Mathematics of Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Two Time-Scale Stochastic Approximation with Controlled Markov Noise and Off-Policy Temporal-Difference Learning

Prasenjit Karmakar, Shalabh Bhatnagar

informs
http://pubsonline.informs.org/journal/moor/
MATHEMATICS OF OPERATIONS RESEARCH
Vol. 43, No. 1, February 2018, pp. 130–151
ISSN 0364-765X (print), ISSN 1526-5471 (online)

Two Time-Scale Stochastic Approximation with Controlled Markov Noise and Off-Policy Temporal-Difference Learning

Prasenjit Karmakar,[a] Shalabh Bhatnagar[a]

[a] Department of Computer Science and Automation, Indian Institute of Science, Bangalore 560012, India

Contact: prasenjitk@csa.iisc.ernet.in, http://orcid.org/0000-0001-6895-2364 (PK); shalabh@csa.iisc.ernet.in (SB)

Received: April 13, 2015
Revised: April 18, 2016; December 21, 2016
Accepted: February 7, 2017
Published Online in Articles in Advance: July 13, 2017

MSC2010 Subject Classification: Primary: 62L20; secondary: 93E35, 68T05
OR/MS Subject Classification: Primary: stochastic model applications; secondary: analysis of algorithms

https://doi.org/10.1287/moor.2017.0855

Abstract. We present for the first time an asymptotic convergence analysis of two time-scale stochastic approximation driven by "controlled" Markov noise. In particular, the faster and slower recursions have nonadditive controlled Markov noise components in addition to martingale difference noise. We analyze the asymptotic behavior of our framework by relating it to limiting differential inclusions in both time scales that are defined in terms of the ergodic occupation measures associated with the controlled Markov processes. Finally, we present a solution to the off-policy convergence problem for temporal-difference learning with linear function approximation, using our results.

Funding: The authors' work was partly supported by the Robert Bosch Centre for Cyber-Physical Systems, Indian Institute of Science, Bangalore, India.

Keywords: Markov noise • two time-scale stochastic approximation • asymptotic convergence • temporal-difference learning

1. Introduction

Stochastic approximation algorithms are sequential nonparametric methods for finding a zero or minimum of a function in the situation where only the noisy observations of the function values are available. Two time-scale stochastic approximation algorithms represent one of the most general subclasses of stochastic approximation methods. These algorithms consist of two coupled recursions, which are updated with different (one is considerably smaller than the other) stepsizes, which, in turn, facilitate convergence for such algorithms.

Two time-scale stochastic approximation algorithms (Borkar [6]) have successfully been applied to several complex problems arising in the areas of reinforcement learning (RL), signal processing, and admission control in communication networks. There are many RL applications (precisely those where parameterization of value function is implemented) where nonadditive Markov noise is present in one or both iterates, thus requiring the current two time-scale framework to be extended to include Markov noise (for example, in Degris et al. [9, p. 5], it is mentioned that to generalize the analysis to Markov noise, the theory of two time-scale stochastic approximation needs to include the latter).

Here, we present a more general framework of two time-scale stochastic approximation with "controlled" Markov noise, i.e., the noise is not simply Markov; rather it is driven by the iterates and an additional control process as well. We analyze the asymptotic behavior of our framework by relating it to limiting differential inclusions in both time scales that are defined in terms of the ergodic occupation measures associated with the controlled Markov processes. Next, using these results for the special case of our framework where the random processes are irreducible Markov chains, we present a solution to the off-policy convergence problem for temporal-difference learning with linear function approximation. While the off-policy convergence problem for RL with linear function approximation has been one of the most interesting problems, there are very few solutions available in the current literature. One such work (Yu [21]) shows the convergence of the least squares temporal-difference learning algorithm with eligibility traces (LSTD(λ)) as well as the TD(λ) algorithm. While the LSTD methods are not feasible when the dimension of the feature vector is large, off-policy TD(λ) is shown to converge only when the eligibility function $\lambda \in [0, 1]$ is very close to 1. Another recent work (Yu [22]) proves weak convergence of several emphatic temporal-difference learning algorithms, which is also designed to solve the off-policy convergence problem. In Sutton et al. [17, 18] and Maei [13], the gradient temporal difference learning (GTD) algorithms were proposed to solve this problem. However, the authors make the assumption that the data is available in the "off-policy" setting (i.e., the off-policy issue is incorporated into the data rather than in the algorithm) whereas, in reality, one has only the "on-policy" Markov trajectory corresponding to a given behavior policy, and we are interested in designing an online learning algorithm. We use one of the algorithms from Maei [13] called TDC with "importance weighting," which takes the "on-policy" data as input and show its convergence using the results we develop. Our convergence analysis can also be extended for the same algorithm with eligibility traces for a sufficiently large range of values of λ. Our results can be used to provide a convergence analysis for RL algorithms such as those in Menache et al. [14] for which convergence proofs have not been provided.

To the best of our knowledge, there are related works such as Tadić [19], Konda and Tsitsiklis [10, 11], and Tadić [20] where two time-scale stochastic approximation algorithms with algorithm iterate dependent non-additive Markov noise is analyzed. In all of them, the Markov noise in the recursion is handled using the classic Poisson equation-based approach of Benveniste et al. [4] and Metivier and Priouret [15] and applied to the asymptotic analysis of many algorithms used in machine learning, system identification, signal processing, image analysis and automatic control. However, we show that our method also works if there is another additional control process as well and if the underlying Markov process has nonunique stationary distributions. Further, the mentioned application does not require strong assumption such as aperiodicity for the underlying Markov chain, which is a sufficient condition if we use Poisson equation-based approach (Ma et al. [12], Tadić [19]). Additionally, our assumptions are quite different from the assumptions made in the mentioned literature and we give a detailed comparison in Section 2.2.

The organization of the paper is as follows. Section 2 formally defines the problem and provides background and assumptions. Section 3 shows the main results. Section 4 discusses how one of our assumptions of Section 2 can be relaxed. Section 5 presents an application of our results to the off-policy convergence problem for temporal-difference learning with linear function approximation. Finally, we conclude by providing some future research directions.

2. Background, Problem Definition, and Assumptions

In the following, we describe the preliminaries and notation used in our proofs. Most of the definitions and notation are from Benaïm et al. [3], Borkar [8], and Aubin and Cellina [1].

2.1. Definition and Notation

Let F denote a set-valued function mapping each point $\theta \in \mathbb{R}^m$ to a set $F(\theta) \subset \mathbb{R}^m$. F is called a *Marchaud map* if the following hold:

(i) F is *upper semicontinuous* in the sense that if $\theta_n \to \theta$ and $w_n \to w$ with $w_n \in F(\theta_n)$ for all $n \geq 1$, then $w \in F(\theta)$. In order words, the graph of F defined as $\{(\theta, w)\colon w \in F(\theta)\}$ is closed.

(ii) $F(\theta)$ is a nonempty compact convex subset of $\mathbb{R}^m$ for all $\theta \in \mathbb{R}^m$.

(iii) $\exists\, c > 0$ such that for all $\theta \in \mathbb{R}^m$,

$$\sup_{z \in F(\theta)} \|z\| \leq c(1 + \|\theta\|),$$

where $\|\cdot\|$ denotes any norm on $\mathbb{R}^m$.

A solution for the differential inclusion (*d.i.*)

$$\dot{\theta}(t) \in F(\theta(t)) \tag{1}$$

with initial point $\theta_0 \in \mathbb{R}^m$ is an absolutely continuous (on compacts) mapping $\theta\colon \mathbb{R} \to \mathbb{R}^m$ such that $\theta(0) = \theta_0$ and

$$\dot{\theta}(t) \in F(\theta(t))$$

for almost every $t \in \mathbb{R}$. If F is a Marchaud map, it is well known that (1) has solutions (possibly nonunique) through every initial point. The differential inclusion (1) induces a *set-valued dynamical system* $\{\Phi_t\}_{t \in \mathbb{R}}$ defined by

$$\Phi_t(\theta_0) = \{\theta(t)\colon \theta(\cdot) \text{ is a solution to (1) with } \theta(0) = \theta_0\}.$$

Consider the autonomous ordinary differential equation (o.d.e.)

$$\dot{\theta}(t) = h(\theta(t)), \tag{2}$$

where h is Lipschitz continuous. One can write (2) in the format of (1) by taking $F(\theta) = \{h(\theta)\}$. It is well known that (2) is well posed, i.e., it has a *unique solution* for every initial point. Hence the set-valued dynamical system induced by the o.d.e. or *flow* is $\{\Phi_t\}_{t \in \mathbb{R}}$ with

$$\Phi_t(\theta_0) = \{\theta(t)\},$$

where $\theta(\cdot)$ is the solution to (2) with $\theta(0)=\theta_0$. It is also well known that $\Phi_t(\cdot)$ is a *continuous function* for all $t\in\mathbb{R}$.

A set $A\subset\mathbb{R}^m$ is said to be *invariant* (for F) if for all $\theta_0\in A$ there exists a solution $\theta(\cdot)$ of (1) with $\theta(0)=\theta_0$ such that $\theta(\mathbb{R})\subset A$.

Given a set $A\subset\mathbb{R}^m$ and $\theta'', w''\in A$, we write $\theta''\hookrightarrow_A w''$ if for every $\epsilon>0$ and $T>0$, there exist n solutions ($n\in\mathbb{N}$) $\theta_1(\cdot),\dots,\theta_n(\cdot)$ to (1) and real numbers $t_1,t_2,\dots,t_n$ greater than T such that

(i) $\theta_i(s)\in A$ for all $0\le s\le t_i$ and for all $i=1,\dots,n$,
(ii) $\|\theta_i(t_i)-\theta_{i+1}(0)\|\le\epsilon$ for all $i=1,\dots,n-1$,
(iii) $\|\theta_1(0)-\theta''\|\le\epsilon$ and $\|\theta_n(t_n)-w''\|\le\epsilon$.

The sequence $(\theta_1(\cdot),\dots,\theta_n(\cdot))$ is called an (ϵ,T) *chain* (in A from θ'' to w'') for F. A set $A\subset\mathbb{R}^m$ is said to be *internally chain transitive*, provided that A is compact and $\theta''\hookrightarrow_A w''$ for all $\theta'',w''\in A$. It can be proved that in the above case, A is an invariant set.

A compact invariant set A is called an *attractor* for Φ, provided that there is a neighborhood U of A (in the induced topology) with the property that $d(\Phi_t(\theta''),A)\to 0$ as $t\to\infty$ *uniformly* in $\theta''\in U$. Here, $d(X,Y)=\sup_{\theta''\in X}\inf_{w''\in Y}\|\theta''-w''\|$ for $X,Y\subset\mathbb{R}^m$. Such a U is called a *fundamental neighborhood* of the attractor A. *An attractor of a well-posed o.d.e.* is an attractor for the set-valued dynamical system induced by the o.d.e.

The set

$$\omega_\Phi(\theta'')=\bigcap_{t\ge 0}\overline{\Phi_{[t,\infty)}(\theta'')}$$

is called the *ω-limit* set of a point $\theta''\in\mathbb{R}^m$. If A is a set, then

$$B(A)=\{\theta''\in\mathbb{R}^m:\ \omega_\Phi(\theta'')\subset A\}$$

denotes its *basin of attraction*. A *global attractor* for Φ is an attractor A whose basin of attraction consists of all $\mathbb{R}^m$. Then, the following lemma will be useful for our proofs, see Benaïm et al. [3] for a proof.

Lemma 1. *Suppose Φ has a global attractor A. Then, every internally chain transitive set lies in A.*

We also require another result, which will be useful to apply our results to the RL application we mention. Before stating it, we recall some definitions from Borkar [8, Appendix 11.2.3].

A point $\theta^*\in\mathbb{R}^m$ is called *Lyapunov stable* for the o.d.e. (2) if for all $\epsilon>0$, there exists a $\delta>0$ such that every trajectory of (2) initiated in the δ-neighborhood of θ^* remains in its ϵ-neighborhood. θ^* is called *globally asymptotically stable* if θ^* is Lyapunov stable and *all* trajectories of the o.d.e. converge to it.

Lemma 2. *Consider the autonomous o.d.e. $\dot\theta(t)=h(\theta(t))$, where h is Lipschitz continuous. Let θ^* be globally asymptotically stable. Then, θ^* is the global attractor of the o.d.e.*

Proof. We refer the readers to Borkar [8, Lemma 1, Chapter 3] for a proof. □

We end this subsection with a notation, which will be used frequently in the convergence statements in the following sections.

Definition 1. For function $\theta(\cdot)$ defined on $[0,\infty)$, the notation "$\theta(t)\to A$ as $t\to\infty$" means that $\bigcap_{t\ge 0}\overline{\{\theta(s):\ s\ge t\}}\subset A$. Similar definition applies for a sequence $\{\theta_n\}$.

2.2. Problem Definition

Our goal is to perform an asymptotic analysis of the following coupled recursions:

$$\theta_{n+1}=\theta_n+a(n)[h(\theta_n,w_n,Z_n^{(1)})+M_{n+1}^{(1)}],\tag{3}$$

$$w_{n+1}=w_n+b(n)[g(\theta_n,w_n,Z_n^{(2)})+M_{n+1}^{(2)}],\tag{4}$$

where $\theta_n\in\mathbb{R}^d$, $w_n\in\mathbb{R}^k$, $n\ge 0$ and $\{Z_n^{(i)}\},\{M_n^{(i)}\}$, $i=1,2$ are random processes that we describe below.

We make the following assumptions:

(A1) $\{Z_n^{(i)}\}$ takes values in a compact metric space $S^{(i)}$, $i=1,2$. Additionally, the processes $\{Z_n^{(i)}\}$, $i=1,2$ are controlled Markov processes that are controlled by three different control processes: the iterate sequences $\{\theta_m\}$, $\{w_m\}$, and a random process $\{A_n^{(i)}\}$ taking values in a compact metric space $U^{(i)}$, respectively, with their individual dynamics specified by

$$P(Z_{n+1}^{(i)}\in B^{(i)}\mid Z_m^{(i)},A_m^{(i)},\theta_m,w_m,m\le n)=\int_{B^{(i)}}p^{(i)}(dy\mid Z_n^{(i)},A_n^{(i)},\theta_n,w_n),\quad n\ge 0$$

for $B^{(i)}$ Borel in $S^{(i)}$, $i=1,2$, respectively.

