Asymptotic and Finite Sample Analysis of Nonexpansive Stochastic Approximations with Markovian Noise

Ethan Blaser[1], **Shangtong Zhang**[1]

[1] Department of Computer Science, University of Virginia
blaser@email.virginia.edu, shangtong@virginia.edu

Abstract

Stochastic approximation is a powerful class of algorithms with celebrated success. However, a large body of previous analysis focuses on stochastic approximations driven by contractive operators, which is not applicable in some important reinforcement learning settings like the average reward setting. This work instead investigates stochastic approximations with merely nonexpansive operators. In particular, we study nonexpansive stochastic approximations with Markovian noise, providing both asymptotic and finite sample analysis. Key to our analysis are novel bounds of noise terms resulting from the Poisson equation. As an application, we prove for the first time that classical tabular average reward temporal difference learning converges to a sample-path dependent fixed point.

Extended version — https://arxiv.org/abs/2409.19546

1 Introduction

Stochastic approximation (SA) algorithms (Robbins and Monro 1951; Kushner and Yin 2003; Borkar 2009) form the foundation of many iterative optimization and learning methods by updating a vector incrementally and stochastically. Prominent examples include stochastic gradient descent (Kiefer and Wolfowitz 1952) and temporal difference (TD) learning (Sutton 1988). These algorithms generate a sequence of iterates $\{x_n\}$ starting from an initial point $x_0 \in \mathbb{R}^d$ through the recursive update:

$$x_{n+1} \doteq x_n + \alpha_{n+1}(H(x_n, Y_{n+1}) - x_n) \tag{SA}$$

where $\{\alpha_n\}$ is a sequence of learning rates, $\{Y_n\}$ is a sequence of random noise in a space $\mathcal{Y}$, and a function $H : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}^d$ maps the current iterate x_n and noise Y_{n+1} to the actual incremental update. We use h to denote the expected update, i.e., $h(x) \doteq \mathbb{E}[H(x, y)]$, where the expectation will be formally defined shortly.

Despite the foundational role of SA in analyzing reinforcement learning (RL, Sutton and Barto (2018)) algorithms, most of the existing literature assumes that the expected mapping h is a contraction. However, in many problems in RL, particularly those involving average reward formulations (Tsitsiklis and Roy 1999; Puterman 2014; Wan, Naik, and Sutton 2021b,a; He, Wan, and Mahmood 2022), h is only guaranteed to be non-expansive, not contractive. Table 1 highlights the relative scarcity of results concerning nonexpansive mappings. As a result, it is surprising that the convergence of some of the simplest and most fundamental RL algorithms, such as tabular average reward TD (Tsitsiklis and Roy 1999), has not been fully settled, despite more than 25 years having passed since its introduction.

One tool for analyzing (SA) with nonexpansive h, which has recently gained renewed attention, is Krasnoselskii-Mann (KM) iterations:

$$x_{n+1} = x_n + \alpha_{n+1}(h(x_n) - x_n). \tag{KM}$$

Under some other restrictive conditions, Krasnosel'skii (1955) first proves the convergence of (KM) to a fixed point of h and this result is further generalized by Edelstein (1966); Ishikawa (1976); Reich (1979); Liu (1995). More recently, Cominetti, Soto, and Vaisman (2014) use a novel fox-and-hare model to connect KM iterations with Bernoulli random variables, providing a sharper convergence rate for $\|x_n - h(x_n)\| \to 0$. Kim and Xu (2007); Cominetti, Soto, and Vaisman (2014); Bravo, Cominetti, and Pavez-Signé (2019) further consider (KM) with some deterministic additive noise.

However, practitioners usually do not have access to h directly. Instead, they only have access to a noisy estimate of h (cf. H in (SA)). As a result, the general SA update (SA) is also called the Stochastic KM (SKM) iterations when h is nonexpansive. Under mild conditions, Bravo and Cominetti (2024) prove the almost sure convergence of SKM, together with the convergence rates of $\mathbb{E}[\|x_n - h(x_n)\|]$. However, one significant limitation of Bravo and Cominetti (2024) is that they assume $\{Y_t\}$ are i.i.d., which significantly restricts their applications in RL because the corresponding $\{Y_t\}$ in many RL algorithms (e.g., the aforementioned tabular average reward TD) is a Markov chain. This is the second gap that this work shall close.

To summarize, we make two contributions in this work to close the two gaps. **First**, Theorem 2.6 proves that the sequence $\{x_n\}$ generated by (SA) with Markovian $\{Y_n\}$ and nonexpansive h, converges almost surely to some random point $x_* \in \mathcal{X}_*$, where $\mathcal{X}_*$ is the set of fixed points of h. Importantly, x_* may depend on the entire sample-path. Theorem 3.1 further provides the convergence rate

	Nonexpansive h	Markovian $\{Y_n\}$	Asymptotic	Non-Asymptotic
Krasnosel'skii (1955)	✓		✓	
Ishikawa (1976)	✓		✓	
Reich (1979)	✓		✓	
Benveniste, Métivier, and Priouret (1990)			✓	
Liu (1995)			✓	
Szepesvári (1997)			✓	
Abounadi, Bertsekas, and Borkar (2002)	✓		✓	
Tadic (2002)		✓		✓
Kushner and Yin (2003)			✓	
Koval and Schwabe (2003)			✓	✓
Tadic (2004)		✓		✓
Kim and Xu (2007)	✓		✓	
Borkar (2009)			✓	
Cominetti, Soto, and Vaisman (2014)	✓		✓	✓
Bravo, Cominetti, and Pavez-Signé (2019)	✓		✓	✓
Chen et al. (2021)		✓		✓
Borkar et al. (2021)		✓	✓	✓
Karandikar and Vidyasagar (2024)		✓	✓	✓
Bravo and Cominetti (2024)	✓		✓	✓
Qian et al. (2024)		✓	✓	✓
Liu, Chen, and Zhang (2025)		✓	✓	
Ours	✓	✓	✓	✓

Table 1: Overview of stochastic approximation methods, with a focus on those that consider non-expansive mappings. "Non-expansive h" refers to works where the expected mapping is non-expansive, as opposed to strictly a contraction. "Markovian $\{Y_n\}$" indicates cases where the noise term $\{Y_n\}$ is Markovian. "Asymptotic" refers to works that prove almost sure convergence, which is not necessarily weaker than non-asymptotic convergence results. Note that we present only a representative subset of results for SA with contractive mappings due to an abundance of literature in the area. For a more comprehensive treatment, see Benveniste, Métivier, and Priouret (1990); Kushner and Yin (2003); Borkar (2009).

