w_2)\| =& \Big\| \textstyle\sum_y d_{\mathcal{Y},w_1}(y) H(w_1,y) - \sum_y d_{\mathcal{Y},w_2}(y) H(w_2,y) \Big\| \\
=& \Big\| \textstyle\sum_y d_{\mathcal{Y},w_1}(y)[H(w_1,y) - H(w_2,y)] + \sum_y [d_{\mathcal{Y},w_1}(y) - d_{\mathcal{Y},w_2}(y)] H(w_2,y) \Big\| \\
\leq& \Big\| \textstyle\sum_y d_{\mathcal{Y},w_1}(y)[H(w_1,y) - H(w_2,y)] \Big\| \\
&+ \Big\| \sum_y [d_{\mathcal{Y},w_1}(y) - d_{\mathcal{Y},w_2}(y)] H(w_2,y) \Big\|.
\end{aligned}$$

Then we have

$$\begin{aligned}
& \Big\| \textstyle\sum_y d_{\mathcal{Y},w_1}(y)[H(w_1,y) - H(w_2,y)] \Big\| \\
\leq& \textstyle\sum_y d_{\mathcal{Y},w_1}(y) \|H(w_1,y) - H(w_2,y)\| && \text{(Assumption 5.3)} \\
\leq& L_h \|w_1 - w_2\|, && \text{(Assumption 5.3)}
\end{aligned}$$

and

$$\Big\| \textstyle\sum_y [d_{\mathcal{Y},w_1}(y) - d_{\mathcal{Y},w_2}(y)] H(w_2,y) \Big\| \leq \|d_{\mathcal{Y},w_1} - d_{\mathcal{Y},w_2}\|_1 \max_y \|H(w_2,y)\|,$$

where $\|\cdot\|_1$ is the l_1 norm. By Lemma 5 of Liu et al. (2025), there exists some constant C_d such that $\|d_{\mathcal{Y},w_1} - d_{\mathcal{Y},w_2}\|_1 \leq \frac{C_d}{1+\|w_1\|+\|w_2\|} \|w_1 - w_2\|$. Combining with Assumption 5.2, we get

$$\begin{aligned}
\Big\| \textstyle\sum_y [d_{\mathcal{Y},w_1}(y) - d_{\mathcal{Y},w_2}(y)] H(w_2,y) \Big\| \leq& \Big(\tfrac{C_d}{1+\|w_1\|+\|w_2\|} \|w_1 - w_2\| \Big) \cdot L_h(\|w_2\| + 1) \\
\leq& C_d L_h \|w_1 - w_2\|.
\end{aligned}$$

Putting the bounds together we get

$$\|h(w_1) - h(w_2)\| \leq L_h \|w_1 - w_2\| + C_d L_h \|w_1 - w_2\| = L_h' \|w_1 - w_2\|,$$

where $L_h' \doteq L_h(1 + C_d)$. ■

With this lemma, we can derive

$$\begin{aligned}
\textstyle\sum_{t=t_m}^{t_{m+1}-1} \alpha_t \big\| \mathbb{E}_{t_m}\big[h\big(w_{t-\tau_{\alpha_t}}\big)\big] - h(w_{t_m}) \big\| =& \textstyle\sum_{t=t_m}^{t_{m+1}-1} \alpha_t \big\| \mathbb{E}_{t_m}\big[h\big(w_{t-\tau_{\alpha_t}}\big) - h(w_{t_m})\big] \big\| \\
\leq& \textstyle\sum_{t=t_m}^{t_{m+1}-1} \alpha_t L_h' \mathbb{E}_{t_m}\big[\big\| w_{t-\tau_{\alpha_t}} - w_{t_m} \big\| \big]. \\
& && \text{(Lemma 30)}
\end{aligned}$$

When $t_m \geq t - \tau_{\alpha_t}$.

$$\begin{aligned}
&\textstyle\sum_{t=t_m}^{t_{m+1}-1} \alpha_t \left\| \mathbb{E}_{t_m}\left[h\left(w_{t-\tau_{\alpha_t}}\right)\right] - h(w_{t_m})\right\| \\
\leq &\textstyle\sum_{t=t_m}^{t_{m+1}-1} \alpha_t L_h' \mathbb{E}_{t_m}\left[\left\| w_{t-\tau_{\alpha_t}} - w_{t_m}\right\|\right] \\
\leq &\textstyle\sum_{t=t_m}^{t_{m+1}-1} \alpha_t L_h' \frac{\alpha_{t-\tau_{\alpha_t},t_m-1} L_h}{1-\alpha_{t-\tau_{\alpha_t},t_m-1} L_h}(\|w_{t_m}\|+1) && \text{(Lemma 29)} \\
\leq &2L_h L_h'(\|w_{t_m}\|+1) \cdot \textstyle\sum_{t=t_m}^{t_{m+1}-1} \alpha_t \max_{t\in[t_m,t_{m+1}-1]} \alpha_{t-\tau_{\alpha_t},t_m-1} \\
\leq &2L_h L_h'(\|w_{t_m}\|+1) \cdot 2T_m \cdot o(T_m) && \text{(By (11), Lemma 20 and (21))} \\
\leq &C_{24,3} T_m^2 (\|w_{t_m}\|+1),
\end{aligned}$$