Remark 1. In this context, one should note that Benveniste et al. [4] and Metivier and Priouret [15] require the Markov process to take values in a normed Polish space.

Remark 2. In Borkar [7], it is assumed that the state space where the controlled Markov process takes values is Polish. This space is then compactified using the fact that a Polish space can be homeomorphically embedded into a dense subset of a compact metric space. The vector field $h(\cdot,\cdot)$: $\mathbb{R}^d \times S \to \mathbb{R}^d$ is considered bounded when the first component lies in a compact set. This would, however, require a continuous extension of h': $\mathbb{R}^d \times \phi(S) \to \mathbb{R}^d$ defined by $h'(x,s') = h(x,\phi^{-1}(s'))$ to $\mathbb{R}^d \times \overline{\phi(S)}$. Here, $\phi(\cdot)$ is the homeomorphism defined by $\phi(s) = (\rho(s,s_1), \rho(s,s_2), \ldots) \in [0,1]^\infty$, and $\{s_i\}$ and ρ is a countable dense subset and metric of the Polish space, respectively. A sufficient condition for the above is h' to be uniformly continuous (Rudin [16, Example 13, p. 99]). However, this is hard to verify. This is the main motivation for us to take the range of the Markov process as compact for our problem. However, there are other reasons for taking compact state space, which will be clear in the proofs of this section and the next.

(A2) h: $\mathbb{R}^{d+k} \times S^{(1)} \to \mathbb{R}^d$ is jointly continuous as well as Lipschitz in its first two arguments uniformly w.r.t. the third. The latter condition means that

$$\forall\, z^{(1)} \in S^{(1)}, \quad \|h(\theta,w,z^{(1)}) - h(\theta',w',z^{(1)})\| \le L^{(1)}(\|\theta-\theta'\| + \|w-w'\|).$$

The same thing is also true for g, where the Lipschitz constant is $L^{(2)}$. Note that the Lipschitz constant $L^{(i)}$ does not depend on $z^{(i)}$ for $i=1,2$.

Remark 3. We later relax the uniformity of the Lipschitz constant w.r.t. the Markov process state space by putting suitable moment assumptions on the Markov process.

(A3) $\{M_n^{(i)}\}$, $i=1,2$ are martingale difference sequences w.r.t. increasing σ-fields

$$\mathcal{F}_n = \sigma(\theta_m, w_m, M_m^{(i)}, Z_m^{(i)}, m \le n, i=1,2), \quad n \ge 0,$$

satisfying

$$E[\|M_{n+1}^{(i)}\|^2 \mid \mathcal{F}_n] \le K(1 + \|\theta_n\|^2 + \|w_n\|^2), \quad i=1,2$$

for $n \ge 0$ and a given constant $K > 0$.

(A4) The stepsizes $\{a(n)\},\{b(n)\}$ are positive scalars satisfying

$$\sum_n a(n) = \sum_n b(n) = \infty, \qquad \sum_n (a(n)^2 + b(n)^2) < \infty, \qquad \frac{a(n)}{b(n)} \to 0.$$

Moreover, $a(n)$, $b(n)$, $n \ge 0$ are nonincreasing.

Before stating the assumption on the transition kernel $p^{(i)}$, $i=1,2$, we need to define the metric in the space of probability measures $\mathcal{P}(S)$. Here, we mention the definitions and main theorems on the spaces of probability measures that we use in our proofs (details can be found in Borkar [5, Chapter 2]). We denote the metric by d and is defined as

$$d(\mu,\nu) = \sum_j 2^{-j} \left| \int f_j d\mu - \int f_j\, d\nu \right|, \quad \mu,\nu \in \mathcal{P}(S),$$

where $\{f_j\}$ are countable dense in the unit ball of $C(S)$. Then, the following are equivalent:

(i) $d(\mu_n,\mu) \to 0$,

(ii) for all bounded f in $C(S)$,

$$\int_S f\, d\mu_n \to \int_S f\, d\mu, \tag{5}$$

(iii) for all f bounded and uniformly continuous,

$$\int_S f\, d\mu_n \to \int_S f\, d\mu.$$

Hence we see that $d(\mu_n,\mu) \to 0$ iff $\int_S f_j\, d\mu_n \to \int_S f_j\, d\mu$ for all j. Any such sequence of functions $\{f_j\}$ is called a convergence determining class in $\mathcal{P}(S)$. Sometimes we also denote $d(\mu_n,\mu) \to 0$ using the notation $\mu_n \Rightarrow \mu$.

Also, we recall the characterization of relative compactness in $\mathcal{P}(S)$ that relies on the definition of tightness. $\mathcal{A} \subset \mathcal{P}(S)$ is a tight set if for any $\epsilon > 0$, there exists a compact $K_\epsilon \subset S$ such that $\mu(K_\epsilon) > 1-\epsilon$ for all $\mu \in \mathcal{A}$. Clearly, if S is compact, then any $\mathcal{A} \subset \mathcal{P}(S)$ is tight. By Prohorov's theorem, $\mathcal{A} \subset \mathcal{P}(S)$ is relatively compact if and only if it is tight.

With the above definitions, we assume the following:

(A5) The map $S^{(i)} \times U^{(i)} \times \mathbb{R}^{d+k} \ni (z^{(i)}, a^{(i)}, \theta, w) \to p^{(i)}(dy \mid z^{(i)}, a^{(i)}, \theta, w) \in \mathcal{P}(S^{(i)})$ is continuous.

Remark 4. **(A5)** is much simpler than the assumptions on n-step transition kernel in Benveniste et al. [4, Part II, Chapter 2, Theorem 6].

Additionally, unlike Borkar [7, p. 140, line 13], we do not require the extra assumption of the continuity in the θ variable of $p(dy \mid z, a, \theta)$ to be uniform on compacts w.r.t. the other variables.

For $\theta_n = \theta$, $w_n = w$ for all n with a fixed deterministic $(\theta, w) \in \mathbb{R}^{d+k}$ and under any stationary randomized control $\pi^{(i)}$, it follows from Borkar [7, Lemmas 2.1 and 3.1] that the time-homogeneous Markov processes $Z_n^{(i)}$, $i = 1, 2$ have (possibly nonunique) invariant distributions $\Psi^{(i)}_{\theta, w, \pi^{(i)}}$, $i = 1, 2$.

Now, it is well known that the ergodic occupation measure defined as

$$\Psi^{(i)}_{\theta, w, \pi^{(i)}}(dz, da) := \Psi^{(i)}_{\theta, w, \pi^{(i)}}(dz)\pi^{(i)}(z, da) \in \mathcal{P}(S^{(i)} \times U^{(i)})$$

satisfies the following:

$$\int_{S^{(i)}} f^{(i)}(z)\, \Psi^{(i)}_{\theta, w, \pi^{(i)}}(dz, U^{(i)}) = \int_{S^{(i)} \times U^{(i)}} \int_{S^{(i)}} f^{(i)}(y) p^{(i)}(dy \mid z, a, \theta, w)\, \Psi^{(i)}_{\theta, w, \pi^{(i)}}(dz, da) \tag{6}$$

for $f^{(i)}\colon S^{(i)} \to \mathcal{R} \in C_b(S^{(i)})$.

We denote by $D^{(i)}(\theta, w)$, $i = 1, 2$, the set of all such ergodic occupation measures for the prescribed θ and w. In the following, we prove some properties of the map $(\theta, w) \to D^{(i)}(\theta, w)$.

Lemma 3. *For all (θ, w), $D^{(i)}(\theta, w)$ is convex and compact.*

Proof. The proof trivially follows from **(A1)**, **(A5)**, and (6). □

Lemma 4. *The map $(\theta, w) \to D^{(i)}(\theta, w)$ is upper semicontinuous.*

Proof. Let $\theta_n \to \theta$, $w_n \to w$ and $\Psi_n^{(i)} \Rightarrow \Psi^{(i)} \in \mathcal{P}(S^{(i)} \times U^{(i)})$ such that $\Psi_n^{(i)} \in D^{(i)}(\theta_n, w_n)$. Let

$$g_n^{(i)}(z, a) = \int_{S^{(i)}} f^{(i)}(y) p^{(i)}(dy \mid z, a, \theta_n, w_n) \quad \text{and}$$
$$g^{(i)}(z, a) = \int_{S^{(i)}} f^{(i)}(y) p^{(i)}(dy \mid z, a, \theta, w).$$

From (6), we get that

$$\begin{aligned}\int_{S^{(i)}} f^{(i)}(z)\, \Psi^{(i)}(dz, U^{(i)}) &= \lim_{n \to \infty} \int_{S^{(i)}} f^{(i)}(z)\, \Psi_n^{(i)}(dz, U^{(i)}) \\ &= \lim_{n \to \infty} \int_{S^{(i)} \times U^{(i)}} \int_{S^{(i)}} f^{(i)}(y) p^{(i)}(dy \mid z, a, \theta_n, w_n)\, \Psi_n^{(i)}(dz, da) \\ &= \lim_{n \to \infty} \int_{S^{(i)} \times U^{(i)}} g_n^{(i)}(z, a)\, \Psi_n^{(i)}(dz, da).\end{aligned}$$

Now, $p^{(i)}(dy \mid z, a, \theta_n, w_n) \Rightarrow p^{(i)}(dy \mid z, a, \theta, w)$ implies $g_n^{(i)}(\cdot, \cdot) \to g^{(i)}(\cdot, \cdot)$ pointwise. We prove that the convergence is indeed uniform. It is enough to prove that this sequence of functions is equicontinuous. Then, along with pointwise convergence, it will imply uniform convergence on compacts (Rudin [16, p. 168, Example 16]). This is also a place where **(A1)** is used.

Define $g'\colon S^{(i)} \times U^{(i)} \times \mathbb{R}^{d+k} \to \mathbb{R}$ by $g'(z', a', \theta', w') = \int_{S^{(i)}} f^{(i)}(y) p^{(i)}(dy \mid z, a', \theta', w')$. Then, g' is continuous. Let $A = S^{(i)} \times U^{(i)} \times (\{\theta_n\} \cup \theta) \times (\{w_n\} \cup w)$. Therefore A is compact and $g'|_A$ is uniformly continuous. This implies that for all $\epsilon > 0$, there exists $\delta > 0$ such that if $\rho'(s_1, s_2) < \delta$, $\mu'(a_1, a_2) < \delta$, $\|\theta_1 - \theta_2\| < \delta$, $\|w_1 - w_2\| < \delta$, then $|g'(s_1, a_1, \theta_1, w_1) - g'(s_2, a_2, \theta_2, w_2)| < \epsilon$, where $s_1, s_2 \in S^{(i)}$, $a_1, a_2 \in U^{(i)}$, $\theta_1, \theta_2 \in (\{\theta_n\} \cup \theta)$, $w_1, w_2 \in (\{w_n\} \cup w)$ and ρ' and μ' denote the metrics in $S^{(i)}$ and $U^{(i)}$, respectively. Now, use this same δ for the $\{g_n^{(i)}(\cdot, \cdot)\}$ to get for all n the following for $\rho'(z_1, z_2) < \delta$, $\mu'(a_1, a_2) < \delta$:

$$|g_n^{(i)}(z_1, a_1) - g_n^{(i)}(z_2, a_2)| = |g'(z_1, a_1, \theta_n, w_n) - g'(z_2, a_2, \theta_n, w_n)| < \epsilon.$$

Hence $\{g_n^{(i)}(\cdot,\cdot)\}$ is equicontinuous. For large n, $\sup_{(z,a)\in S^{(i)}\times U^{(i)}} |g_n^{(i)}(z,a) - g^{(i)}(z,a)| < \epsilon/2$ because of uniform convergence of $\{g_n^{(i)}(\cdot,\cdot)\}$, hence $\int_{S^{(i)}\times U^{(i)}} |g_n^{(i)}(z,a) - g^{(i)}(z,a)|\,\Psi_n^{(i)}(dz,da) < \epsilon/2$. Now, (for n large),

$$\begin{aligned}
&\left|\int_{S^{(i)}\times U^{(i)}} g_n^{(i)}(z,a)\,\Psi_n^{(i)}(dz,da) - \int_{S^{(i)}\times U^{(i)}} g^{(i)}(z,a)\,\Psi^{(i)}(dz,da)\right|\\
&\quad= \left|\int_{S^{(i)}\times U^{(i)}} [g_n^{(i)}(z,a) - g^{(i)}(z,a)]\,\Psi_n^{(i)}(dz,da) + \int_{S^{(i)}\times U^{(i)}} g^{(i)}(z,a)\,\Psi_n^{(i)}(dz,da) - \int_{S^{(i)}\times U^{(i)}} g^{(i)}(z,a)\,\Psi^{(i)}(dz,da)\right|\\
&\quad< \epsilon/2 + \left|\int_{S^{(i)}\times U^{(i)}} g^{(i)}(z,a)\,\Psi_n^{(i)}(dz,da) - \int_{S^{(i)}\times U^{(i)}} g^{(i)}(z,a)\,\Psi^{(i)}(dz,da)\right|\\
&\quad< \epsilon.
\end{aligned} \tag{7}$$

The last inequality follows the fact that $\Psi_n^{(i)} \Rightarrow \Psi^{(i)}$. Hence, from (7), we get

$$\int_{S^{(i)}} f^{(i)}(z)\,\Psi^{(i)}(dz,U^{(i)}) = \int_{S^{(i)}\times U^{(i)}} \int_{S^{(i)}} f^{(i)}(y) p^{(i)}(dy \mid z,a,\theta,w)\,\Psi^{(i)}(dz,da)$$

proving that the map is upper semicontinuous. □

Define $\tilde{g}(\theta,w,\nu) = \int g(\theta,w,z)\nu(dz,U^{(2)})$ for $\nu \in P(S^{(2)}\times U^{(2)})$ and $\hat{g}(\theta,w) = \{\tilde{g}(\theta,w,\nu)\colon \nu \in D^{(2)}(\theta,w)\}$.