of the expected residuals $\mathbb{E}[\|x_n - h(x_n)\|]$. Both only assume $\{Y_t\}$ is a Markov chain. Table 1 highlights the improvement of this work over those prior. The key idea of our approach is to use Poisson's equation to decompose the error $\{H(x_n, Y_{n+1}) - h(x_n)\}$ into boundable error terms (Benveniste, Métivier, and Priouret 1990). While Poisson's equation has been previously used for handling Markovian noise, our method departs from prior arts in how we bound the resulting error terms. Specifically, Benveniste, Métivier, and Priouret (1990) and Konda and Tsitsiklis (1999) use stopping times, while Borkar et al. (2021) employ a Lyapunov function and use the scaled iterates technique. By contrast, we leverage a 1-Lipschitz continuity assumption on H to directly control the growth of error terms. **Second**, Theorem 4.2 uses our novel SKM results to provide the first proof of almost sure convergence of tabular average reward TD to a possibly sample-path dependent fixed point.

Notations In this paper, all vectors are column. We use $\|\cdot\|$ to denote a generic operator norm. We use $\|\cdot\|_2$ and $\|\cdot\|_\infty$ to denote ℓ_2 norm and infinity norm respectively. We use $\mathcal{O}(\cdot)$ to hide deterministic constants for simplifying presentation, while the letter ζ is reserved for sample-path dependent constants.

2 Asymptotic Analysis of SKM Iterations

To broaden the applicability of our result, we future allow (SA) to have additional additive noise. Namely, we consider the following SKM updates

$$x_{n+1} = x_n + \alpha_{n+1}\left(H(x_n, Y_{n+1}) - x_n + \epsilon^{(1)}_{n+1}\right), \quad \text{(SKM)}$$

where $\{x_n\}$ are stochastic vectors evolving in $\mathbb{R}^d$, $\{Y_n\}$ is a Markov chain evolving in a finite state space $\mathcal{Y}$, $H : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}^d$ defines the update, $\left\{\epsilon^{(1)}_{n+1}\right\}$ is a sequence of stochastic noise evolving in $\mathbb{R}^d$, and $\{\alpha_n\}$ is a sequence of deterministic learning rates. We make the following assumptions.

Assumption 2.1 (Ergodicity)**.** The Markov chain $\{Y_n\}$ is irreducible and aperiodic.

The Markov chain $\{Y_n\}$ thus adopts a unique invariant distribution, denoted d_μ. We use P to denote the transition matrix of $\{Y_n\}$.

Assumption 2.2 (1-Lipschitz)**.** The function H is 1-Lipschitz continuous in its first argument w.r.t. some operator norm $\|\cdot\|$ and uniformly in its second argument, i.e., for any x, x', y, it holds that

$$\|H(x, y) - H(x', y)\| \leq \|x - x'\|.$$

This assumption has two important implications. First, it implies that $H(x,y)$ can grow at most linearly. Indeed, let $x'=0$, we get $\|H(x,y)\| \leq \|H(0,y)\| + \|x\|$. Define $C_H \doteq \max_y \|H(0,y)\|$, we get

$$\|H(x,y)\| \leq C_H + \|x\|. \tag{1}$$

Second, define the function $h:\mathbb{R}^d \to \mathbb{R}^d$ as the expectation of H over the stationary distribution d_μ:

$$h(x) \doteq \mathbb{E}_{y\sim d_\mu}[H(x,y)].$$

We then have that h is non-expansive. Namely,

$$\begin{aligned}\|h(x)-h(x')\| &\leq \textstyle\sum_y d_\mu(y)\|H(x,y)-H(x',y)\| \\ &\leq \|x-x'\|. \end{aligned}\tag{2}$$

We need to assume that the problem is solvable.

Assumption 2.3 (Fixed Points)**.** The non-expansive operator h adopts at least one fixed point.

We use $\mathcal{X}_* \neq \emptyset$ to denote the set of fixed points of h.

Assumption 2.4 (Learning Rate)**.** The learning rate $\{\alpha_n\}$ has the form

$$\alpha_n = \frac{1}{(n+1)^b}, \alpha_0 = 0,$$

where $b \in (\frac{4}{5}, 1]$.

The primary motivation for requiring $b \in (\frac{4}{5}, 1]$ is that our learning rates α_n need to decrease quickly enough for certain key terms in the proof to be finite. The specific need for $b > \frac{4}{5}$ can be seen in the proof of (30) in Lemma B.1. We now impose assumptions on the additive noise.

Assumption 2.5 (Additive Noise)**.**

$$\sum_{k=1}^{\infty} \alpha_k \left\| \epsilon_k^{(1)} \right\| < \infty \quad \text{a.s.}, \tag{3}$$

$$\mathbb{E}\left[\left\| \epsilon_n^{(1)} \right\|^2 \right] = \mathcal{O}(1/n). \tag{4}$$

The first part of Assumption 2.5 can be interpreted as a requirement that the total amount of additive noise remains finite. Additionally, we impose a condition on the second moment of this noise, requiring it to converge at the rate $\mathcal{O}\left(\frac{1}{n}\right)$. While these assumptions on $\epsilon_n^{(1)}$ may seem restrictive, it should be noted that even if $\epsilon_n^{(1)}$ were absent, our work would still extend the results of Bravo and Cominetti (2024) to cases involving Markovian noise, as the Markovian noise component is already incorporated in Y_n, which represents a significant result. For most RL applications involving algorithms which have only one set of learnable weights, the additional noise $\epsilon_k^{(1)}$ will simply be 0. We are now ready to present the asymptotic analysis of (SKM).

Theorem 2.6. *Let Assumptions 2.1 - 2.5 hold. Then the iterates $\{x_n\}$ generated by*
(SKM) *satisfy*

$$\lim_{n\to\infty} x_n = x_* \quad \textit{a.s.},$$

where $x_* \in \mathcal{X}_*$ is a possibly sample-path dependent fixed point. Or more precisely speaking, let ω denote a sample path $(w_0, Y_0, Y_1, \dots)$ and write $x_n(\omega)$ to emphasize the dependence of x_n on ω. Then there exists a set Ω of sample paths with $\Pr(\Omega) = 1$ such that for any $\omega \in \Omega$, the limit $\lim_{n\to\infty} x_n(\omega)$ exists, denoted as $x_*(\omega)$, and satisfies $x_*(\omega) \in \mathcal{X}_*$.