where the second inequality holds because for sufficiently large t, we have $1-\alpha_{t-\tau_{\alpha_t},t_m-1} L_h > \frac{1}{2}$.

When $t_m < t - \tau_{\alpha_t}$.

$$\begin{aligned}
&\textstyle\sum_{t=t_m}^{t_{m+1}-1} \alpha_t \left\| \mathbb{E}_{t_m}\left[h\left(w_{t-\tau_{\alpha_t}}\right)\right] - h(w_{t_m})\right\| \\
\leq &\textstyle\sum_{t=t_m}^{t_{m+1}-1} \alpha_t L_h' \mathbb{E}_{t_m}\left[\left\| w_{t-\tau_{\alpha_t}} - w_{t_m}\right\|\right] \\
\leq &\textstyle\sum_{t=t_m}^{t_{m+1}-1} \alpha_t L_h' C_{28} \cdot 2T_m(\|w_{t_m}\|+1) && \text{(Lemma 28)} \\
\leq &2T_m L_h' C_{28} \cdot 2T_m(\|w_{t_m}\|+1) && \text{(By (11) and Lemma 20)} \\
\leq &T_m^2 C_{24,4}(\|w_{t_m}\|+1).
\end{aligned}$$

Combining all the bounds then completes the proof. ■

C.7 Proof of Lemma 25

Proof We first bound the last two terms in (13). For the first of the two, we have

$$\begin{aligned}
&\langle w_{t_m}, \mathbb{E}_{t_m}[s_{1,m} + s_{2,m} + s_{4,m}]\rangle \\
\leq &\|w_{t_m}\| \|\mathbb{E}_{t_m}[s_{1,m} + s_{2,m} + s_{4,m}]\| \\
\leq &\|w_{t_m}\| (\mathbb{E}_{t_m}[\|s_{1,m}\| + \|s_{4,m}\|] + \|\mathbb{E}_{t_m} s_{2,m}\|) \\
\leq &\|w_{t_m}\| T_m^2 (C_{21} + C_{24} + C_{22,1})(\|w_{t_m}\|+1) \\
= &\mathcal{O}\left(T_m^2(\|w_{t_m}\|^2+1)\right).
\end{aligned}$$

For the second of the two, we notice from Assumption 5.3 that $\|h(0)\| \leq L_h < L_h'$, Lemma 30 then gives

$$\|h(w_{t_m})\| \leq L_h'(\|w_{t_m}\|+1).$$

It then follows easily that

$$\|\bar{\alpha}_m h(w_{t_m})\| = \mathcal{O}(T_m(\|w_{t_m}\|+1)).$$

We then have

$$\|\bar{\alpha}_m h(w_{t_m}) + s_{1,m} + s_{2,m} + s_{3,m} + s_{4,m}\|^2 = \mathcal{O}\left(T_m^2(\|w_{t_m}\|^2+1)\right).$$

We are now ready to refine (13) as

$$\begin{aligned}
&\mathbb{E}_{t_m}\left[\left\|w_{t_{m+1}}\right\|^2\right] \\
\leq&(1-2\bar{\alpha}_m C_{5.4,1})\|w_{t_m}\|^2 + 2\bar{\alpha}_m C_{5.4,2} + \mathcal{O}\Big(T_m^2\Big(\|w_{t_m}\|^2+1\Big)\Big) \\
\leq&(1-T_m C_{5.4,1})\|w_{t_m}\|^2 + \mathcal{O}\left(T_m\right) + \mathcal{O}\Big(T_m^2(\|w_{t_m}\|^2+1)\Big).
\end{aligned}$$

Then for sufficiently large m, we have

$$\mathbb{E}_{t_m}\left[\left\|w_{t_{m+1}}\right\|^2\right] \leq (1-C_{25,1}T_m)\|w_{t_m}\|^2 + C_{25,2}T_m.$$

■

References

Jinane Abounadi, Dimitri Bertsekas, and Vivek S. Borkar. Learning algorithms for markov decision processes with average cost. *SIAM Journal on Control and Optimization*, 2001.