Lemma 5. *$\hat{g}(\cdot,\cdot)$ is a Marchaud map.*

Proof. (i) Convexity and compactness follow trivially from the same for the map $(\theta,w)\to D^{(2)}(\theta,w)$. (ii)

$$\begin{aligned}
\|\tilde{g}(\theta,w,\nu)\| &= \left\|\int g(\theta,w,z)\nu(dz,U^{(2)})\right\| \le \int \|g(\theta,w,z)\|\nu(dz,U^{(2)})\\
&\le \int L^{(2)}(\|\theta\| + \|w\| + \|g(0,0,z)\|)\nu(dz,U^{(2)})\\
&\le \max\left(L^{(2)}, L^{(2)}\int \|g(0,0,z)\|\nu(dz,U^{(2)})\right)(1 + \|\theta\| + \|w\|).
\end{aligned}$$

Clearly, $K(\theta) = \max(L^{(2)}, L^{(2)}\int \|g(0,0,z)\|\nu(dz,U^{(2)})) > 0$. The above is true for all $\tilde{g}(\theta,w,\nu) \in \hat{g}(\theta,w)$, $\nu \in D^{(2)}(\theta,w)$. (iii) Let $(\theta_n,w_n)\to(\theta,w)$, $\tilde{g}(\theta_n,w_n,\nu_n)\to m$, $\nu_n \in D^{(2)}(\theta_n,w_n)$. Now, $\{\nu_n\}$ is tight, hence has a convergent subsequence $\{\nu_{n_k}\}$ with ν being the limit. Then, using the arguments similar to the proof of Lemma 4, one can show that $m = \tilde{g}(\theta,w,\nu)$, whereas $\nu \in D^{(2)}(\theta,w)$ follows directly from the upper semicontinuity of the map $(\theta,w)\to D^{(2)}(\theta,w)$ for all θ. □

Note that the map $\hat{h}(\cdot,\cdot)$ can be defined similarly and can be shown to be a Marchaud map using the exact same technique.

2.3. Other Assumptions Needed for Two Time-Scale Convergence Analysis

We now list the other assumptions required for two time-scale convergence analysis:

(A6) for all $\theta \in \mathbb{R}^d$, the differential inclusion

$$\dot{w}(t) \in \hat{g}(\theta, w(t)) \tag{8}$$

has a singleton global attractor $\lambda(\theta)$, where $\lambda\colon \mathbb{R}^d \to \mathbb{R}^k$ is a Lipschitz map with constant K. Additionally, there exists a continuous function $V\colon \mathbb{R}^{d+k}\to[0,\infty)$ satisfying the hypothesis of Benaïm et al. [3, Corollary 3.28] with $\Lambda = \{(\theta,\lambda(\theta))\colon \theta\in\mathbb{R}^d\}$. This is the most important assumption as it links the fast and slow iterates.

(A7) Stability of the iterates: $\sup_n(\|\theta_n\| + \|w_n\|) < \infty$, a.s.

Let $\bar{\theta}(\cdot), t\ge 0$ be the continuous, piecewise linear trajectory defined by $\bar{\theta}(t(n)) = \theta_n$, $n \ge 0$, with linear interpolation on each interval $[t(n), t(n+1))$, i.e.,

$$\bar{\theta}(t) = \theta_n + (\theta_{n+1} - \theta_n)\frac{t - t(n)}{t(n+1) - t(n)}, \quad t \in [t(n), t(n+1)).$$

The following theorem is our main result.

Theorem 1 (Slower Time-Scale Result). *Under assumptions* **(A1)–(A7)**,

$$(\theta_n, w_n) \to \bigcup_{\theta^* \in A_0} (\theta^*, \lambda(\theta^*)), \quad \text{a.s. as } n \to \infty,$$

where $A_0 = \bigcap_{t \geq 0} \overline{\{\bar{\theta}(s)\colon s \geq t\}}$ is almost everywhere an internally chain transitive set of the differential inclusion

$$\dot{\theta}(t) \in \hat{h}(\theta(t)), \tag{9}$$

where $\hat{h}(\theta) = \{\tilde{h}(\theta, \lambda(\theta), \nu)\colon \nu \in D^{(1)}(\theta, \lambda(\theta))\}$. We call (8) and (9) as the faster and slower d.i. to correspond with faster and slower recursions, respectively.

Corollary 1. *Under the additional assumption that the inclusion*

$$\dot{\theta}(t) \in \hat{h}(\theta(t))$$

has a global attractor set A_1,

$$(\theta_n, w_n) \to \bigcup_{\theta^* \in A_1} (\theta^*, \lambda(\theta^*)), \quad \text{a.s. as } n \to \infty.$$

Remark 5. In case where the set $D^{(2)}(\theta, w)$ is singleton, we can relax **(A6)** to local attractors also. The relaxed assumption will be

(A6)′ The function $\hat{g}(\theta, w) = \int g(\theta, w, z)\Gamma^{(2)}_{\theta, w}(dz)$ is Lipschitz continuous where $\Gamma^{(2)}_{\theta, w}$ is the only element of $D^{(2)}(\theta, w)$. Further, for all $\theta \in \mathbb{R}^d$, the o.d.e.

$$\dot{w}(t) = \hat{g}(\theta, w(t)) \tag{10}$$

has an asymptotically stable equilibrium $\lambda(\theta)$ with domain of attraction G_θ, where $\lambda\colon \mathbb{R}^d \to \mathbb{R}^k$ is a Lipschitz map with constant K. Also, assume that $\bigcap_\theta G_\theta$ is nonempty. Moreover, the function $V'\colon G \to [0, \infty)$ defined by $V'(\theta, w) = V_\theta(w)$ is continuously differentiable where $V_\theta(\cdot)$ is the Lyapunov function (for the definition, see Borkar [8, Chapter 11.2.3]) for the o.d.e. (10) with $\lambda(\theta)$ as its attractor, and $G = \bigcup_{\theta \in \mathbb{R}^d} \{\{\theta\} \times G_\theta\}$. This extra condition is needed so that the set graph(λ) $:= \{(\theta, \lambda(\theta))\colon \theta \in \mathbb{R}^d\}$ becomes an asymptotically stable set of the coupled o.d.e.

$$\dot{w}(t) = \hat{g}(\theta(t), w(t)), \qquad \dot{\theta}(t) = 0.$$

Note that **(A6)′** allows multiple attractors (at least one of them have to be a point, others can be sets) for the faster o.d.e. for every θ.

Then, the statement of Theorem 1 will be modified as in the following.

Theorem 2 (Slower Time-Scale Result When $\lambda(\theta)$ Is a Local Attractor). *Under assumptions* **(A1)–(A5)**, **(A6)′**, *and* **(A7)**, *on the event "*$\{w_n\}$ *belongs to a compact subset B (depending on the sample point) of* $\bigcap_{\theta \in \mathbb{R}^d} G_\theta$ **eventually**,*"*

$$(\theta_n, w_n) \to \bigcup_{\theta^* \in A_0} (\theta^*, \lambda(\theta^*)), \quad \text{a.s. as } n \to \infty.$$

The requirement on $\{w_n\}$ is much stronger than the usual local attractor statement for Kushner-Clarke lemma (Metivier and Priouret [15, Section II.C]), which requires the iterates to enter a compact set in the domain for attraction of the local attractor *infinitely often* only. The reason for imposing this strong assumption is that graph(λ) is not a subset of any compact set in $\mathbb{R}^{d+k}$, and hence the usual tracking lemma kind of arguments do not go through directly. One has to relate the limit set of the coupled iterate (θ_n, w_n) to graph(λ) (see the proof of Lemma 11).

We present the proof of our main results in the next section.

3. Main Results

We first discuss an extension of the single time-scale controlled Markov noise framework of Borkar [7] under our assumptions to prove our main results. Note that the results of Borkar [7] assume that the state space of the controlled Markov process is Polish, which may impose additional conditions that are hard to verify. In this section, other than proving our two time-scale results, we prove many of the results in Borkar [7] (which were only stated there) assuming the state space to be compact.

We begin by describing the intuition behind the proof techniques in Borkar [7].

The space $C([0,\infty);\mathbb{R}^d)$ of continuous functions from $[0,\infty)$ to $\mathbb{R}^d$ is topologized with the coarsest topology such that the map that takes any $f \in C([0,\infty);\mathbb{R}^d)$ to its restriction to $[0,T]$ when viewed as an element of the space $C([0,T];\mathbb{R}^d)$, is continuous for all $T>0$. In other words, $f_n \to f$ in this space iff $f_n|_{[0,T]} \to f|_{[0,T]}$. The other notations used below are the same as those in Borkar [7, 8]. We present a few for easy reference.

Consider the single time-scale stochastic approximation recursion with controlled Markov noise

$$x_{n+1} = x_n + a(n)[h(x_n, Y_n) + M_{n+1}]. \tag{11}$$

Define time instants $t(0)=0$, $t(n)=\sum_{m=0}^{n-1} a(m)$, $n \geq 1$. Let $\bar{x}(t)$, $t \geq 0$ be the continuous, piecewise linear trajectory defined by $\bar{x}(t(n)) = x_n$, $n \geq 0$, with linear interpolation on each interval $[t(n), t(n+1))$, i.e.,

$$\bar{x}(t) = x_n + (x_{n+1} - x_n)\frac{t - t(n)}{t(n+1) - t(n)}, \quad t \in [t(n), t(n+1)).$$

Now, define $\tilde{h}(x,\nu) = \int h(x,z)\,\nu(dz, U)$ for $\nu \in P(S \times U)$. Let $\mu(t)$, $t \geq 0$ be the random process defined by $\mu(t) = \delta_{(Y_n, Z_n)}$ for $t \in [t(n), t(n+1))$, $n \geq 0$, where $\delta_{(y,a)}$ is the Dirac measure corresponding to (y,a). Consider the nonautonomous o.d.e.

$$\dot{x}(t) = \tilde{h}(x(t), \mu(t)). \tag{12}$$

Let $x^s(t)$, $t \geq s$ denote the solution to (12) with $x^s(s) = \bar{x}(s)$ for $s \geq 0$. Note that $x^s(t)$, $t \in [s, s+T]$ and $x^s(t)$, $t \geq s$ can be viewed as elements of $C([0,T];\mathbb{R}^d)$ and $C([0,\infty);\mathbb{R}^d)$, respectively. With this abuse of notation, it is easy to see that $\{x^s(\cdot)|_{[s,s+T]}, s \geq 0\}$ is a pointwise bounded and equicontinuous family of functions in $C([0,T];\mathbb{R}^d)$ $\forall\, T > 0$. By Arzelà-Ascoli theorem, it is relatively compact. From Borkar [7, Lemma 2.2], one can see that for all $s(n) \uparrow \infty$, $\{\bar{x}(s(n)+\cdot)|_{[s(n),s(n)+T]}, n \geq 1\}$ has a limit point in $C([0,T];\mathbb{R}^d)$ $\forall\, T > 0$. With the above topology for $C([0,\infty);\mathbb{R}^d)$, $\{x^s(\cdot), s \geq 0\}$ is also relatively compact in $C([0,\infty);\mathbb{R}^d)$ and for all $s(n) \uparrow \infty$, $\{\bar{x}(s(n)+\cdot), n \geq 1\}$ has a limit point in $C([0,\infty);\mathbb{R}^d)$.

One can write from (11) the following:

$$\bar{x}(u(n)+t) = \bar{x}(u(n)) + \int_0^t h(\bar{x}(u(n)+\tau), \nu(u(n)+\tau))\,d\tau + W^n(t),$$

where $u(n) \uparrow \infty, \bar{x}(u(n)+\cdot) \to \tilde{x}(\cdot), \nu(t) = (Y_n, Z_n)$ for $t \in [t(n), t(n+1))$, $n \geq 0$ and $W^n(t) = W(t+u(n)) - W(u(n))$, $W(t) = W_n + (W_{n+1} - W_n)((t - t(n))/(t(n+1) - t(n)))$, $W_n = \sum_{k=0}^{n-1} a(k) M_{k+1}$, $n \geq 0$. From here, one cannot directly take limit on both sides as finding limit points of $\nu(s+\cdot)$ as $s \to \infty$ is not meaningful. Now, $h(x,y) = \int h(x,z)\delta_{(y,a)}(dz \times U)$. Hence, by defining $\tilde{h}(x,\rho) = \int h(x,z)\rho(dz)$ and $\mu(t) = \delta_{\nu(t)}$, one can write the above as

$$\bar{x}(u(n)+t) = \bar{x}(u(n)) + \int_0^t \tilde{h}(\bar{x}(u(n)+\tau), \mu(u(n)+\tau))\,d\tau + W^n(t). \tag{13}$$

The advantage is that the space $\mathscr{U}$ of measurable functions from $[0,\infty)$ to $\mathscr{P}(S \times U)$ is compact metrizable, so subsequential limits exist. Note that $\mu(\cdot)$ is not a member of $\mathscr{U}$, rather we need to fix a sample point, i.e., $\mu(\cdot,\omega) \in \mathscr{U}$. For ease of understanding, we abuse the terminology and talk about the limit points $\tilde{\mu}(\cdot)$ of $\mu(s+\cdot)$.

From (13), one can infer that the limit $\tilde{x}(\cdot)$ of $\bar{x}(u(n)+\cdot)$ satisfies the o.d.e. $\dot{x}(t) = \tilde{h}(x(t), \mu(t))$ with $\mu(\cdot)$ replaced by $\tilde{\mu}(\cdot)$. Here, each $\tilde{\mu}(t), t \in \mathbb{R}$ in $\tilde{\mu}(\cdot)$ is generated through different limiting processes each one associated with the compact metrizable space U_t = space of measurable functions from $[0,t]$ to $\mathscr{P}(S \times U)$. This will be problematic if we want to further explore the process $\tilde{\mu}(\cdot)$ and convert the nonautonomous o.d.e. into an autonomous one.