Proof. We first define two useful shorthands,

$$\alpha_{k,n} \doteq \alpha_k \prod_{j=k+1}^{n} (1-\alpha_j),\ \alpha_{n,n} \doteq \alpha_n, \tag{5}$$

$$\tau_n \doteq \sum_{k=1}^{n} \alpha_k(1-\alpha_k). \tag{6}$$

We then start with a decomposition of the error $H(x,Y_{n+1}) - h(x)$ using Poisson's equation akin to Métivier and Priouret (1987); Benveniste, Métivier, and Priouret (1990). Namely, thanks to the finiteness of $\mathcal{Y}$, it is well known (see, e.g., Theorem 17.4.2 of Meyn and Tweedie (2012) or Theorem 8.2.6 of Puterman (2014)) that there exists a function $\nu(x,y): \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}^d$ such that

$$H(x,y) - h(x) = \nu(x,y) - (P\nu)(x,y). \tag{7}$$

Here, we use $P\nu$ to denote the function $(x,y) \mapsto \sum_{y'} P(y,y')\nu(x,y')$. The error can then be decomposed as

$$H(x,Y_{n+1}) - h(x) = M_{n+1} + \epsilon_{n+1}^{(2)} + \epsilon_{n+1}^{(3)}, \tag{8}$$

where

$$M_{n+1} \doteq \nu(x_n, Y_{n+2}) - (P\nu)(x_n, Y_{n+1}), \tag{9}$$

$$\epsilon_{n+1}^{(2)} \doteq \nu(x_n, Y_{n+1}) - \nu(x_{n+1}, Y_{n+2}), \tag{10}$$

$$\epsilon_{n+1}^{(3)} \doteq \nu(x_{n+1}, Y_{n+2}) - \nu(x_n, Y_{n+2}). \tag{11}$$

Here $\{M_{n+1}\}$ is a Martingale difference sequence. We then use

$$\xi_{n+1} \doteq \epsilon_{n+1}^{(1)} + \epsilon_{n+1}^{(2)} + \epsilon_{n+1}^{(3)}, \tag{12}$$

to denote all the non-Martingale noise, yielding

$$x_{n+1} = (1-\alpha_{n+1})x_n + \alpha_{n+1}(h(x_n) + M_{n+1} + \xi_{n+1}).$$

We now define an auxiliary sequence $\{U_n\}$ to capture how the noise evolves

$$U_{n+1} \doteq (1-\alpha_{n+1})U_n + \alpha_{n+1}(M_{n+1} + \xi_{n+1}),\ U_0 \doteq 0. \tag{13}$$

If we can prove that the total noise is well controlled in the following sense

$$\sum_{k=1}^{\infty} \alpha_k \|U_{k-1}\| < \infty \quad \text{a.s.}, \tag{14}$$

$$\lim_{n\to\infty} \|U_n\| = 0 \quad \text{a.s.}, \tag{15}$$

then a result from Bravo and Cominetti (2024) can be applied on each sample path to complete the almost sure convergence proof. The remainder of the proof is dedicated to the verification of these two conditions.

Telescoping (13) yields

$$U_n = \underbrace{\sum_{k=1}^{n} \alpha_{k,n} M_k}_{\overline{M}_n} + \underbrace{\sum_{k=1}^{n} \alpha_{k,n} \epsilon_k^{(1)}}_{\overline{\epsilon}_n^{(1)}} + \underbrace{\sum_{k=1}^{n} \alpha_{k,n} \epsilon_k^{(2)}}_{\overline{\epsilon}_n^{(2)}} + \underbrace{\sum_{k=1}^{n} \alpha_{k,n} \epsilon_k^{(3)}}_{\overline{\epsilon}_n^{(3)}}. \tag{16}$$

Then, we can upper-bound (14) as

$$\sum_{k=1}^{n} \alpha_k \|U_{k-1}\| \leq \underbrace{\sum_{k=1}^{n} \alpha_k \left\|\overline{M}_{k-1}\right\|}_{\overline{\overline{M}}_n} + \underbrace{\sum_{k=1}^{n} \alpha_k \left\|\overline{\epsilon}_{k-1}^{(1)}\right\|}_{\overline{\overline{\epsilon}}_n^{(1)}} + \underbrace{\sum_{k=1}^{n} \alpha_k \left\|\overline{\epsilon}_{k-1}^{(2)}\right\|}_{\overline{\overline{\epsilon}}_n^{(2)}} + \underbrace{\sum_{k=1}^{n} \alpha_k \left\|\overline{\epsilon}_{k-1}^{(3)}\right\|}_{\overline{\overline{\epsilon}}_n^{(3)}}. \tag{17}$$

Here we bound only $\overline{\overline{\epsilon}}_n^{(2)}$ to demonstrate the novelty of our approach to handling these error terms. The almost sure bounds for $\overline{\overline{M}}_n, \overline{\overline{\epsilon}}_n^{(1)}$, and $\overline{\overline{\epsilon}}_n^{(3)}$ are provided in Lemmas B.8, B.9, and B.10 respectively. Starting with the definition of $\overline{\epsilon}_n^{(2)}$ from (16), and substituting the definition of $\epsilon_n^{(2)}$ from (10) we have,

$$\begin{aligned}
&\overline{\epsilon}_n^{(2)} \\
&= -\sum_{k=1}^{n} \alpha_{k,n} (\nu(x_k, Y_{k+1}) - \nu(x_{k-1}, Y_k)), \\
&= -\sum_{k=1}^{n} \alpha_{k,n} \nu(x_k, Y_{k+1}) - \alpha_{k-1,n} \nu(x_{k-1}, Y_k) \\
&\quad + \alpha_{k-1,n} \nu(x_{k-1}, Y_k) - \alpha_{k,n} \nu(x_{k-1}, Y_k), \\
&= -\alpha_n \nu(x_n, Y_{n+1}) - \sum_{k=1}^{n} (\alpha_{k-1,n} - \alpha_{k,n}) \nu(x_{k-1}, Y_k),
\end{aligned}$$

where the last equality holds because $\alpha_0 \doteq 0$ and $\alpha_{n,n} = \alpha_n$. Taking the norm gives

$$\begin{aligned}
\left\|\overline{\epsilon}_n^{(2)}\right\| &\leq \alpha_n \|\nu(x_n, Y_{n+1})\| \\
&\quad + \sum_{k=1}^{n} |\alpha_{k-1,n} - \alpha_{k,n}| \|\nu(x_{k-1}, Y_k)\|, \qquad (18) \\
&\leq \zeta_{B.5} (\alpha_n \tau_n + \sum_{k=1}^{n} |\alpha_{k-1,n} - \alpha_{k,n}| \, \tau_{k-1}), \\
&\leq 2\zeta_{B.5} \alpha_n \tau_n,
\end{aligned}$$

where the second inequality holds by Lemma B.5 with $\zeta_{B.5}$ denoting a sample-path dependent constant defined in Lemma B.5, and the last inequality holds because $\alpha_0 \doteq 0$, and that $\alpha_{i,n}$ and τ_i are monotonically increasing (Lemma A.2).