Sebastian Allmeier and Nicolas Gast. Computing the bias of constant-step stochastic approximation with Markovian noise. In *Advances in Neural Information Processing Systems*, 2024.

David Antrobius and Shangtong Zhang. On the divergence of differential temporal difference learning without local clocks. In *Advances in Neural Information Processing Systems*, 2026.

Leemon C. Baird. Residual algorithms: Reinforcement learning with function approximation. In *Proceedings of the International Conference on Machine Learning*, 1995.

Anas Barakat and Pascal Bianchi. Convergence and dynamical behavior of the ADAM algorithm for nonconvex stochastic optimization. *SIAM Journal on Optimization*, 2021.

Richard Bellman. A Markovian decision process. *Journal of Mathematics and Mechanics*, 1957.

Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. *Operations Research*, 2021.

Ethan Blaser, Jiuqi Wang, and Shangtong Zhang. Almost sure convergence of differential temporal difference learning for average reward Markov decision processes. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2026.

Vivek Borkar, Shuhang Chen, Adithya Devraj, Ioannis Kontoyiannis, and Sean Meyn. The ODE method for asymptotic statistics in stochastic approximation and reinforcement learning. *The Annals of Applied Probability*, 2025.

Vivek S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Hindustan Book Agency, 2008.

Qi Cai, Zhuoran Yang, Jason D. Lee, and Zhaoran Wang. Neural temporal difference and Q learning provably converge to global optima. *Mathematics of Operations Research*, 2024.

Diogo Carvalho, Francisco S. Melo, and Pedro Santos. A new convergent variant of Q-learning with linear function approximation. In *Advances in Neural Information Processing Systems*, 2020.

Siddharth Chandak and Vivek S. Borkar. A concentration bound for TD(0) with function approximation. *Stochastic Systems*, 2026.

Siddharth Chandak, Vivek S. Borkar, and Parth Dodhia. Concentration of contractive stochastic approximation and reinforcement learning. *Stochastic Systems*, 2022.

Zaiwei Chen, Siva Theja Maguluri, Sanjay Shakkottai, and Karthikeyan Shanmugam. Finite-sample analysis of contractive stochastic approximation using smooth convex envelopes. In *Advances in Neural Information Processing Systems*, 2020.

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*, 2022.

Zaiwei Chen, Siva Theja Maguluri, and Martin Zubeldia. Concentration of contractive stochastic approximation: Additive and multiplicative noise. *The Annals of Applied Probability*, 2025.

Edwin K.P. Chong, I-Jeng Wang, and Sanjeev R. Kulkarni. Noise conditions for prespecified convergence rates of stochastic approximation algorithms. *IEEE Transactions on Information Theory*, 1999.

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*, 2018a.

Gal Dalal, Balázs Szörényi, Gugan Thoppe, and Shie Mannor. Finite sample analysis of two-timescale stochastic approximation with applications to reinforcement learning. In *Proceedings of the Conference on Learning Theory*, 2018b.

Aymeric Dieuleveut and Francis Bach. Nonparametric stochastic approximation with large step-sizes. *The Annals of Statistics*, 2016.

Joseph L. Doob. *Stochastic Processes*. Wiley, 1953.

Alain Durmus, Eric Moulines, Alexey Naumov, Sergey Samsonov, Kevin Scaman, and Hoi-To Wai. Tight high probability bounds for linear stochastic approximation with fixed stepsize. In *Advances in Neural Information Processing Systems*, 2021.

Eyal Even-Dar and Yishay Mansour. Learning rates for Q-learning. *Journal of Machine Learning Research*, 2003.

Aditya Gopalan and Gugan Thoppe. Does DQN learn? *IEEE Transactions on Automatic Control*, 2026.

Geoffrey J. Gordon. Chattering in SARSA(λ). *Internal report, Carnegie Mellon University, CMU Learning Lab*, 1996.

Geoffrey J. Gordon. Reinforcement learning with function approximation converges to a region. In *Advances in Neural Information Processing Systems*, 2000.

Dongyan Huo, Yudong Chen, and Qiaomin Xie. Bias and extrapolation in Markovian linear stochastic approximation with constant step sizes. *Mathematics of Operations Research*, 2026.

Rajeeva Laxman Karandikar and Mathukumalli Vidyasagar. Convergence rates for stochastic approximation: Biased noise with unbounded variance, and applications. *Journal of Optimization Theory and Applications*, 2024.