Hence the main result is proved using an auxiliary lemma (Borkar [7, Lemma 2.3]) other than the tracking lemma (Borkar [7, Lemma 2.2]). Let $u(n(k)) \uparrow \infty$ be such that $\bar{x}(u(n(k))+\cdot) \to \tilde{x}(\cdot)$ and $\mu(u(n(k))+\cdot) \to \tilde{\mu}(\cdot)$, then by using Borkar [7, Lemma 2.2], one can show that $x^{u(n(k))}(\cdot) \to \tilde{x}(\cdot)$. Then, the auxiliary lemma shows that the o.d.e. trajectory $x^{u(n(k))}(\cdot)$ associated with $\mu(u(n(k))+\cdot)$ tracks (in the limit) the o.d.e. trajectory associated with $\tilde{\mu}(\cdot)$. Hence Borkar [7, Lemma 2.3] links the two limiting processes $\tilde{x}(\cdot)$ and $\tilde{\mu}(\cdot)$ in some sense. Note that Borkar [7, Lemma 2.3] involves only the o.d.e. trajectories, not the interpolated trajectory of the algorithm.

Consider the iteration

$$\theta_{n+1} = \theta_n + a(n)[h(\theta_n, Y_n) + \epsilon_n + M_{n+1}], \tag{14}$$

where $\epsilon_n \to 0$ and the rest of the notations are the same as Borkar [7]. Specifically, $\{Y_n\}$ is the controlled Markov process driven by $\{\theta_n\}$ and M_{n+1}, $n \geq 0$ is a martingale difference sequence. Let $\bar{\theta}(t)$, $t \geq 0$ be the continuous, piecewise linear trajectory of (14) defined by $\bar{\theta}(t(n)) = \theta_n$, $n \geq 0$, with linear interpolation on each interval $[t(n), t(n+1))$. Also, let $\theta^s(t)$, $t \geq s$ denote the solution to (12) with $\theta^s(s) = \bar{\theta}(s)$ for $s \geq 0$.

The convergence analysis of (14) requires some changes in Borkar [7, Lemmas 2.2 and 3.1]. The modified versions of them are precisely the following two lemmas.

Lemma 6. *For any $T > 0$, $\sup_{t\in[s,s+T]} \|\bar{\theta}(t) - \theta^s(t)\| \to 0$, a.s. as $s \to \infty$.*

Proof. The proof follows from the Borkar [7, Lemma 2.2, Remark 3 (p. 144)]. □

Now, μ can be viewed as a random variable taking values in $\mathcal{U}$ = the space of measurable functions from $[0,\infty)$ to $\mathcal{P}(S \times U)$. This space is topologized with the coarsest topology such that the map

$$\nu(\cdot) \in \mathcal{U} \to \int_0^T g(t) \int f\, d\nu(t)\, dt \in \mathbb{R}$$

is continuous for all $f \in C(S)$, $T > 0$, $g \in L_2[0,T]$. Note that $\mathcal{U}$ is compact metrizable.

Lemma 7. *Almost surely every limit point of $(\mu(s+\cdot), \bar{\theta}(s+\cdot))$ as $s \to \infty$ is of the form $(\tilde{\mu}(\cdot), \tilde{\theta}(\cdot))$, where $\tilde{\mu}(\cdot)$ satisfies $\tilde{\mu}(t) \in D(\tilde{\theta}(t))$ a.e. t.*

Proof. Suppose that $u(n) \uparrow \infty$, $\mu(u(n)+\cdot) \to \tilde{\mu}(\cdot)$, and $\bar{\theta}(u(n)+\cdot) \to \tilde{\theta}(\cdot)$. Let $\{f_i\}$ be countable dense in the unit ball of $C(S)$, hence a separating class, i.e., $\forall\, i$, $\int f_i\, d\mu = \int f_i\, d\nu$ implies $\mu = \nu$. For each i,

$$\zeta_n^i = \sum_{m=1}^{n-1} a(m)\left(f_i(Y_{m+1}) - \int f_i(y) p\,(dy \mid Y_m, Z_m, \theta_m)\right), \quad n \geq 1,$$

is a zero-mean martingale with $\mathcal{F}_n = \sigma(\theta_m, Y_m, Z_m, m \leq n)$. Moreover, it is a square integrable martingale because f_i's are bounded and each ζ_n^i is a finite sum. Its quadratic variation process

$$A_n = \sum_{m=0}^{n-1} a(m)^2 E\left[\left(f_i(Y_{m+1}) - \int f_i(y) p\,(dy \mid Y_m, Z_m, \theta_m)\right)^2 \middle| \mathcal{F}_m\right] + E[(\zeta_0^i)^2]$$

is almost surely convergent. By the martingale convergence theorem, ζ_n^i, $n \geq 0$ converges a.s. for all i. As before let $\tau(n,t) = \min\{m \geq n \colon t(m) \geq t(n) + t\}$ for $t \geq 0$, $n \geq 0$. Then, as $n \to \infty$,

$$\sum_{m=n}^{\tau(n,t)} a(m)\left(f_i(Y_{m+1}) - \int f_i(y) p\,(dy \mid Y_m, Z_m, \theta_m)\right) \to 0, \quad \text{a.s.}$$

for $t > 0$. By our choice of $\{f_i\}$ and the fact that $\{a(n)\}$ is an eventually nonincreasing sequence (the latter property is used only here and in Lemma 14), we have

$$\sum_{m=n}^{\tau(n,t)} (a(m) - a(m+1)) f_i(Y_{m+1}) \to 0, \quad \text{a.s.}$$

From the foregoing,

$$\sum_{m=n}^{\tau(n,t)} \left(a(m+1) f_i(Y_{m+1}) - a(m) \int f_i(y) p\,(dy \mid Y_m, Z_m, \theta_m)\right) \to 0, \quad \text{a.s.}$$

for all $t > 0$, which implies

$$\sum_{m=n}^{\tau(n,t)} a(m)\left(f_i(Y_m) - \int f_i(y) p\,(dy \mid Y_m, Z_m, \theta_m)\right) \to 0, \quad \text{a.s.}$$

for all $t > 0$ because $a(n) \to 0$ and $f_i(\cdot)$ are bounded. This implies

$$\int_{t(n)}^{t(n)+t} \left(\int \left(f_i(z) - \int f_i(y) p\,(dy \mid z, a, \hat{\theta}(s))\right) \mu\,(s, dz\,da)\right) ds \to 0, \quad \text{a.s.}$$

and that, in turn, implies

$$\int_{u(n)}^{u(n)+t}\left(\int\left(f_i(z)-\int f_i(y)p\,(dy\mid z,a,\hat{\theta}(s))\right)\mu(s,dzda)\right)ds\to 0,\quad \text{a.s.}$$

(this is true because $a(n)\to 0$ and $f_i(\cdot)$ is bounded) where $\hat{\theta}(s)=\theta_n$ when $s\in[t(n),t(n+1))$ for $n\geq 0$. Now, one can claim from the above that

$$\int_{u(n)}^{u(n)+t}\left(\int\left(f_i(z)-\int f_i(y)\,p(dy\mid z,a,\bar{\theta}(s))\right)\mu(s,dz\,da)\right)ds\to 0,\quad \text{a.s.}$$

This is because the map $S\times U\times\mathbb{R}^d\ni(z,a,\theta)\to\int f(y)p\,(dy\mid z,a,\theta)$ is continuous and hence uniformly continuous on the compact set $A=S\times U\times M$, where M is the compact set s.t. $\theta_n\in M$ for all n. Here, we also use the fact that $\|\bar{\theta}(s)-\theta_m\|=\|h(\theta_m,Y_m)+\epsilon_m+M_{m+1}\|(s-s_m)\to 0$, $s\in[t_m,t_{m+1})$ as the first two terms inside the norm in the right-hand side (RHS) are bounded. The above convergence is equivalent to

$$\int_0^t\left(\int\left(f_i(z)-\int f_i(y)\,p\big(dy\mid z,a,\bar{\theta}(s+u(n))\big)\right)\mu\big(s+u(n),dz\,da\big)\right)ds\to 0,\quad \text{a.s.}$$

Fix a sample point in the probability one set on which the convergence above holds for all i. Then, the convergence above leads to

$$\int_0^t\left(\int f_i(z)-\int f_i(y)\,p(dy\mid z,a,\tilde{\theta}(s))\right)\tilde{\mu}(s,dz\,da)\,ds=0,\quad \forall\, i. \tag{15}$$

Here, we use one part of the proof from Lemma 2.3 of Borkar [7] that if $\mu^n(\cdot)\to\mu^\infty(\cdot)\in\mathcal{U}$, then for any $t>0$,

$$\int_0^t\int\tilde{f}(s,z,a)\,\mu^n(s,dz\,da)\,ds-\int_0^t\int\tilde{f}(s,z,a)\,\mu^\infty(s,dz\,da)\,ds\to 0$$

for all $\tilde{f}\in C([0,t]\times S\times A)$ and the fact that $\tilde{f}_n(s,z,a)=\int f_i(y)p\,(dy\mid z,a,\bar{\theta}(s+u(n)))$ converges uniformly to $\tilde{f}(s,z,a)=\int f_i(y)p\,(dy\mid z,a,\tilde{\theta}(s))$. To prove the latter, define g: $C([0,t])\times[0,t]\times S\times A\to\mathbb{R}$ by $g(\theta(\cdot),s,z,a)=\int f_i(y)p\,(dy\mid z,a,\theta(s)))$. To see that g is continuous, we need to check that if $\theta_n(\cdot)\to\theta(\cdot)$ uniformly and $s(n)\to s$, then $\theta_n(s(n))\to\theta(s)$. This is because $\|\theta_n(s(n))-\theta(s)\|=\|\theta_n(s(n))-\theta(s(n))+\theta(s(n))-\theta(s)\|\leq\|\theta_n(s(n))-\theta(s(n))\|+\|\theta(s(n))-\theta(s)\|$. The first and second terms go to zero due to the uniform convergence of $\theta_n(\cdot)$, $n\geq 0$ and continuity of $\theta(\cdot)$, respectively. Let $A=\{\bar{\theta}(u(n)+\cdot)|_{[u(n),\,u(n)+t]},n\geq 1\}\cup\tilde{\theta}(\cdot)|_{[0,t]}$. A is compact as it is the union of a sequence of functions and their limit. Therefore $g|_{(A\times[0,t]\times S\times U)}$ is uniformly continuous. Then using the same arguments as in Lemma 4, we can show equicontinuity of $\{\tilde{f}_n(\cdot,\cdot)\}$, that results in uniform convergence and thereby (15). An application of Lebesgue's theorem in conjunction with (15) shows that

$$\int\left(f_i(z)-\int f_i(y)p\,(dy\mid z,a,\tilde{\theta}(t))\right)\tilde{\mu}\,(t,dz\,da)=0,\quad \forall\, i$$

for a.e. t. By our choice of $\{f_i\}$, this leads to

$$\tilde{\mu}(t,dy\times U)=\int p\,(dy\mid z,a,\tilde{\theta}(t))\,\tilde{\mu}(t,dz\,da)$$

a.e. t. Therefore the conclusion follows by disintegrating such measure as the product of marginal on S and the regular conditional law on U (Borkar [7, p. 140]). □

Remark 6. Note that the above invariant distribution does not come "naturally"; rather it arises from the assumption made to match the natural time-scale intuition for the controlled Markov noise component, i.e., the slower iterate should see the average effect of the Markov component.

The proof of the following lemma, in this case, will be unchanged from its original version, therefore we just mention it for completeness and refer the reader to Borkar [7, Lemma 2.3] for its proof.

Lemma 8. *Let $\mu^n(\cdot) \to \mu^\infty(\cdot) \in \mathcal{U}$. Let $\theta^n(\cdot)$, $n = 1, 2, \ldots, \infty$ denote solutions to (12) corresponding to the case where $\mu(\cdot)$ is replaced by $\mu^n(\cdot)$ for $n = 1, 2, \ldots \infty$. Suppose $\theta^n(0) \to \theta^\infty(0)$. Then,*

$$\lim_{n\to\infty} \sup_{t\in[0,T]} \|\theta^n(t) - \theta^\infty(t)\| = 0$$

for every $T > 0$.

Lemma 9. *Almost surely, $\{\theta_n\}$ converges to an internally chain transitive set of the differential inclusion*

$$\dot{\theta}(t) \in \hat{h}(\theta(t)), \tag{16}$$

where $\hat{h}(\theta) = \{\tilde{h}(\theta, \nu)\colon \nu \in D(\theta)\}$.