Then, from the definition of $\overline{\overline{\epsilon}}_n^{(2)}$ in (14), we have

$$\overline{\overline{\epsilon}}_n^{(2)} = \sum_{k=1}^{n} \alpha_k \left\|\overline{\epsilon}_{k-1}^{(2)}\right\| \leq 2\zeta_{B.5} \sum_{k=1}^{n} \alpha_k^2 \tau_k,$$

where the inequality holds because $\alpha_0 \doteq 0$ and α_k is decreasing. Then, by Lemma B.1, we have $\sup_n \sum_{k=1}^{n} \alpha_k^2 \tau_k < \infty$, which when combined with the monotone convergence theorem proves that $\lim_{n\to\infty} \overline{\overline{\epsilon}}_n^{(2)} < \infty$, verifying (14).

We now verify (15). This time, rewrite U_n as

$$U_n = -\sum_{k=1}^{n} \alpha_k U_{k-1} + \alpha_k \Big(M_k + \epsilon_k^{(1)} + \epsilon_k^{(2)} + \epsilon_k^{(3)} \Big).$$

Lemma B.11, Assumption 2.5, and Lemmas B.12, B.13 prove that $\sup_n \|\sum_{k=1}^{n} \alpha_k M_k\| < \infty$ and $\sup_n \left\|\sum_{k=1}^{n} \alpha_k \epsilon_k^{(j)}\right\| < \infty$ for $j \in \{1, 2, 3\}$ respectively.

Together with (16), this means that $\sup_n \|U_n\| < \infty$. In other words, we have established the stability of (13). Then, it can be shown (Lemma B.14), using an extension of Theorem 2.1 of Borkar (2009) (Lemma D.7), that $\{U_n\}$ converges to the globally asymptotically stable equilibrium of the ODE $\frac{\mathrm{d}U(t)}{\mathrm{d}t} = -U(t)$, which is 0. This verifies (15). Lemma B.15 then invokes a result from Bravo and Cominetti (2024) and completes the proof. □

Remark 2.7. We want to highlight that the technical novelty of our work comes from two sources. The first is that while the use of Poisson's equation for handling Markovian noise is well-established, including the noise representation in (8), previous works with such error decomposition (e.g., Benveniste, Métivier, and Priouret (1990); Konda and Tsitsiklis (1999); Borkar et al. (2025)) usually only need to bound terms like $\sum_k \alpha_k \epsilon_k^{(1)}$. In contrast, our setup requires the bounding of additional terms such as $\overline{\epsilon}_n^{(1)} = \sum_k \alpha_{k,n} \epsilon_k^{(1)}$ and $\overline{\overline{\epsilon}}_n^{(1)} = \sum_i \alpha_i \left\|\overline{\epsilon}_{k-1}^{(1)}\right\|$ that appear novel and more challenging. Specifically, Benveniste, Métivier, and Priouret (1990); Konda and Tsitsiklis (1999) consider the stopping time when $\|x_n\|$ first exceeds some threshold. Borkar et al. (2021) develop a contractive and recursive bound for $\|\nu(x_k, Y_{k+1})\|$. Both are highly complicated and do not apply to our problem of bounding $\overline{\overline{\epsilon}}^{(1)}$. We instead leverage the 1-Lipschitzness of H and use the sample-path dependent direct bound (cf. Lemma B.5) for $\|\nu(x_k, Y_{k+1})\|$. Second, our work extends Theorem 2.1 of Borkar (2009) by relaxing an assumption on the convergence of the deterministic noise term. Instead of requiring the noise to converge to 0, we only require a more mild condition on the asymptotic rate of change of this noise term. This extension, detailed in Appendix D, has independent utility beyond this work.

3 Finite Sample Analysis of SKM Iterations

The previous analysis not only guarantees the almost sure convergence of the iterates, but can also be used to obtain estimates of the expected fixed-point residuals.

Theorem 3.1. *Consider the iteration (SKM) and let Assumptions 2.1 − 2.5 hold. There exists a constant $C_{3.1}$ such that*

$$\mathbb{E}\left[\|x_n - h(x_n)\|\right] \leq \frac{C_{3.1}}{\sqrt{\tau_n}} = \begin{cases} \mathcal{O}\big(1/\sqrt{n^{1-b}}\big) \text{ if } \frac{4}{5} < b < 1, \\ \mathcal{O}\big(1/\sqrt{\log n}\big) \text{ if } b = 1. \end{cases}$$

Proof. Considering the sequence $z_n \doteq x_n - U_n$ we have,

$$\begin{aligned} \|x_n - h(x_n)\| &\leq \|z_n - h(z_n)\| + 2\|z_n - x_n\|, \\ &= \|z_n - h(z_n)\| + 2\|U_n\|. \end{aligned}$$

where the inequality holds due to the non-expansivity of h as proven in (2). Then, our proof of Theorem 2.6 guarantees the conditions under which the $\{z_n\}$ is bounded. Specifically, we proved in Lemma B.15 that if $\sum_{n=1}^{\infty} \alpha_k \|U_{k-1}\| < \infty$ (14) and $\|U_n\| \to 0$ (15) almost surely, then a result from Bravo and Cominetti (2024) (included as Lemma A.1 for completeness) can be invoked to bound $\|z_n - h(z_n)\|$. Specifically, by identifying $e_k = U_{k-1}$ in Lemma A.1, we get

$$\begin{aligned} &\|x_n - h(x_n)\| \\ &\leq \zeta_{A.1}\sigma(\tau_n) + \sum_{k=2}^{n} 2\alpha_k \sigma(\tau_n - \tau_k)\|U_{k-1}\| + 4\|U_n\|. \end{aligned}$$

for $\zeta_{A.1} = 2\text{dist}(x_0, \mathcal{X}_*) + \sum_{k=2}^{\infty} \alpha_k \|U_{k-1}\|$. However, $\zeta_{A.1}$ is a sample-path dependent constant whose order is unknown, and the random sequence $\|U_n\|$ may occasionally become very large. Therefore, we compute the non-asymptotic error bound of the expected residuals $\mathbb{E}\left[\|x_n - h(x_n)\|\right]$, which gives,

$$\begin{aligned} \mathbb{E}[\|x_n - h(x_n)\|] \leq \underbrace{\mathbb{E}[\zeta_{A.1}]\sigma(\tau_n)}_{R_1} \\ + \underbrace{\sum_{k=2}^{n} 2\alpha_k \sigma(\tau_n - \tau_k)\mathbb{E}[\|U_{k-1}\|]}_{R_2} + \underbrace{4\mathbb{E}[\|U_n\|]}_{R_3}. \end{aligned}$$