Sajad Khodadadian and Martin Zubeldia. A general-purpose theorem for high-probability bounds of stochastic approximation with Polyak averaging. In *Advances in Neural Information Processing Systems*, 2025.

Vijaymohan Konda. *Actor-Critic Algorithms*. PhD thesis, Massachusetts Institute of Technology, 2002.

Nathaniel Korda and L. A. Prashanth. On TD(0) with function approximation: Concentration bounds and a centered variant with exponential convergence. In *Proceedings of the International Conference on Machine Learning*, 2015.

Michael A. Kouritzin and Samira Sadeghi. Convergence rates and decoupling in linear stochastic approximation algorithms. *SIAM Journal on Control and Optimization*, 2015.

Valery Koval and Rainer Schwabe. A law of the iterated logarithm for stochastic approximation procedures in d-dimensional Euclidean space. *Stochastic Processes and their Applications*, 2003.

Harold J. Kushner and G. George Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2003.

Caio Kalil Lauand and Sean Meyn. Revisiting step-size assumptions in stochastic approximation. *arXiv Preprint*, 2024.

Caio Kalil Lauand, Ioannis Kontoyiannis, and Sean Meyn. The case for and against fixed step-size: Stochastic approximation algorithms in optimization and machine learning. *arXiv Preprint*, 2025.

Donghwan Lee and Niao He. A unified switching system perspective and convergence analysis of Q-Learning algorithms. In *Advances in Neural Information Processing Systems*, 2020.

Gen Li, Changxiao Cai, Yuxin Chen, Yuting Wei, and Yuejie Chi. Is Q-learning minimax optimal? A tight sample complexity analysis. *Operations Research*, 2024.

Xiang Li, Wenhao Yang, Jiadong Liang, Zhihua Zhang, and Michael I. Jordan. A statistical analysis of Polyak-Ruppert averaged Q-learning. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2023.

Han-Dong Lim and Donghwan Lee. Regularized Q-learning. In *Advances in Neural Information Processing Systems*, 2024.

Long-Ji Lin. Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine Learning*, 1992.

Jun Liu and Ye Yuan. On almost sure convergence rates of stochastic gradient methods. In *Proceedings of the Conference on Learning Theory*, 2022.

Xinyu Liu, Zixuan Xie, and Shangtong Zhang. Linear Q-learning does not diverge in L^2: Convergence rates to a bounded set. In *Proceedings of the International Conference on Machine Learning*, 2025.

Francisco S. Melo, Sean P. Meyn, and M. Isabel Ribeiro. An analysis of reinforcement learning with function approximation. In *Proceedings of the International Conference on Machine Learning*, 2008.

Sean Meyn. The projected Bellman equation in reinforcement learning. *IEEE Transactions on Automatic Control*, 2024.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 2015.

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 the Conference on Learning Theory*, 2020.

Wenlong Mou, Koulik Khamaru, Martin J. Wainwright, Peter L. Bartlett, and Michael I. Jordan. Optimal variance-reduced stochastic approximation in Banach spaces. *arXiv Preprint*, 2022.

Wenlong Mou, Ashwin Pananjady, Martin J. Wainwright, and Peter L. Bartlett. Optimal and instance-dependent guarantees for Markovian linear stochastic approximation. *Mathematical Statistics and Learning*, 2024.

Eric Moulines and Francis R. Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, 2011.

Hoang Huy Nguyen, Quang Dinh Thien Nguyen, Duc Anh Nguyen, and Siva Theja Maguluri. Almost sure convergence of stochastic approximation: An interplay of noise and step size. *arXiv Preprint*, 2026.

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*, 2021.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, 1994.

Xiaochi Qian and Shangtong Zhang. Revisiting a design choice in gradient temporal difference learning. In *Proceedings of the International Conference on Learning Representations*, 2025.

Xiaochi Qian, Zixuan Xie, Xinyu Liu, and Shangtong Zhang. Almost sure convergence rates and concentration of stochastic approximation and reinforcement learning with Markovian noise. *Journal of Machine Learning Research*, 2026.

Guannan Qu and Adam Wierman. Finite-time analysis of asynchronous stochastic approximation and Q-learning. In *Proceedings of the Conference on Learning Theory*, 2020.

Herbert Robbins and David Siegmund. A convergence theorem for non-negative almost supermartingales and some applications. In *Optimizing Methods in Statistics*. 1971.

Gavin A. Rummery and Mahesan Niranjan. On-line Q-learning using connectionist systems. Technical report, Department of Engineering, University of Cambridge, 1994.