Proof. Lemma 8 shows that every limit point $(\tilde{\mu}(\cdot), \tilde{\theta}(\cdot))$ of $(\mu(s + \cdot), \bar{\theta}(s + \cdot))$ as $s \to \infty$ is such that $\tilde{\theta}(\cdot)$ satisfies (12) with $\mu(\cdot) = \tilde{\mu}(\cdot)$. Hence $\tilde{\theta}(\cdot)$ is absolutely continuous. Moreover, using Lemma 7, one can see that it satisfies (16) a.e. t, hence is a solution to the differential inclusion (16). Hence the proof follows. □

Lemma 10 (Faster Time-Scale Result)**.** *$(\theta_n, w_n) \to \{(\theta, \lambda(\theta))\colon \theta \in \mathbb{R}^d\}$, a.s.*

Proof. We first rewrite (3) as

$$\theta_{n+1} = \theta_n + b(n)[\epsilon_n + M^{(3)}_{n+1}],$$

where $\epsilon_n = (a(n)/b(n))h(\theta_n, w_n, Z^{(1)}_n) \to 0$ as $n \to \infty$, a.s. and $M^{(3)}_{n+1} = (a(n)/b(n))M^{(1)}_{n+1}$ for $n \geq 0$. Let $\alpha_n = (\theta_n, w_n)$, $\alpha = (\theta, w) \in \mathbb{R}^{d+k}$, $G(\alpha, z) = (0, g(\alpha, z))$, $\epsilon'_n = (\epsilon_n, 0)$, $M^{(4)}_{n+1} = (M^{(3)}_{n+1}, M^{(2)}_{n+1})$. Then, one can write (3) and (4) in the framework of (14) as

$$\alpha_{n+1} = \alpha_n + b(n)[G(\alpha_n, Z^{(2)}_n) + \epsilon'_n + M^{(4)}_{n+1}] \tag{17}$$

with $\epsilon'_n \to 0$ as $n \to \infty$. $\alpha_n, n \geq 0$ converges almost surely to an internally chain transitive set of the differential inclusion

$$\dot{\alpha}(t) \in \hat{G}(\alpha(t)),$$

where $\hat{G}(\alpha) = \{\tilde{G}(\alpha, \nu)\colon \nu \in D^{(2)}(\theta, w)\}$ with $\tilde{G}(\alpha, \nu) = (0, \tilde{g}(\theta, w, \nu))$. In other words, (θ_n, w_n), $n \geq 0$ converges to an internally chain transitive set of the differential inclusion

$$\dot{w}(t) \in \hat{g}(\theta(t), w(t)), \dot{\theta}(t) = 0.$$

The rest follows from the second part of **(A6)**. □

Remark 7. Under the conditions mentioned in Remark 4, the above faster time-scale result should be modified as follows.

Lemma 11 (Faster Time-Scale Result When $\lambda(\theta)$ Is a Local Attractor)**.** *Under assumptions* **(A1)–(A5)**, **(A6)′**, *and* **(A7)**, *on the event "$\{w_n\}$ belongs to a compact subset B (depending on the sample point) of $\bigcap_{\theta\in\mathbb{R}^d} G_\theta$ eventually,"*

$$(\theta_n, w_n) \to \{(\theta, \lambda(\theta))\colon \theta \in \mathbb{R}^d\}, \quad a.s.$$

Proof. Fix a sample point ω. The proof follows from these observations:

1. continuity of flow for the coupled o.d.e. around the initial point,
2. $\sup_n \|\theta_n\| = M_1 < \infty$,
3. the fact that the set graph(λ) is Lyapunov stable ($V'(\cdot)$ as mentioned in **(A6)′** will be a Lyapunov function for this set), and
4. the fact that $\bigcap_{t\geq 0} \overline{\bar{\alpha}(s)\colon s \geq t}$ is an internally chain transitive set of the coupled o.d.e.

$$\dot{w}(t) = \hat{g}(\theta(t), w(t)), \qquad \dot{\theta}(t) = 0, \tag{18}$$

where $\bar{\alpha}(\cdot)$ is the interpolated trajectory of the coupled iterate $\{\alpha_n\}$.
As $\{\theta\colon \|\theta\| \leq M_1\} \times B \subset \bigcup_{\theta\in\mathbb{R}^d} \{\{\theta\} \times G_\theta\}$, the first three observations show that for all $\epsilon > 0$, there exists a $T_\epsilon > 0$ such that any o.d.e. trajectory for (18) with starting point on the compact set $\{\theta\colon \|\theta\| \leq M_1\} \times B$ reaches the ϵ-neighborhood of graph(λ) after time T_ϵ. Further,

$$\bigcap_{t\geq 0} \overline{\bar{\alpha}(s)\colon s \geq t} \subset \{\theta\colon \|\theta\| \leq M_1\} \times B.$$

Then, one can use the last observation by choosing $T > T_\epsilon$ to show the required convergence to the set graph(λ). □

Remark 8. One interesting question in this context is to analyze whether one can extend the single time-scale local attractor convergence statements to the two time-scale setting under some *verifiable conditions*. More specifically, if there is a global attractor A_1 for

$$\dot{\theta}(t) \in \hat{h}(\theta(t)),$$

then one can provide verifiable conditions to show

$$P\Big[(\theta_n, w_n) \to \bigcup_{\theta \in A_1} (\theta, \lambda(\theta))\Big] > 0.$$

Here, $\lambda(\theta)$ is a local attractor as mentioned in **(A6)′**.

There are two ways in which this could possibly be tried:

1. Use Theorem 2 where we show that on the event $\{w_n\}$ belongs to a compact subset B (depending on the sample point) of $\bigcap_{\theta \in \mathbb{R}^d} G_\theta$ "eventually,"

$$(\theta_n, w_n) \to \bigcup_{\theta^* \in A_1} (\theta^*, \lambda(\theta^*)), \quad \text{a.s. as } n \to \infty,$$

which is an extension of the Kushner-Clarke lemma to the two time-scale case. Therefore the task would be to impose verifiable assumptions so that $P(\{w_n\}$ belongs to a compact subset B (depending on the sample point) of $\bigcap_{\theta \in \mathbb{R}^d} G_\theta$ "eventually"$) > 0$. In a stochastic approximation scenario, it is not immediately clear how one could possibly impose verifiable assumptions so that such a probabilistic statement becomes true.

2. The second approach would be to extend the analysis of Benaïm [2], Benaïm et al. [3] to the two time-scale case. In our opinion, this is very hard as this analysis is based on the attractor introduced by Benaim et al. whereas the coupled o.d.e. (18), which tracks the coupled iterate (θ_n, w_n) (therefore the interpolated trajectory of the coupled iterate will be an asymptotic pseudotrajectory (Benaïm [2]) for (18)) has no attractor. The reason is that one cannot obtain a fundamental neighborhood for sets like $\bigcup_{\theta \in A_1} (\theta, \lambda(\theta))$ as the θ component will remain constant for any trajectory of the above coupled o.d.e.

Thus, it is immediately not clear as to how this question can be addressed and this will be an interesting future direction.

From the faster time-scale results we get, $\|w_n - \lambda(\theta_n)\| \to 0$, a.s., i.e., $\{w_n\}$ asymptotically tracks $\{\lambda(\theta_n)\}$, a.s.

Now, consider the nonautonomous o.d.e.

$$\dot{\theta}(t) = \tilde{h}(\theta(t), \lambda(\theta(t)), \mu(t)), \tag{19}$$

where $\mu(t) = \delta_{Z_n^{(1)}, A_n^{(1)}}$ when $t \in [t(n), t(n+1))$ for $n \geq 0$ and $\tilde{h}(\theta, w, \nu) = \int h(\theta, w, z)\nu(dz)$. Let $\theta^s(t)$, $t \geq s$ denote the solution to (19) with $\theta^s(s) = \bar{\theta}(s)$ for $s \geq 0$. Then, we have the following lemma.

Lemma 12. *For any* $T > 0$, $\sup_{t \in [s, s+T]} \|\bar{\theta}(t) - \theta^s(t)\| \to 0$, *a.s.*

Proof. The slower recursion corresponds to

$$\theta_{n+1} = \theta_n + a(n)[h(\theta_n, w_n, Z_n^{(1)}) + M_{n+1}^{(1)}].$$

Let $t(n+m) \in [t(n), t(n)+T]$. Let $[t] = \max\{t(k)\colon t(k) \leq t\}$. Then, by construction,

$$\begin{aligned}
\bar{\theta}(t(n+m)) &= \bar{\theta}(t(n)) + \sum_{k=0}^{m-1} a(n+k) h(\bar{\theta}(t(n+k)), w_{n+k}, Z_{n+k}^{(1)}) + \delta_{n,n+m} \\
&= \bar{\theta}(t(n)) + \sum_{k=0}^{m-1} a(n+k) h(\bar{\theta}(t(n+k)), \lambda(\bar{\theta}(t(n+k))), Z_{n+k}^{(1)}) \\
&\quad + \sum_{k=0}^{m-1} a(n+k)(h(\bar{\theta}(t(n+k)), w_{n+k}, Z_{n+k}^{(1)}) - h(\bar{\theta}(t(n+k)), \lambda(\theta_{n+k}), Z_{n+k}^{(1)})) + \delta_{n,n+m},
\end{aligned}$$

where $\delta_{n,n+m} = \zeta_{n+m} - \zeta_n$ with $\zeta_n = \sum_{m=0}^{n-1} a(m) M_{m+1}^{(1)}$, $n \geq 1$.

$$\begin{aligned}
\theta^{t(n)}(t(m+n)) &= \bar{\theta}(t(n)) + \int_{t(n)}^{t(n+m)} \tilde{h}(\theta^{t(n)}(t), \lambda(\theta^{t(n)}(t)), \mu(t))\, dt \\
&= \bar{\theta}(t(n)) + \sum_{k=0}^{m-1} a(n+k) h(\theta^{t(n)}(t(n+k)), \lambda(\theta^{t(n)}(t(n+k))), Z_{n+k}^{(1)}) \\
&\quad + \int_{t(n)}^{t(n+m)} (h(\theta^{t(n)}(t), \lambda(\theta^{t(n)}(t), \mu(t))) - h(\theta^{t(n)}([t]), \lambda(\theta^{t(n)}([t]), \mu([t])))) \, dt.
\end{aligned}$$

Let $t(n) \le t \le t(n+m)$. Now, if $0 \le k \le (m-1)$ and $t \in (t(n+k), t(n+k+1)]$,

$$\begin{aligned}\|\theta^{t(n)}(t)\| &\le \|\bar{\theta}(t(n))\| + \left\|\int_{t(n)}^{t} \tilde{h}(\theta^{t(n)}(\tau), \lambda(\theta^{t(n)}(\tau)), \mu(\tau))\,d\tau\right\| \\ &\le \|\theta_n\| + \sum_{l=0}^{k-1}\int_{t(n+l)}^{t(n+l+1)} (\|h(0,0,Z^{(1)}_{n+l})\| + L^{(1)}(\|\lambda(0)\| + (K+1)\|\theta^{t(n)}(\tau)\|))\,d\tau \\ &\quad + \int_{t(n+k)}^{t} (\|h(0,0,Z^{(1)}_{n+k})\| + L^{(1)}(\|\lambda(0)\| + (K+1)\|\theta^{t(n)}(\tau)\|))\,d\tau \\ &\le C_0 + (M + L^{(1)}\|\lambda(0)\|)T + L^{(1)}(K+1)\int_{t(n)}^{t}\|\theta^{t(n)}(\tau)\|\,d\tau,\end{aligned}$$

where $C_0 = \sup_n \|\theta_n\| < \infty$, $\sup_{z\in S^{(1)}}\|h(0,0,z)\| = M$. By Gronwall's inequality, it follows that

$$\|\theta^{t(n)}(t)\| \le (C_0 + (M + L^{(1)}\|\lambda(0)\|)T)e^{L^{(1)}(K+1)T}$$

$$\begin{aligned}\|\theta^{t(n)}(t) - \theta^{t(n)}(t(n+k))\| &\le \int_{t(n+k)}^{t} \|h(\theta^{t(n)}(s), \lambda(\theta^{t(n)}(s)), Z^{(1)}_{n+k})\|\,ds \\ &\le (\|h(0,0,Z^{(1)}_{n+k})\| + L^{(1)}\|\lambda(0)\|)(t - t(n+k)) + L^{(1)}(K+1)\int_{t(n+k)}^{t}\|\theta^{t(n)}(s)\|\,ds \\ &\le C_T a(n+k),\end{aligned}$$

where $C_T = (M + L^{(1)}\|\lambda(0)\|) + L^{(1)}(K+1)(C_0 + (M + L^{(1)}\|\lambda(0)\|)T)e^{L^{(1)}(K+1)T}$. Thus

$$\begin{aligned}&\left\|\int_{t(n)}^{t(n+m)} (h(\theta^{t(n)}(t), \lambda(\theta^{t(n)}(t)), \mu(t)) - h(\theta^{t(n)}([t]), \lambda(\theta^{t(n)}([t])), \mu([t])))\,dt\right\| \\ &\le \sum_{k=0}^{m-1}\int_{t(n+k)}^{t(n+k+1)} \|h(\theta^{t(n)}(t), \lambda(\theta^{t(n)}(t)), Z^{(1)}_{n+k}) - h(\theta^{t(n)}([t]), \lambda(\theta^{t(n)}([t])), Z^{(1)}_{n+k})\|\,dt \\ &\le L\sum_{k=0}^{m-1}\int_{t(n+k)}^{t(n+k+1)} \|\theta^{t(n)}(t) - \theta^{t(n)}(t(n+k))\|\,dt \\ &\le C_T L\sum_{k=0}^{m-1} a(n+k)^2 \le C_T L\sum_{k=0}^{\infty} a(n+k)^2 \to 0, \quad \text{as } n\to\infty, \text{ where } L = L^{(1)}(K+1).\end{aligned}$$

Hence

$$\begin{aligned}\|\bar{\theta}(t(n+m))\| - \theta^{t(n)}(t(n+m)) &\le L\sum_{k=0}^{m-1} a(n+k)\|\bar{\theta}(t(n+k)) - \theta^{t(n)}(t(n+k))\| \\ &\quad + C_T L\sum_{k=0}^{\infty} a(n+k)^2 + \sup_{k\ge 0}\|\delta_{n,n+k}\| + L^{(1)}\sum_{k=0}^{m-1} a(n+k)\|w_{n+k} - \lambda(\theta_{n+k})\| \\ &\le L\sum_{k=0}^{m-1} a(n+k)\|\bar{\theta}(t(n+k)) - \theta^{t(n)}(t(n+k))\| \\ &\quad + C_T L\sum_{k=0}^{\infty} a(n+k)^2 + \sup_{k\ge 0}\|\delta_{n,n+k}\| + L^{(1)}T\sup_{k\ge 0}\|w_{n+k} - \lambda(\theta_{n+k})\|, \quad \text{a.s.}\end{aligned}$$

Define

$$K_{T,n} = C_T L\sum_{k=0}^{\infty} a(n+k)^2 + \sup_{k\ge 0}\|\delta_{n,n+k}\| + L^{(1)}T\sup_{k\ge 0}\|w_{n+k} - \lambda(\theta_{n+k})\|.$$

Note that $K_{T,n} \to 0$, a.s. The remainder of the proof follows in the exact same manner as the tracking lemma, see Borkar [8, Lemma 1, Chapter 2]. □

Lemma 13. *Suppose $\mu^n(\cdot) \to \mu^\infty(\cdot) \in U^{(1)}$. Let $\theta^n(\cdot)$, $n = 1, 2, \ldots, \infty$ denote solutions to (19) corresponding to the case where $\mu(\cdot)$ is replaced by $\mu^n(\cdot)$ for $n = 1, 2, \ldots, \infty$. Suppose $\theta^n(0) \to \theta^\infty(0)$. Then,*

$$\lim_{n\to\infty}\sup_{t\in[0,T]}\|\theta^n(t) - \theta^\infty(t)\| \to 0$$

for every $T > 0$.