Recalling that $\sigma(y) \doteq \min\left\{1, 1/\sqrt{\pi y}\right\}$, we can see that if there exists a deterministic constant $C_{3.1}$ such that $\mathbb{E}[\zeta_{A.1}] \leq C_{3.1}$, we obtain that $R_1 = \mathcal{O}\big(1/\sqrt{\tau_n}\big)$. Therefore, in order to prove the Theorem, it is sufficient to find such a constant $C_{3.1}$ such that $\mathbb{E}[\zeta_{A.1}] \leq C_{3.1}$, and prove that R_2, and R_3 are also $\mathcal{O}\big(1/\sqrt{\tau_n}\big)$.

We proceed by first upper-bounding R_3, i.e., $\mathbb{E}[\|U_n\|]$. Taking the expectation of (16), we have,

$$\begin{aligned} &\mathbb{E}[\|U_n\|] \\ \leq &\mathbb{E}\left[\left\|\overline{M}_n\right\|\right] + \mathbb{E}\left[\left\|\overline{\epsilon}_n^{(1)}\right\|\right] + \mathbb{E}\left[\left\|\overline{\epsilon}_n^{(2)}\right\|\right] + \mathbb{E}\left[\left\|\overline{\epsilon}_n^{(3)}\right\|\right] \\ \leq &C_{C.1}\tau_n\sqrt{\alpha_{n+1}} + \sum_{i=1}^{n} \alpha_{i,n}\mathbb{E}\left[\left\|\epsilon_i^{(1)}\right\|\right] + C_{C.2}\alpha_n\tau_n \\ &+ C_{C.3}\alpha_n \sum_{i=1}^{n} \alpha_i\tau_i \quad \text{(Corollaries C.1, C.2, C.3)} \\ \doteq &\omega_n \end{aligned} \tag{19}$$

It can be shown (Lemma C.4) that $\omega_n = \mathcal{O}(\tau_n\sqrt{\alpha_{n+1}})$, which is dominated by $1/\sqrt{\tau_n}$.

For R_2, Lemma C.5 proves, similarly to Theorems 2.11 and 3.1 of Bravo and Cominetti (2024), that $R_2 = \mathcal{O}\big(1/\sqrt{\tau_n}\big)$.

For R_1. We first observe that

$$\sum_{k=2}^{\infty} \alpha_k \mathbb{E}[\|U_{k-1}\|] \leq \sum_{k=2}^{\infty} \alpha_k \omega_{k-1} = \mathcal{O}\left(\sum_{k=2}^{\infty} \alpha_k^{3/2}\tau_{k-1}\right),$$

which is finite by Lemma B.1. It is then obvious to see that there exists a $C_{3.1}$ such that $\mathbb{E}[\zeta_{A.1}] = 2\text{dist}(x_0, \mathcal{X}_*) + \sum_{k=2}^{\infty} \alpha_k \mathbb{E}[\|U_{k-1}\|] \leq C_{3.1}$, which completes the proof. □

Remark 3.2. While the convergence rate is relatively slow, especially compared to the discounted setting (e.g., Chen et al. (2021)), it matches the rate in the i.i.d. noise case for nonexpansive operators (Bravo and Cominetti 2024). This slow rate is inherent due to the nonexpansive nature of h (Cominetti, Soto, and Vaisman 2014) and is not a limitation of our analysis.

4 Application in Average Reward Temporal Difference Learning

In this section, we provide the first proof of almost sure convergence to a fixed point for average reward TD in its simplest tabular form. Remarkably, this convergence result has remained unproven for over 25 years despite the algorithm's fundamental importance and simplicity.

4.1 Reinforcement Learning Background

In reinforcement learning (RL), we consider a Markov Decision Process (MDP; Bellman (1957); Puterman (2014)) 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 $p_0 : \mathcal{S} \to [0, 1]$. At time step 0, an initial state S_0 is sampled from p_0. At time t, given the state S_t, the agent samples an action $A_t \sim \pi(\cdot|S_t)$, where $\pi : \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the policy being followed by the agent. A reward $R_{t+1} \doteq r(S_t, A_t)$ is then emitted and the agent proceeds to a successor state $S_{t+1} \sim p(\cdot|S_t, A_t)$. In the rest of the paper, we will assume the Markov chain $\{S_t\}$ induced by the policy π is irreducible and thus adopts a unique stationary distribution d_μ. The average reward (a.k.a. gain, Puterman (2014)) is defined as $\bar{J}_\pi \doteq \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[R_t\right]$. Correspondingly, the differential value function (a.k.a. bias, Puterman (2014)) is defined as

$$v_\pi(s) \doteq \lim_{T\to\infty} \frac{1}{T}\sum_{\tau=1}^{T} \mathbb{E}\left[\sum_{i=1}^{\tau} (R_{t+i} - \bar{J}_\pi) \mid S_t = s\right].$$

The corresponding Bellman equation (a.k.a. Poisson's equation) is then

$$v = r_\pi - \bar{J}_\pi e + P_\pi v, \tag{20}$$

where $v \in \mathbb{R}^{|\mathcal{S}|}$ is the free variable, e denotes an all-one vector, $r_\pi \in \mathbb{R}^{|\mathcal{S}|}$ is the reward vector induced by the policy

π, i.e., $r_\pi(s) \doteq \sum_a \pi(a|s) r(s,a)$, and $P_\pi \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ is the transition matrix induced by the policy π, i.e., $P_\pi(s,s') \doteq \pi(a|s)p(s'|s,a)$. It is known (Puterman 2014) that all solutions to (20) form a set

$$\mathcal{V}_* \doteq \{v_\pi + ce \mid c \in \mathbb{R}\}. \tag{21}$$

The policy evaluation problem in average reward MDPs is to estimate v_π, perhaps up to a constant offset ce.