Sergey Samsonov, Eric Moulines, Qi-Man Shao, Zhuo-Song Zhang, and Alexey Naumov. Gaussian approximation and multiplier bootstrap for Polyak-Ruppert averaged linear stochastic approximation with applications to TD learning. In *Advances in Neural Information Processing Systems*, 2024.

Othmane Sebbouh, Robert M. Gower, and Aaron Defazio. Almost sure convergence rates for stochastic gradient descent and stochastic heavy ball. In *Proceedings of the Conference on Learning Theory*, 2021.

Rayadurgam Srikant and Lei Ying. Finite-time error bounds for linear stochastic approximation and TD learning. In *Proceedings of the Conference on Learning Theory*, 2019.

Richard S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 1988.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018.

Richard S. Sutton, David A. McAllester, Satinder P. Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, 1999.

Richard S. Sutton, Csaba Szepesvári, and Hamid Reza Maei. A convergent o(n) temporal-difference algorithm for off-policy learning with linear function approximation. In *Advances in Neural Information Processing Systems*, 2008.

Csaba Szepesvári. The asymptotic convergence-rate of Q-learning. In *Advances in Neural Information Processing Systems*, 1997.

Vladislav B. Tadić. On the almost sure rate of convergence of temporal-difference learning algorithms. In *Proceedings of the IFAC World Congress*, 2002.

Vladislav B. Tadić. On the almost sure rate of convergence of linear stochastic approximation algorithms. *IEEE Transactions on Information Theory*, 2004.

Gugan Thoppe and Vivek Borkar. A concentration bound for stochastic approximation via Alekseev's formula. *Stochastic Systems*, 2019.

Panos Toulis and Edoardo M. Airoldi. Asymptotic and finite-sample properties of estimators based on stochastic gradients. *The Annals of Statistics*, 2017.

John N. Tsitsiklis and Benjamin Van Roy. An analysis of temporal-diffference learning with function approximation. In *IEEE Transactions on Automatic Control*, 1997.

Mathukumalli Vidyasagar. Convergence of stochastic approximation via martingale and converse Lyapunov methods. *Mathematics of Control, Signals, and Systems*, 2023.

Jiuqi Wang and Shangtong Zhang. Almost sure convergence of linear temporal difference learning with arbitrary features. *Journal of Machine Learning Research*, 2026.

Christopher J. C. H. Watkins and Peter Dayan. Technical note: Q-learning. *Machine Learning*, 1992.

Christopher John Cornish Hellaby Watkins. *Learning from Delayed Rewards*. PhD thesis, University of Cambridge, 1989.

Zixuan Xie, Xinyu Liu, Rohan Chandra, and Shangtong Zhang. Finite sample analysis of linear temporal difference learning with arbitrary features. In *Advances in Neural Information Processing Systems*, 2025.

Ran Xin, Usman A. Khan, and Soummya Kar. An improved convergence analysis for decentralized online stochastic non-convex optimization. *IEEE Transactions on Signal Processing*, 2021.

Pan Xu and Quanquan Gu. A finite-time analysis of Q-Learning with neural network function approximation. In *Proceedings of the International Conference on Machine Learning*, 2020.

Kaiqing Zhang, Alec Koppel, Hao Zhu, and Tamer Başar. Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM Journal on Control and Optimization*, 2020.

Shangtong Zhang. *Breaking the Deadly Triad in Reinforcement Learning*. PhD thesis, University of Oxford, 2022.

Shangtong Zhang. Towards formalizing reinforcement learning theory: A Robbins-Siegmund approach. *Reinforcement Learning Journal*, 2026.

Shangtong Zhang and Shimon Whiteson. Truncated emphatic temporal difference methods for prediction and control. *Journal of Machine Learning Research*, 2022.

Shangtong Zhang, Yi Wan, Richard S. Sutton, and Shimon Whiteson. Average-reward off-policy policy evaluation with function approximation. In *Proceedings of the International Conference on Machine Learning*, 2021a.

Shangtong Zhang, Hengshuai Yao, and Shimon Whiteson. Breaking the deadly triad with a target network. In *Proceedings of the International Conference on Machine Learning*, 2021b.

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

Shangtong Zhang, Remi Tachet des Combes, and Romain Laroche. On the convergence of SARSA with linear function approximation. In *Proceedings of the International Conference on Machine Learning*, 2023.

Yixuan Zhang and Qiaomin Xie. Constant stepsize Q-learning: Distributional convergence, bias and extrapolation. *Reinforcement Learning Journal*, 2024.

Shaofeng Zou, Tengyu Xu, and Yingbin Liang. Finite-sample analysis for SARSA with linear function approximation. In *Advances in Neural Information Processing Systems*, 2019.