Proof. It is shown in Borkar [7, Lemma 2.3] that

$$\int_0^t \int \tilde{f}(s,z)\mu^n(s,dz)\,ds - \int_0^t \int \tilde{f}(s,z)\mu^\infty(s,dz)\,ds \to 0$$

for any $\tilde{f} \in C([0,T]\times S)$. Using this, one can see that

$$\left\| \int_0^t (\tilde{h}(\theta^\infty(s),\lambda(\theta^\infty(s)),\mu^n(s)) - \tilde{h}(\theta^\infty(s),\lambda(\theta^\infty(s)),\mu^\infty(s)))\,ds \right\| \to 0.$$

This follows because λ is continuous and h is jointly continuous in its arguments. As a function of t, the integral on the left is equicontinuous and pointwise bounded. By the Arzela-Ascoli theorem, this convergence must, in fact, be uniform for t in a compact set. Now, for $t>0$,

$$\begin{aligned}\|\theta^n(t)-\theta^\infty(t)\| &\le \|\theta^n(0)-\theta^\infty(0)\| + \int_0^t \|\tilde{h}(\theta^n(s),\lambda(\theta^n(s)),\mu^n(s)) - \tilde{h}(\theta^\infty(s),\lambda(\theta^\infty(s)),\mu^\infty(s))\|\,ds \\ &\le \|\theta^n(0)-\theta^\infty(0)\| + \int_0^t (\|\tilde{h}(\theta^n(s),\lambda(\theta^n(s)),\mu^n(s)) - \tilde{h}(\theta^\infty(s),\lambda(\theta^\infty(s)),\mu^n(s))\|)\,ds \\ &\quad + \int_0^t (\|\tilde{h}(\theta^\infty(s),\lambda(\theta^\infty(s)),\mu^n(s)) - \tilde{h}(\theta^\infty(s),\lambda(\theta^\infty(s)),\mu^\infty(s))\|)\,ds.\end{aligned}$$

Now, using the fact that λ is Lipschitz with constant K, the remaining part of the proof follows in the same manner as Borkar [7, Lemma 2.3]. □

Note that Lemma 13 shows that every limit point $(\tilde{\mu}(\cdot),\tilde{\theta}(\cdot))$ of $(\mu(s+\cdot),\bar{\theta}(s+\cdot))$ as $s\to\infty$ is such that $\tilde{\theta}(\cdot)$ satisfies (19) with $\mu(\cdot)=\tilde{\mu}(\cdot)$.

Lemma 14. *Almost surely every limit point of* $(\mu(s+\cdot),\bar{\theta}(s+\cdot))$ *as* $s\to\infty$ *is of the form* $(\tilde{\mu}(\cdot),\tilde{\theta}(\cdot))$*, where* $\tilde{\mu}(\cdot)$ *satisfies* $\tilde{\mu}(t)\in D^{(1)}(\tilde{\theta}(t),\lambda(\tilde{\theta}(t)))$.

Proof. Suppose that $u(n)\uparrow\infty$, $\mu(u(n)+\cdot)\to\tilde{\mu}(\cdot)$ and $\bar{\theta}(u(n)+\cdot)\to\tilde{\theta}(\cdot)$. Let $\{f_i\}$ be countable dense in the unit ball of $C(S)$, hence it is a separating class, i.e., for all i, $\int f_i\,d\mu=\int f_i\,d\nu$ implies $\mu=\nu$. For each i,

$$\zeta_n^i = \sum_{m=1}^{n-1} a(m)\left(f_i(Z_{m+1}^{(1)}) - \int f_i(y)\,p(dy\mid Z_m^{(1)},A_m^{(1)},\theta_m,w_m)\right),$$

is a zero mean martingale with $\mathcal{F}_n=\sigma(\theta_m,w_m,Z_m^{(1)},A_m^{(1)},m\le n)$, $n\ge 1$. Moreover, it is a square integrable martingale because f_i's are bounded and each ζ_n^i is a finite sum. Its quadratic variation process

$$A_n = \sum_{m=0}^{n-1} a(m)^2 E\left[\left(f_i(Z_{m+1}^{(1)}) - \int f_i(y)\,p(dy\mid Z_m^{(1)},A_m^{(1)},\theta_m,w_m)\right)^2 \,\middle|\, \mathcal{F}_m\right] + E[(\zeta_0^i)^2]$$

is almost surely convergent. By the martingale convergence theorem, $\{\zeta_n^i\}$ converges a.s. Let $\tau(n,t)=\min\{m\ge n\colon t(m)\ge t(n)+t\}$ for $t\ge 0, n\ge 0$. Then as $n\to\infty$,

$$\sum_{m=n}^{\tau(n,t)} a(m)\left(f_i(Z_{m+1}^{(1)}) - \int f_i(y)\,p(dy\mid Z_m^{(1)},A_m^{(1)},\theta_m,w_m)\right) \to 0, \quad \text{a.s.},$$

for $t>0$. By our choice of $\{f_i\}$ and the fact that $\{a(n)\}$ are eventually nonincreasing,

$$\sum_{m=n}^{\tau(n,t)} (a(m)-a(m+1))f_i(Z_{m+1}^{(1)}) \to 0, \quad \text{a.s.}$$

Thus

$$\sum_{m=n}^{\tau(n,t)} a(m)\left(f_i(Z_m^{(1)}) - \int f_i(y)\,p(dy\mid Z_m^{(1)},A_m^{(1)},\theta_m,w_m)\right) \to 0, \quad \text{a.s.},$$

which implies

$$\int_{t(n)}^{t(n)+t} \left(\int \left(f_i(z) - \int f_i(y)\,p(dy\mid z,a,\hat{\theta}(s),\hat{w}(s))\right)\mu(s,dz\,da)\right)ds \to 0, \quad \text{a.s.}$$

Recall that $u(n)$ can be any general sequence other than $t(n)$. Therefore

$$\int_{u(n)}^{u(n)+t}\left(\int\left(f_i(z)-\int f_i(y)\,p(dy\mid z,a,\hat{\theta}(s),\hat{w}(s))\right)\mu(s,dz\,da)\right)ds\to 0,\quad \text{a.s.}$$

(this follows from the fact that $a(n)\to 0$ and f_i's are bounded), where $\hat{\theta}(s)=\theta_n$ and $\hat{w}(s)=w_n$ when $s\in[t(n),t(n+1)), n\geq 0$. Now, one can claim from the above that

$$\int_{u(n)}^{u(n)+t}\left(\int\left(f_i(z)-\int f_i(y)\,p(dy\mid z,a,\bar{\theta}(s),\lambda(\bar{\theta}(s)))\right)\mu(s,dzda)\right)ds\to 0,\quad \text{a.s.}$$

This is because the map $S^{(1)}\times U^{(1)}\times\mathbb{R}^{d+k}\ni(z,a,\theta,w)\to\int f_i(y)\,p(dy\mid z,a,\theta,w)$ is continuous and hence uniformly continuous on the compact set $A=S^{(1)}\times U^{(1)}\times M_1\times M_2$, where M_1 is the compact set s.t. $\theta_n\in M_1$ for all n and $M_2=\{w\colon \|w\|\leq\max(\sup\|w_n\|,K')\}$, where K' is the bound for the compact set $\lambda(M_1)$. Here, we also use the fact that $\|w_m-\lambda(\bar{\theta}(s))\|\to 0$ for $s\in[t_m,t_{m+1})$ as λ is Lipschitz and $\|w_m-\lambda(\theta_m)\|\to 0$. The above convergence is equivalent to

$$\int_0^t\left(\int\left(f_i(z)-\int f_i(y)\,p(dy\mid z,a,\bar{\theta}(s+u(n)),\lambda(\bar{\theta}(s+u(n))))\right)\mu(s+u(n),dz\,da)\right)ds\to 0,\quad\text{a.s.}$$

Fix a sample point in the probability one set on which the convergence above holds for all i. Then, the convergence above leads to

$$\int_0^t\left(\int f_i(z)-\int f_i(y)\,p(dy\mid z,a,\tilde{\theta}(s),\lambda(\tilde{\theta}(s)))\right)\tilde{\mu}(s,dz\,da)\,ds=0,\quad\forall\, i. \tag{20}$$

For showing the above, we use one part of the proof from Borkar [7, Lemma 2.3] that if $\mu^n(\cdot)\to\mu^\infty(\cdot)\in\mathcal{U}$, then for any t,

$$\int_0^t\int\tilde{f}(s,z,a)\,\mu^n(s,dz\,da)\,ds-\int_0^t\int\tilde{f}(s,z,a)\mu^\infty(s,dz\,da)\,ds\to 0$$

for all $\tilde{f}\in C([0,t]\times S^{(1)}\times U^{(1)})$. In addition, we make use of the fact that $\tilde{f}_n(s,z,a)=\int f_i(y)p(dy\mid z,a,\bar{\theta}(s+u(n)),\lambda(\bar{\theta}(s+u(n))))$ converges uniformly to $\tilde{f}(s,z,a)=\int f_i(y)p(dy\mid z,a,\tilde{\theta}(s),\lambda(\tilde{\theta}(s)))$. To prove this, define $g\colon C([0,t])\times[0,t]\times S^{(1)}\times U^{(1)}\to\mathbb{R}$ by $g(\theta(\cdot),s,z,a)=\int f_i(y)p(dy\mid z,a,\theta(s),\lambda(\theta(s)))$. Let $A'=\{\bar{\theta}(u(n)+\cdot)|_{[u(n),u(n)+t]},n\geq 1\}\cup\tilde{\theta}(\cdot)|_{[0,t]}$. Using the same argument as in Lemma 7 and **(A6)**, i.e., λ is Lipschitz (the latter helps to claim that if $\theta_n(\cdot)\to\theta(\cdot)$ uniformly, then $\lambda(\theta_n(\cdot))\to\lambda(\theta(\cdot))$ uniformly), it can be seen that g is continuous. Then, A' is compact as it is a union of a sequence of functions and its limit. Therefore $g|_{(A'\times[0,t]\times S^{(1)}\times U^{(1)})}$ is uniformly continuous. Then, a similar argument as in Lemma 4 shows equicontinuity of $\{\tilde{f}_n(\cdot,\cdot)\}$ that results in uniform convergence and thereby (20). An application of Lebesgue's theorem in conjunction with (20) shows that

$$\int\left(f_i(z)-\int f_i(y)p\,(dy\mid z,a,\tilde{\theta}(t),\lambda(\tilde{\theta}(t)))\right)\tilde{\mu}(t,dzda)=0,\quad\forall\, i$$

for a.e. t. By our choice of $\{f_i\}$, this leads to

$$\tilde{\mu}(t,dy\times U^{(1)})=\int p\,(dy\mid z,a,\tilde{\theta}(t),\lambda(\tilde{\theta}(t)))\tilde{\mu}(t,dz\,da),\quad\text{a.e. } t.\quad\square$$

Lemma 13 shows that every limit point $(\tilde{\mu}(\cdot),\tilde{\theta}(\cdot))$ of $(\mu(s+\cdot),\bar{\theta}(s+\cdot))$ as $s\to\infty$ is such that $\tilde{\theta}(\cdot)$ satisfies (19) with $\mu(\cdot)=\tilde{\mu}(\cdot)$. Hence $\tilde{\theta}(\cdot)$ is absolutely continuous. Moreover, using Lemma 14, one can see that it satisfies (9) a.e. t, hence is a solution to the differential inclusion (9).

Proof of Theorems 1 and 2. From the previous three lemmas, it is easy to see that $A_0=\bigcap_{t\geq 0}\overline{\{\bar{\theta}(s)\colon s\geq t\}}$ is almost everywhere an internally chain transitive set of (9). $\square$

Proof of Corollary 1. Follows directly from Theorem 1 and Lemma 1. $\square$

4. Discussion on the Assumptions: Relaxation of (A2)

We discuss relaxation of the uniformity of the Lipschitz constant w.r.t. state of the controlled Markov process for the vector field. The modified assumption here is

(A2)′ $h: \mathbb{R}^{d+k} \times S^{(1)} \to \mathbb{R}^d$ is jointly continuous as well as Lipschitz in its first two arguments with the third argument fixed to same value and Lipschitz constant is a function of this value. The latter condition means that

$$\forall z^{(1)} \in S^{(1)} \quad \|h(\theta, w, z^{(1)}) - h(\theta', w', z^{(1)})\| \le L^{(1)}(z^{(1)})(\|\theta - \theta'\| + \|w - w'\|).$$

A similar condition holds for g, where the Lipschitz constant is $L^{(2)}: S^{(2)} \to \mathbb{R}^+$.

Note that this allows $L^{(i)}(\cdot)$ to be an unbounded measurable function making it discontinuous due to **(A1)**. The straightforward solution for implementing this is to additionally assume the following:

(A8) $\sup_n L^{(i)}(Z_n^{(i)}) < \infty$, a.s., still allowing $L^{(i)}(\cdot)$ to be an unbounded function. As all our proofs in Section 3 are shown for every sample point of a probability 1 set, our proofs will go through. In the following, we give such an example for the case where the Markov process is uncontrolled.

It is enough to consider examples with locally compact $S^{(i)}$ (because then we can take the standard one-point compactification and define $L^{(i)}$ arbitrarily at the extra point).