4.2 Average Reward Temporal Difference Learning

Temporal Difference learning (TD; Sutton (1988)) is a foundational algorithm in RL (Sutton and Barto 2018). Inspired by its success in the discounted setting, Tsitsiklis and Roy (1999) proposed using the update rule (Average Reward TD) to estimate v_π (up to a constant offset) for average reward MDPs. The updates are given by:

$$J_{t+1} = J_t + \beta_{t+1}(R_{t+1} - J_t), \quad \text{(Average Reward TD)}$$
$$v_{t+1}(S_t) = v_t(S_t) + \alpha_{t+1}(R_{t+1} - J_t + v_t(S_{t+1}) - v_t(S_t)),$$

where $\{S_0, R_1, S_1, \dots\}$ is a trajectory of states and rewards from an MDP under a fixed policy in a finite state space $\mathcal{S}$, $J_t \in \mathbb{R}$ is the scalar estimate of the average reward $\bar{J}_\pi$, $v_t \in \mathbb{R}^{|\mathcal{S}|}$ is the tabular value estimate, and $\{\alpha_t, \beta_t\}$ are learning rates.

To utilize Theorem 2.6 to prove the almost sure convergence of (Average Reward TD), we first rewrite it in a compact form to match that of (SKM). Define the augmented Markov chain $Y_{t+1} \doteq (S_t, A_t, S_{t+1})$. It is easy to see that $\{Y_t\}$ evolves in the finite space $\mathcal{Y} \doteq \{(s,a,s') \mid \pi(a|s) > 0, p(s'|s,a) > 0\}$. We then define a function $H : \mathbb{R}^{|\mathcal{S}|} \times \mathcal{Y} \to \mathbb{R}^{|\mathcal{S}|}$ by defining the s-th element of $H(v,(s_0,a_0,s_1))$ as

$$H(v,(s_0,a_0,s_1))[s] \doteq$$
$$\mathbb{I}_{\{s=s_0\}}(r(s_0,a_0) - \bar{J}_\pi + v(s_1) - v(s_0)) + v(s).$$

Then, the update to $\{v_t\}$ in (Average Reward TD) can then be expressed as

$$v_{t+1} = v_t + \alpha_{t+1}(H(v_t, Y_{t+1}) - v_t + \epsilon_{t+1}). \tag{22}$$

Here, $\epsilon_{t+1} \in \mathbb{R}^{|\mathcal{S}|}$ is the random noise vector defined as $\epsilon_{t+1}(s) \doteq \mathbb{I}\{s = S_t\}(J_t - \bar{J}_\pi)$. This ϵ_{t+1} is the current estimate error of the average reward estimator J_t. Intuitively, the indicator $\mathbb{I}\{s = S_t\}$ reflects the asynchronous nature of (Average Reward TD). For each t, only the S_t-indexed element in v_t is updated.

Throughout the rest of the section, we utilize the following assumption.

Assumption 4.1 (Ergodicity)**.** Both $\mathcal{S}$ and $\mathcal{A}$ are finite. The Markov chain $\{S_t\}$ induced by the policy π is aperiodic and irreducible.

Theorem 4.2. *Let Assumption 4.1 hold. Consider the learning rates in the form of $\alpha_t = \frac{1}{(t+1)^b}, \beta_t = \frac{1}{t}$ with $b \in (\frac{4}{5}, 1]$. Then the iterates $\{v_t\}$ generated by* (Average Reward TD) *satisfy*

$$\lim_{t\to\infty} v_t = v_* \quad \textit{a.s.},$$

where $v_* \in \mathcal{V}_*$ is a possibly sample-path dependent fixed point.

Proof. We proceed via verifying assumptions of Theorem 2.6. In particular, we consider the compact form (22).

Under Assumption 4.1, it is obvious that $\{Y_t\}$ is irreducible and aperiodic and adopts a unique stationary distribution.

To verify Assumption 2.2, we demonstrate that H is $1-$Lipschitz in v w.r.t $\|\cdot\|_\infty$. For notation simplicity, let $y = (s_0, a_0, s_1)$. Separating by cases based on s, we have

$$|H(v,y)[s] - H(v',y)[s]| = \begin{cases} |v(s) - v'(s)|, & s \neq s_0, \\ |v(s_1) - v'(s_1)|, & s = s_0, \end{cases}$$

and in both cases the right side is at most $\|v - v'\|_\infty$. Thus,

$$\|H(v,y) - H(v',y)\|_\infty = \max_{s\in\mathcal{S}} |H(v,y)[s] - H(v',y)[s]|$$
$$\leq \|v - v'\|_\infty.$$

It is well known that the set of solutions to Poisson's equation $\mathcal{V}_*$ defined in (21) is non-empty (Puterman 2014), verifying Assumption 2.3. Assumption 2.4 is directly met by the definition of α_t.

To verify Assumption 2.5, we first notice that for (Average Reward TD), we have $\left\|\epsilon_t^{(1)}\right\|_\infty = \left|\bar{J}_\pi - J_t\right|$. It is well-known from the ergodic theorem that J_t converges to $\bar{J}_\pi$ almost surely. Assumption 2.5, however, requires both an almost sure convergence rate and an L^2 convergence rate. To this end, we rewrite the update of $\{J_t\}$ as

$$J_{t+1} = J_t + \beta_{t+1}\left(R_{t+1} + \gamma J_t \phi(S_{t+1}) - J_t\phi(S_t)\right)\phi(S_t),$$

where we define $\gamma \doteq 0$ and $\phi(s) \doteq 1\,\forall s$. It is now clear that the update of $\{J_t\}$ is a special case of linear TD in the discounted setting (Sutton 1988). Given our choice of $\beta_t = \frac{1}{t}$, the general result about the almost sure convergence rate of linear TD (Theorem 1 of Tadic (2002)) ensures that

$$\left|J_t - \bar{J}_\pi\right| \leq \frac{\zeta_{4.2}\sqrt{\ln\ln t}}{\sqrt{t}} \quad \text{a.s.},$$

where $\zeta_{4.2}$ is a sample-path dependent constant. This immediately verifies (3). We do note that this almost sure convergence rate can also be obtained via a law of the iterated logarithm for Markov chains (Theorem 17.0.1 of Meyn and Tweedie (2012)). The general result about the L^2 convergence rate of linear TD (Theorem 11 of Srikant and Ying (2019)) ensures that

$$\mathbb{E}\left[\left|J_t - \bar{J}_\pi\right|^2\right] = \mathcal{O}(1\,/\,t).$$

This immediately verifies (4) and completes the proof. □

Remark 4.3. The convergence rate we established in Theorem 3.1 also applies directly to the update in (Average Reward TD), and yields a bound on the expected residuals. However, this rate does not improve upon the existing result in Zhang, Zhang, and Maguluri (2021), and thus we omit it here. A further discussion on the significance of Theorem 4.2 in comparison to the results in Zhang, Zhang, and Maguluri (2021) is deferred to the subsequent section.