Let $S^{(i)} = \mathbb{Z}$ and let $Z_n^{(i)}$, $n \ge 0$ be the Markov chain on $\mathbb{Z}$ starting at 0 with transition probabilities $p(n, n+1) = p$ and $p(n, n-1) = 1 - p$. We assume $1/2 < p < 1$. Let $L^{(i)}(n) = ((1-p)/p)^n$.

Note that $Z_n^{(i)}$, $n \ge 0$ is a transient Markov chain with $Z_n^{(i)} \to +\infty$, a.s. From this, it follows that $\inf_n Z_n^{(i)} > -\infty$, and thus $\sup_n L^{(i)}(Z_n^{(i)}) < \infty$ almost surely. It follows that $(L^{(i)}(Z_n^{(i)}))_{n\in\mathbb{N}}$ is a bounded sequence with probability 1, but this bound is clearly not deterministic since there is a nonzero probability that the sample path reaches large negative values.

However, in the following, we discuss on the idea of using moment assumptions to analyze the convergence of single time-scale controlled Markov noise framework of Borkar [7]. We show that the iterates (14) (with $\epsilon_n = 0$) converge to an internally chain transitive set of the o.d.e. (12). For this we prove Lemma 6 under the following assumptions: for all $T > 2$, $i = 1, 2$,

(S1) The controlled Markov process Y_n as described in Borkar [7] takes values in a compact metric space.

(S2) For all $n > 0$, $0 < a(n) \le 1$, $\sum_n a(n) = \infty$, $\sum_n a(n)^2 < \infty$, and $a(n+1) \le a(n)$, $n \ge 0$.

(S3) $h: \mathbb{R}^d \times S \to \mathbb{R}^d$ Lipschitz in its first argument, w.r.t. the second. The condition means that

$$\forall z \in S, \quad \|h(\theta, z) - h(\theta', z)\| \le L(z)(\|\theta - \theta'\|).$$

Let $\phi(n, T) = \max(m: a(n) + a(n+1) + \cdots + a(n+m) \le T)$ with the bound depending on T. With this notation we assume the following two moment assumptions:

(S4)
$$\sup_n E\left[\left(\sup_{0\le m\le\phi(n,T)} L(Y_{n+m})\right)^{16}\right] < \infty.$$

(S5)
$$\sup_n E\left[e^{8\sum_{m=0}^{\phi(n,T)} a(n+m) L(Y_{n+m})}\right] < \infty.$$

Note that **(S4)** and **(S5)** are trivially satisfied in the case when $L(z) = L$ for all $z \in S$, i.e., the case of Section 2.

Remark 9. As long as one can prove Lemma 6 for all $T > 2$ it will hold for all $T > 0$, thus one can combine **(S4)** and **(S5)** into the following assumption:

$$\sup_n E[e^{8T \sup_{0\le m\le\phi(n,T)} L(Y_{n+m})}] < \infty.$$

As an instance where such an assumption is verified, consider the Markov process of Metivier and Priouret [15, Equation (3.4)] defined by

$$Y_{n+1} = A(\theta_n) Y_n + B(\theta_n) W_{n+1},$$

where $A(\theta)$, $B(\theta)$, $\theta \in \mathbb{R}^d$, are $k \times k$-matrices and $(W_n)_{n\ge O}$ are independent and identically distributed $\mathbb{R}^k$-valued random variables. Assume that the following conditions hold true for all $x, y \in S$:

(a) $L(Y_n)$ is a nondecreasing sequence.

(b) For $r > 0, R > 0$,

$$\sup_{\|\theta\|\le R} e^{rL(A(\theta)x + B(\theta)y)} \le L_R {\alpha_R}^r e^{rL(x)} + M_R e^{C_R L(y)}$$

for some $C_R, M_R, L_R > 0$, and $\alpha_R < 1$.

Then,

$$E[e^{rL(Y_n)} \mid Y_{n-1} = x, \theta_{n-1} = \theta] \le \int e^{rL(A(\theta)x + B(\theta)y)} \mu_n(dy) \le L_R {\alpha_R}^r e^{rL(x)} + M_R E[e^{C_R L(W_n)}] = L_R {\alpha_R}^r e^{rL(x)} + K_R,$$

with $K_R = M_R E[e^{C_R L(W_n)}]$ (this follows from the fact that W_n are i.i.d. if we assume that $E[e^{C_R L(W_1)}] < \infty$). Choosing large values of r, one can show that

$$E[e^{rL(Y_n)} \mid Y_{n-1} = x, \theta_{n-1} = \theta] \le \beta_R e^{rL(x)} + K_R,$$

where $\beta_R = L_R {\alpha_R}^r < 1$. Using the above, for large r,

$$E[e^{rL(Y_n)}] = E[E[e^{rL(Y_n)} \mid Y_{n-1}, \theta_{n-1}]] \le \beta_R E[e^{rL(Y_{n-1})}] + K_R,$$

which shows that

$$\sup_n E[e^{rL(Y_n)}] < \infty.$$

Choosing $r > 8T$,

$$\sup_n E[e^{8TL(Y_n)}] < \infty.$$

Note that this is a much weaker assumption that **(A8)**.

(S6) The noise sequence M_n, $n \ge 0$ (need not be a martingale difference sequence) satisfies the following condition:

$$\sup_n E\left[\left(\sum_{m=0}^{\phi(n,T)} \|M_{n+m+1}\|\right)^4\right] < \infty.$$

(S7) $\sup_n \|\theta_n\| < \infty$.

With the above assumptions, we prove the following tracking lemma.

Lemma 15. *For any $T > 0$, $\sup_{t \in [s, s+T]} \|\bar{\theta}(t) - \theta^s(t)\| \to 0$,, a.s.*

Proof. Let $t(n) \le t \le t(n+m)$. Now, if $0 \le k \le (m-1)$ and $t \in (t(n+k), t(n+k+1)]$,

$$\begin{aligned}
\|\theta^{t(n)}(t)\| &\le \|\bar{\theta}(t(n))\| + \left\|\int_{t(n)}^{t} \tilde{h}(\theta^{t(n)}(\tau), \mu(\tau))\, d\tau\right\| \\
&\le \|\theta_n\| + \sum_{l=0}^{k-1} \int_{t(n+l)}^{t(n+l+1)} (\|h(0, Y_{n+l})\| + L(Y_{n+l})\|\theta^{t(n)}(\tau)\|)\, d\tau + \int_{t(n+k)}^{t} (\|h(0, Y_{n+k})\| + L(Y_{n+k})\|\theta^{t(n)}(\tau)\|)\, d\tau \\
&\le C_0 + MT + \int_{t(n)}^{t} L(Y(\tau))\|\theta^{t(n)}(\tau)\|\, d\tau
\end{aligned}$$

where $Y(\tau) = Y_n$ if $\tau \in [t(n), t(n+1))$. Then, it follows from an application of Gronwall inequality that

$$\|\theta^{t(n)}(t)\| \le C e^{\int_{t(n)}^{t} L(Y(\tau))\, d\tau} \quad \text{a.e. } t,$$

where $C = C_0 + MT$. Next,

$$\begin{aligned}
\|\theta^{t(n)}(t) - \theta^{t(n)}(t(n+k))\| &\le \int_{t(n+k)}^{t} \|h(\theta^{t(n)}(s), Y_{n+k})\|\, ds \\
&\le \|h(0, Y_{n+k})\|(t - t(n+k)) + L(Y_{n+k}) \int_{t(n+k)}^{t} \|\theta^{t(n)}(s)\|\, ds \\
&\le Ma(n+k) + CL(Y_{n+k}) \int_{t(n+k)}^{t} e^{\int_{t(n)}^{s} L(Y(\tau))\, d\tau}\, ds.
\end{aligned}$$

Then,

$$\begin{aligned}
&\left\|\int_{t(n)}^{t(n+m)} (h(\theta^{t(n)}(t), \mu(t)) - h(\theta^{t(n)}([t]), \mu([t])))\, dt\right\| \\
&\quad \le \sum_{k=0}^{m-1} \int_{t(n+k)}^{t(n+k+1)} \|h(\theta^{t(n)}(t), Y_{n+k}) - h(\theta^{t(n)}([t]), Y_{n+k})\|\, dt \\
&\quad \le \sum_{k=0}^{m-1} L(Y_{n+k}) \int_{t(n+k)}^{t(n+k+1)} \|\theta^{t(n)}(t) - \theta^{t(n)}(t(n+k))\|\, dt \le \sum_{k=0}^{m-1} c_k,
\end{aligned}$$

where

$$c_k = L(Y_{n+k})a(n+k)^2[M + CL(Y_{n+k})e^{\sum_{i=0}^{k} a(n+i)L(Y_{n+i})}].$$

$$\|\bar\theta(t(n+m)) - \theta^{t(n)}(t(n+m))\| \le \sum_{k=0}^{m-1} L(Y_{n+k})a(n+k)\|\bar\theta(t(n+k)) - \theta^{t(n)}(t(n+k))\| + \sum_{k=0}^{m-1} c_k + \|\delta_{n,n+m}\|,$$

where $\delta_{n,n+m} = \sum_{k=n}^{n+m-1} a(k)M_{k+1}$.

Therefore using discrete Gronwall inequality, we get

$$\|\bar\theta(t(n+m)) - \theta^{t(n)}(t(n+m))\| \le r(m,n)e^{\sum_{k=0}^{m-1} a(n+k)L(Y_{n+k})}$$

where $r(m,n) = \sum_{k=0}^{m-1}(c_k + a(n+k)\|M_{n+k+1}\|)$.

Now, for some $\lambda \in [0,1]$,

$$\begin{aligned}&\|\theta^{t(n)}(t) - \bar\theta(t)\|\\ &\quad\le (1-\lambda)\|\theta^{t(n)}(t(n+m+1)) - \bar\theta(t(n+m+1))\| + \lambda\|\theta^{t(n)}(t(n+m)) - \bar\theta(t(n+m))\|\\ &\qquad + \max(\lambda, 1-\lambda)\int_{t(n+m)}^{t(n+m+1)} \|\tilde h(\theta^{t(n)}(s), \mu(s))\|\,ds\\ &\quad\le r(m+1,n)e^{\sum_{k=0}^{m} a(n+k)L(Y_{n+k})} + a(n+m)[M + CL(Y_{n+m})e^{\sum_{k=0}^{m} a(n+k)L(Y_{n+k})}].\end{aligned}$$

Therefore

$$\begin{aligned}\rho(n,T) &:= \sup_{t\in[t(n),t(n)+T]} \|\theta^{t(n)}(t) - \bar\theta(t)\|\\ &\le r(\phi(n,T+1),n)e^{\sum_{k=0}^{\phi(n,T)} a(n+k)L(Y_{n+k})} + a(n)\Big[M + C\sup_{0\le m\le\phi(n,T)} L(Y_{n+m})e^{\sum_{k=0}^{\phi(n,T)} a(n+k)L(Y_{n+k})}\Big].\end{aligned}$$

Now, to prove the a.s. convergence of the quantity in the left-hand side as $n\to\infty$, we have using Cauchy-Schwartz inequality:

$$\begin{aligned}\sum_{n=1}^{\infty} E[\rho(n,T)^2] &\le 2K_T\sum_{n=1}^{\infty}(E[(r(\phi(n,T+1),n))^4])^{1/2} + 4M^2\sum_{n=0}^{\infty} a(n)^2\\ &\quad + 4C^2\sum_{n=1}^{\infty} a(n)^2 E\Big[\Big(\sup_{0\le m\le\phi(n,T)} L(Y_{n+m})\Big)^2 e^{2\sum_{k=0}^{\phi(n,T)} a(n+k)L(Y_{n+k})}\Big],\end{aligned}$$

where $K_T = \sqrt{\sup_n E[e^{4\sum_{k=0}^{\phi(n,T)} a(n+k)L(Y_{n+k})}]}$, which depends only on T due to **(S5)**. Now, the third term in the RHS is clearly finite from the assumptions **(S4)** and **(S5)**. Now, we analyze the first term, i.e.,

$$\sum_{n=1}^{\infty}(E[r(\phi(n,T+1),n)^4])^{1/2} \le 2\sqrt{2}\sum_{n=1}^{\infty}\Big(E\Big[\Big(\sum_{k=0}^{\phi(n,T)} c_k\Big)^4\Big]\Big)^{1/2} + 2\sqrt{2}\sum_{n=1}^{\infty}\Big(E\Big[\Big(\sum_{k=0}^{\phi(n,T)} a(n+k)\|M_{n+k+1}\|\Big)^4\Big]\Big)^{1/2}. \tag{21}$$

Next, we analyze the first term in the RHS of (21) again using Cauchy-Schwartz inequality:

$$\begin{aligned}\sum_{n=1}^{\infty}\Big(E\Big[\Big(\sum_{k=0}^{\phi(n,T)} c_k\Big)^4\Big]\Big)^{1/2} &\le 8M^2\sum_{n=1}^{\infty}\phi(n,T)^2 a(n)^4\Big(E\Big[\Big(\sup_{0\le k\le\phi(n,T)} L(Y_{n+k})\Big)^4\Big]\Big)^{1/2}\\ &\quad + 8C^2\sum_{n=1}^{\infty}\phi(n,T)^2 a(n)^4\Big(E\Big[\Big(\sup_{0\le k\le\phi(n,T)} L(Y_{n+k})\Big)^8 e^{4\sum_{i=0}^{\phi(n,T)} a(n+i)L(Y_{n+i})}\Big]\Big)^{1/2}.\end{aligned}$$

Therefore the the RHS will be finite if we can show that $\sum_{n=1}^{\infty}\phi(n,T)^2 a(n)^4$ is finite. For common stepsize sequence $a(n) = 1/n$, $\phi(n,T) = O(n)$ thus the above series converges clearly. One can make the series converge for all $a(n) = 1/n^k$ with $\frac{1}{2} < k \le 1$ by putting assumptions on higher moments in **(S4)** and **(S5)**.