4.3 Significance of Theorem 4.2

Since (Average Reward TD) has been previously studied, we highlight the significance of Theorem 4.2, which provides the first proof of almost sure convergence of (Average Reward TD) to a (possibly sample-path dependent) fixed point in the tabular setting.

Tsitsiklis and Roy (1999) proves the almost sure convergence for linear function approximation, where $v(s)$ is approximated by $\phi(s)^\top w$ with feature matrix $\Phi \in \mathbb{R}^{|\mathcal{S}|\times K}$. This setting reduces to the tabular case when $\Phi = I$. However, their result requires assumptions like linear independence of Φ's columns and $\Phi w \neq ce$ for any scalar c. The latter unfortunately does not hold in the tabular case (e.g., $Ie = e$). With a non-trivial construction of Φ, it is possible to adapt their result to show that the $\{v_t\}$ in (Average Reward TD) converge almost surely to some (possibly sample-path dependent) subset of $\mathcal{V}_*$. Even so, it is not clear whether $\{v_t\}$ itself converges. It is possible that $\{v_t\}$ oscillates inside or around $\mathcal{V}_*$. Our result rules out this possibility by showing that on every sample-path $\{v_t\}$ must converge to a single fixed point, although different sample-paths may converge to different fixed points.

Zhang, Zhang, and Maguluri (2021) later established L^2 convergence for the linear case without requiring $\Phi w \neq ce$, and derived convergence rates. However, L^2 convergence does not imply almost sure convergence, and even if one could strengthen their result to almost sure convergence, it would still only guarantee convergence to a set rather than a fixed point.

Chen (2025) studies average reward TD using a seminorm contraction argument and show that the seminorm distance of the iterates to the fixed point set converges to zero. This does not imply convergence of the iterates themselves, since distinct points can have zero seminorm distance, so oscillations within $\mathcal{V}_*$ are not ruled out. Theorem 4.2 provides a stronger result by proving almost sure convergence of the iterates to a fixed point.

5 Related Work

ODE and Lyapunov Methods for Asymptotic Convergence A large body of research has employed ODE-based methods to establish almost sure convergence of SA algorithms (Benveniste, Métivier, and Priouret 1990; Kushner and Yin 2003; Borkar 2009). These methods typically begin by proving the stability of the iterates $\{x_n\}$ (i.e., $\sup_n \|x_n\| < \infty$). Abounadi, Bertsekas, and Borkar (2002) use this ODE method to study the convergence of SKM iterations, but they require the additive noise sequence to be uniformly bounded, and that the set of fixed points of the nonexpansive map be a singleton to prove the stability of the iterates.

The ODE@∞ technique (Borkar and Meyn 2000; Borkar et al. 2021; Meyn 2024; Liu, Chen, and Zhang 2025) is a powerful stability technique in RL. If the so-called "ODE@∞ is globally asymptotically stable, existing results such as Meyn (2022); Borkar et al. (2021); Liu, Chen, and Zhang (2025) can be used to establish the desired stability of $\{x_t\}$. However, if we consider a generic non-expansive operator h which may admit multiple fixed points or induce oscillatory behavior, we cannot guarantee the global asymptotic stability of the ODE@∞ without additional assumptions. This limits the utility of the ODE@∞ method in analyzing (SKM).

In addition to ODE methods, there are other works that use Lyapunov methods such as (Bertsekas and Tsitsiklis 1996; Konda and Tsitsiklis 1999; Srikant and Ying 2019; Borkar et al. 2021; Chen et al. 2021; Zhang, des Combes, and Laroche 2022; Zhang, Des Combes, and Laroche 2023) to provide asymptotic and non-asymptotic results of various RL algorithms. Both the ODE and Lyapunov based methods are distinct from the fox-and-hare based approach for (KM) with additive noise introduced by (Cominetti, Soto, and Vaisman 2014) upon which our work is built.

Average Reward RL The (Average Reward TD) algorithm introduced by Tsitsiklis and Roy (1999) is the most fundamental policy evaluation algorithm in average reward settings. In addition to the tabular setting we study here, (Average Reward TD) has also been extended to linear function approximation (Tsitsiklis and Roy 1999; Konda and Tsitsiklis 1999; Wu et al. 2020; Zhang, Zhang, and Maguluri 2021; Xie et al. 2025).

Furthermore, the (Average Reward TD) algorithm has inspired the design of many other TD algorithms for average reward MDPs, for both policy evaluation and control, including Konda and Tsitsiklis (1999); Yang et al. (2016); Wan, Naik, and Sutton (2021a); Zhang and Ross (2021); Wan, Naik, and Sutton (2021b); He, Wan, and Mahmood (2022); Saxena et al. (2023). Because the operators in the average reward setting are not contractive, we envision that our work will shed light on the almost sure convergence of these algorithms.

6 Conclusion

In this work, we provide the first proof of almost sure convergence as well as non-asymptotic finite sample analysis of stochastic approximations under nonexpansive maps with Markovian noise. As an application, we provide the first proof of almost sure convergence of (Average Reward TD) to a potentially sample-path dependent fixed point. This result highlights the underappreciated strength of SKM iterations, a tool whose potential is often overlooked in the RL community. Addressing several follow-up questions could open the door to proving the convergence of many other RL algorithms. Do SKM iterations converge in L^p? Do they follow a central limit theorem or a law of the iterated logarithm? Can they be extended to two-timescale settings? Resolving these questions could pave the way for significant advancements in RL theory. We leave them for future investigation.

Acknowledgments

EB acknowledges support from the NSF Graduate Research Fellowship under award 1842490. This work is also 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, a Cisco Faculty Research Award, and an Nvidia academic grant program award.

References

Abounadi, J.; Bertsekas, D. P.; and Borkar, V. 2002. Stochastic approximation for nonexpansive maps: Application to Q-learning algorithms. *SIAM Journal on Control and Optimization*.

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

Benveniste, A.; Métivier, M.; and Priouret, P. 1990. *Adaptive Algorithms and Stochastic Approximations*. Springer.

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

Borkar, V.; Chen, S.; Devraj, A.; Kontoyiannis, I.; and Meyn, S. 2021. The ODE method for asymptotic statistics in stochastic approximation and reinforcement learning. *ArXiv Preprint*.

Borkar, V.; Chen, S.; Devraj, A.; Kontoyiannis, I.; and Meyn, S. 2025. The ODE method for asymptotic statistics in stochastic approximation and reinforcement learning. *The Annals of Applied Probability*.

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

Borkar, V. S.; and Meyn, S. P. 2000. The ODE method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*.

Bravo, M.; and Cominetti, R. 2024. Stochastic fixed-point iterations for nonexpansive maps: Convergence and error bounds. *SIAM Journal on Control and Optimization*.

Bravo, M.; Cominetti, R.; and Pavez-Signé, M. 2019. Rates of convergence for inexact Krasnosel'skii–Mann iterations in Banach spaces. *Mathematical Programming*.

Chen, Z. 2025. Non-Asymptotic Guarantees for Average-Reward Q-Learning with Adaptive Stepsizes. *ArXiv Preprint*.

Chen, Z.; Maguluri, S. T.; Shakkottai, S.; and Shanmugam, K. 2021. A Lyapunov theory for finite-sample guarantees of asynchronous Q-learning and TD-learning variants. *ArXiv Preprint*.

Cominetti, R.; Soto, J. A.; and Vaisman, J. 2014. On the rate of convergence of Krasnosel'skii-Mann iterations and their connection with sums of Bernoullis. *Israel Journal of Mathematics*.

Edelstein, M. 1966. A Remark on a Theorem of M. A. Krasnoselski. *American Mathematical Monthly*.

Folland, G. B. 1999. *Real analysis: modern techniques and their applications*. John Wiley & Sons.

He, J.; Wan, Y.; and Mahmood, A. R. 2022. The Emphatic Approach to Average-Reward Policy Evaluation. In *Deep Reinforcement Learning Workshop NeurIPS 2022*.

Ishikawa, S. 1976. Fixed points and iteration of a nonexpansive mapping in a Banach space. *Proceedings of the American Mathematical Society*.

Karandikar, R. L.; and Vidyasagar, M. 2024. Convergence Rates for Stochastic Approximation: Biased Noise with Unbounded Variance, and Applications. *Journal of Optimization Theory and Applications*.

Kiefer, J.; and Wolfowitz, J. 1952. Stochastic Estimation of the Maximum of a Regression Function. *Annals of Mathematical Statistics*.

Kim, T.-H.; and Xu, H.-K. 2007. Robustness of Mann's algorithm for nonexpansive mappings. *Journal of Mathematical Analysis and Applications*.

Konda, V. R.; and Tsitsiklis, J. N. 1999. Actor-Critic Algorithms. In *Advances in Neural Information Processing Systems*.

Koval, V.; and Schwabe, R. 2003. A law of the iterated logarithm for stochastic approximation procedures in d-dimensional Euclidean space. *Stochastic Processes and Their Applications*.

Krasnosel'skii, M. A. 1955. Two remarks on the method of successive approximations. *Uspekhi Matematicheskikh Nauk*.

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

Liu, L.-S. 1995. Ishikawa and Mann iterative process with errors for nonlinear strongly accretive mappings in Banach spaces. *Journal of Mathematical Analysis and Applications*.

Liu, S.; Chen, S.; and Zhang, S. 2025. The ODE Method for Stochastic Approximation and Reinforcement Learning with Markovian Noise. *Journal of Machine Learning Research*.

Métivier, M.; and Priouret, P. 1987. Théorèmes de convergence presque sure pour une classe d'algorithmes stochastiques à pas décroissant. *Probability Theory and Related Fields*.

Meyn, S. 2022. *Control systems and reinforcement learning*. Cambridge University Press.

Meyn, S. 2024. The Projected Bellman Equation in Reinforcement Learning. *IEEE Transactions on Automatic Control*.

Meyn, S. P.; and Tweedie, R. L. 2012. *Markov chains and stochastic stability*. Springer Science & Business Media.

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

Qian, X.; Xie, Z.; Liu, X.; and Zhang, S. 2024. Almost Sure Convergence Rates and Concentration of Stochastic Approximation and Reinforcement Learning with Markovian Noise. *ArXiv Preprint*.

Reich, S. 1979. Weak convergence theorems for nonexpansive mappings in Banach spaces. *Journal of Mathematical Analysis and Applications*.

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

Saxena, N.; Khastagir, S.; Kolathaya, S.; and Bhatnagar, S. 2023. Off-policy average reward actor-critic with deterministic policy search. In *Proceedings of the International Conference on Machine Learning*.

Srikant, R.; and Ying, L. 2019. Finite-time error bounds for linear stochastic approximation andtd learning. In *Proceedings of the Conference on Learning Theory*.

Sutton, R. S. 1988. Learning to Predict by the Methods of Temporal Differences. *Machine Learning*.

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

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

Tadic, V. B. 2002. On The Almost Sure Rate Of Convergence Of Temporal-Difference Learning Algorithms. *IFAC Proceedings Volumes*.

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

Tsitsiklis, J. N.; and Roy, B. V. 1999. Average cost temporal-difference learning. *Automatica*.

Wan, Y.; Naik, A.; and Sutton, R. 2021a. Average-reward learning and planning with options. In *Advances in Neural Information Processing Systems*.

Wan, Y.; Naik, A.; and Sutton, R. S. 2021b. Learning and Planning in Average-Reward Markov Decision Processes. In *Proceedings of the International Conference on Machine Learning*.

Wu, Y.; Zhang, W.; Xu, P.; and Gu, Q. 2020. A Finite-Time Analysis of Two Time-Scale Actor-Critic Methods. In *Advances in Neural Information Processing Systems*.

Xie, Z.; Liu, X.; Chandra, R.; and Zhang, S. 2025. Finite Sample Analysis of Linear Temporal Difference Learning with Arbitrary Features. In *Advances in Neural Information Processing Systems*.

Yang, S.; Gao, Y.; An, B.; Wang, H.; and Chen, X. 2016. Efficient average reward reinforcement learning using constant shifting values. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zhang, S.; des Combes, R. T.; and Laroche, R. 2022. Global Optimality and Finite Sample Analysis of Softmax Off-Policy Actor Critic under State Distribution Mismatch. *Journal of Machine Learning Research*.

Zhang, S.; Des Combes, R. T.; and Laroche, R. 2023. On the convergence of SARSA with linear function approximation. In *Proceedings of the International Conference on Machine Learning*.

Zhang, S.; Zhang, Z.; and Maguluri, S. T. 2021. Finite Sample Analysis of Average-Reward TD Learning and Q-Learning. In *Advances in Neural Information Processing Systems*.

Zhang, Y.; and Ross, K. W. 2021. On-policy deep reinforcement learning for the average-reward criterion. In *Proceedings of the International Conference on Machine Learning*.