In the above, we have used the following inequality repeatedly for nonnegative random variables X and Y:

$$\sqrt{E[(X+Y)^{2^n}]} \le 2^{(2n-1)/2}\big[\sqrt{E[X^{2^n}]} + \sqrt{E[Y^{2^n}]}\big]$$

with $n\in\mathbb{N}$.

Now,

$$\sum_{n=1}^{\infty}\Big(E\Big[\Big(\sum_{k=0}^{\phi(n,T)} a(n+k)\|M_{n+k+1}\|\Big)^4\Big]\Big)^{1/2} \le \sum_{n=1}^{\infty} a(n)^2\Big(E\Big[\Big(\sum_{k=0}^{\phi(n,T)}\|M_{n+k+1}\|\Big)^4\Big]\Big)^{1/2},$$

which is finite under assumption **(S5)** and the fact that $a(n)$ are nonincreasing. □

5. Application: Off-Policy Temporal-Difference Learning with Linear Function Approximation

In this section, we present an application of our results in the setting of off-policy Temporal-difference learning with linear function approximation. In this framework, we need to estimate the value function for a target policy π given the continuing evolution of the underlying Markov decision process (MDP) (with finite state and action spaces S and A respectively, specified by expected reward $r(\cdot,\cdot,\cdot)$ and transition probability kernel $p(\cdot\mid\cdot,\cdot)$) for a behavior policy π_b with $\pi\neq\pi_b$. The authors of Sutton et al. [17, 18], Maei [13] have proposed two approaches to solve the problem:

(i) Subsampling: In this approach, the transitions that are relevant to deterministic target policy are kept and the rest of the data is discarded from the given "on-policy" trajectory. We use the triplet (S,R,S') to represent (current state, reward, next state). Therefore one has "off-policy" data $(X'_n,R_n,W_n), n\geq 0$, where $E[R_n\mid X'_n=s, W_n=s']=r(s,a,s')$, $P(W_n=s'\mid X'_n=s)=p(s'\mid s,a)$ with $\pi(s)=a$, π being the target policy and X'_n, $n\geq 0$ is a random process generated by sampling the "on-policy" trajectory at increasing stopping times.

(ii) Importance weighting: In this approach, unlike subsampling, all the data from the given "on-policy" trajectory is used. One advantage of this method is that we can allow the policy to be randomized in case of behavior and target policies unlike the subsampling scenario, where one can use only deterministic policy as a target policy.

Then, they introduce GTD learning algorithms (GTD) (Sutton et al. [17, 18], Maei [13]) for both approaches.

Currently, all GTD algorithms make the assumption that data is available in the "off-policy" setting, i.e., of the form (X'_n,R_n,W_n), $n\geq 0$, where $\{X'_n\}$ are i.i.d., $E[R_n\mid X'_n=s,W_n=s']=r(s,a,s')$ and $P(W_n=s'\mid X'_n=s)=p(s'\mid s,a)$ with $\pi(s)=a$, π being the deterministic target policy. Additionally, the distribution of $\{X'_n\}$ is assumed to be sampled according to the stationary distribution of the Markov chain corresponding to the behavior policy. However, such data cannot be generated from subsampling given only the "on-policy" trajectory. The reason is that a Markov chain sampled at increasing stopping times cannot be i.i.d. In the following, we show how GTD learning along with importance weighting can be used to solve the off-policy convergence problem stated above for TD when only the "on-policy" trajectory is available.

5.1. Problem Definition

Suppose we are given an on-policy trajectory (X_n,A_n,R_n,X_{n+1}), $n\geq 0$, where $\{X_n\}$ is a time-homogeneous irreducible Markov chain with unique stationary distribution ν and generated from a behavior policy $\pi_b\neq\pi$. Here, the quadruplet (S,A,R,S') represents (current state, action, reward, next state). Also, assume that $\pi_b(a\mid s)>0$ for all $s\in S$, $a\in A$. We need to find the solution θ^* for the following:

$$0=\sum_{s,a,s'}\nu(s)\pi(a\mid s)p(s'\mid s,a)\delta(\theta;s,a,s')\phi(s)=E[\rho_{X,A_n}\delta_{X,R_n,X_{n+1}}(\theta)\phi(X)]=b-A\theta, \tag{22}$$

where

(i) $\theta\in\mathbb{R}^d$ is the parameter for value function,
(ii) $\phi\colon S\to\mathbb{R}^d$ is a vector of state features,
(iii) $X\sim\nu$,
(iv) $0<\gamma<1$ is the discount factor,
(v) $E[R_n\mid X_n=s,X_{n+1}=s']=\sum_{a\in A}\pi_b(a\mid s)r(s,a,s')$,
(vi) $P(X_{n+1}=s'\mid X=s)=\sum_{a\in A}\pi_b(a\mid s)p(s'\mid s,a)$,
(vii) $\delta(\theta;s,a,s')=r(s,a,s')+\gamma\theta^T\phi(s')-\theta^T\phi(s)$ is the temporal-difference term with expected reward,
(viii) $\rho_{X,A_n}=\pi(A_n\mid X)/\pi_b(A_n\mid X)$,
(ix) $\delta_{X,R_n,X_{n+1}}=R_n+\gamma\theta^T\phi(X_{n+1})-\theta^T\phi(X)$ is the online temporal difference,
(x) $A=E[\rho_{X,A_n}\phi(X)(\phi(X)-\gamma\phi(X_{n+1}))^T]$,
(xi) $b=E[\rho_{X,A_n}R_n\phi(X)]$.

Hence, the desired approximate value function under the target policy π is $V_\pi^*=\theta^{*T}\phi$. Let $V_\theta=\theta^T\phi$. It is well known (Maei [13]) that θ^* satisfies the projected fixed-point equation; namely,

$$V_\theta=\Pi_{\mathcal{G},\nu}T^\pi V_\theta,$$

where

$$\Pi_{\mathcal{G},\nu}\hat{V}=\arg\min_{f\in\mathcal{G}}(\|\hat{V}-f\|_\nu)$$

with $\mathcal{G} = \{V_\theta \mid \theta \in \mathbb{R}^d\}$ and the Bellman operator

$$T^\pi V_\theta(i) = \sum_{j\in S}\sum_{a\in A}\pi(a \mid i)p(j \mid i,a)[\gamma V_\theta(i) + r(i,a,j)].$$

Therefore, to find θ^*, the idea is to minimize the mean square projected Bellman error $J(\theta) = \|V_\theta - \Pi_{\mathcal{G},\nu}T^\pi V_\theta\|_\nu^2$ using stochastic gradient descent. It can be shown that the expression of gradient contains product of multiple expectations. Such framework can be modeled by two time-scale stochastic approximation, where one iterate stores the quasi-stationary estimates of some of the expectations and the other iterate is used for sampling.

5.2. The TDC Algorithm with Importance Weighting

We consider the TDC algorithm with importance-weighting from Sections 4.2 and 5.2 of Maei [13]. The gradient in this case can be shown to satisfy

$$-\tfrac{1}{2}\nabla J(\theta) = E[\rho_{X,R_n}\delta_{X,R_n,X_{n+1}}(\theta)\phi(X)] - \gamma E[\rho_{X,R_n}\phi(X_{n+1})\phi(X)^T]w(\theta),$$
$$w(\theta) = E[\phi(X)\phi(X)^T]^{-1}E[\rho_{X,R_n}\delta_{X,R_n,X_{n+1}}(\theta)\phi(X)].$$

Define $\phi_n = \phi(X_n)$, $\phi'_n = \phi(X_{n+1})$, $\delta_n(\theta) = \delta_{X_n,R_n,X_{n+1}}(\theta)$, and $\rho_n = \rho_{X_n,A_n}$. Therefore the associated iterations in this algorithm are

$$\theta_{n+1} = \theta_n + a(n)\rho_n[\delta_n(\theta_n)\phi_n - \gamma\phi'_n\phi_n^T w_n], \qquad w_{n+1} = w_n + b(n)[(\rho_n\delta_n(\theta_n) - \phi_n^T w_n)\phi_n] \tag{23}$$

with $\{a(n)\},\{b(n)\}$ satisfying **(A4)**.

5.3. Convergence Proof

Theorem 3 (Convergence of Temporal Difference with Correction (TDC) with Importance Weighting). *Consider the iterations (23) of the TDC. Assume the following:*

(i) $\{a(n)\},\{b(n)\}$ *satisfy* **(A4)**.

(ii) $\{(X_n,R_n,X_{n+1}), n\ge 0\}$ *is such that* $\{X_n\}$ *is a time-homogeneous finite-state irreducible Markov chain generated from the behavior policy* π_b *with unique stationary distribution* ν. $E[R_n \mid X_n = s, X_{n+1} = s'] = \sum_{a\in A}\pi_b(a \mid s)r(s,a,s')$ *and* $P(X_{n+1} = s' \mid X_n = s) = \sum_{a\in A}\pi_b(a \mid s)p(s' \mid s,a)$, *where* π_b *is the behavior policy,* $\pi \ne \pi_b$. *Also,* $E[R_n^2 \mid X_n, X_{n+1}] < \infty$ *for all* n *almost surely, and*

(iii) $C = E[\phi(X)\phi(X)^T]$ *and* $A = E[\rho_{X,R_n}\phi(X)(\phi(X) - \gamma\phi(X_{n+1}))^T]$ *are nonsingular where* $X\sim\nu$.

(iv) $\pi_b(a \mid s) > 0$ *for all* $s\in S$, $a\in A$.

(v) $\sup_n(\|\theta_n\| + \|w_n\|) < \infty$ *w.p.* 1.

Then, the parameter vector θ_n *converges with probability one as* $n\to\infty$ *to the TD(0) solution (22).*

Proof. The iterations (23) can be cast into the framework of Section 2.2 with

(i) $Z_n^{(i)} = X_{n-1}$,

(ii) $h(\theta,w,z) = E[(\rho_n(\delta_n(\theta)\phi_n - \gamma\phi'_n\phi_n^T w)) \mid X_{n-1} = z, \theta_n = \theta, w_n = w]$,

(iii) $g(\theta,w,z) = E[((\rho_n\delta_n(\theta) - \phi_n^T w)\phi_n) \mid X_{n-1} = z, \theta_n = \theta, w_n = w]$,

(iv) $M_{n+1}^{(1)} = \rho_n(\delta_n(\theta_n)\phi_n - \gamma\phi'_n\phi_n^T w_n) - E[\rho_n(\delta_n(\theta_n)\phi_n - \gamma\phi'_n\phi_n^T w_n) \mid X_{n-1},\theta_n,w_n]$,

(v) $M_{n+1}^{(2)} = (\rho_n\delta_n(\theta_n) - \phi_n^T w_n)\phi_n - E[(\rho_n\delta_n(\theta_n) - \phi_n{}^T w_n)\phi_n \mid X_{n-1},\theta_n,w_n]$,

(vi) $\mathcal{F}_n = \sigma(\theta_m, w_m, R_{m-1}, X_{m-1}, A_{m-1}, m\le n, i = 1,2)$, $n\ge 0$.

Note that in (ii) and (iii), we can define h and g independent of n due to time homogeneity of $\{X_n\}$.

Now, we verify the assumptions **(A1)–(A7)** (mentioned in Sections 2.2 and 2.3) for our application:

(i) **(A1)**: $Z_n^{(i)}$, $\forall n, i = 1,2$ takes values in compact metric space as $\{X_n\}$ is a finite-state Markov chain.

(ii) **(A5)**: Continuity of transition kernel follows trivially from the fact that we have a finite-state MDP.

Remark 10. In fact, we don't have to verify this assumption for the special case when the Markov chain is uncontrolled and has unique stationary distribution. The reason is that in such case **(A5)** will be used only in the proof of Lemma 3. However, if the Markov chain has unique stationary distribution Lemma 3 trivially follows:

(iii) **(A2)**

(a) In the following we prove the Lipschitz continuity of h in the first two arguments uniformly w.r.t. the third.

$$\begin{aligned}\|h(\theta,w,z)-h(\theta',w',z)\| &= \|E[\rho_n(\theta-\theta')^T(\gamma\phi(X_{n+1})-\phi(X_n))\phi(X_n)-\gamma\rho_n\phi(X_{n+1})\phi(X_n)^T(w-w')\mid X_{n-1}=z]\| \\ &\le L(2\|\theta-\theta'\|M^2+\|w-w'\|M^2),\end{aligned}$$

where $M=\max_{s\in S}\|\phi(s)\|$ with S being the state space of the MDP and $L=\max_{(s,a)\in(S\times A)}(\pi(a\mid s)/\pi_b(a\mid s))$. Hence h is Lipschitz continuous in the first two arguments uniformly w.r.t. the third. In the last inequality above, we use the Cauchy-Schwarz inequality.

(b) As with the case of h, g can be shown to be Lipschitz continuous in the first two arguments uniformly w.r.t. the third.

(c) Joint continuity of h and g follows from (iii)(a) and (b), respectively, as well as the finiteness of S.

(iv) **(A3)**: Clearly, $\{M^{(i)}_{n+1}\}$, $i=1,2$ are martingale difference sequences w.r.t. increasing σ-fields $\mathcal{F}_n$. Note that $E[\|M^{(i)}_{n+1}\|^2\mid\mathcal{F}_n]\le K(1+\|\theta_n\|^2+\|w_n\|^2)$, a.s., $n\ge 0$ since $E[R_n^2\mid X_n,X_{n+1}]<\infty$ for all n almost surely and S is finite.

(v) **(A4)**: This follows from the conditions (i) in the statement of Theorem 3.

Now, one can see that the faster o.d.e. becomes

$$\dot{w}(t)=E[\rho_{X,A_n}\delta_{X,R_n,X_{n+1}}(\theta)\phi(X)]-E[\phi(X)\phi(X)^T]w(t).$$

Clearly, $C^{-1}E[\rho_{X,A_n}\delta_{X,R_n,X_{n+1}}(\theta)\phi(X)]$ is the globally asymptotically stable equilibrium of the o.d.e. Moreover, $V'(\theta,w)=\frac{1}{2}\|Cw-E[\rho_{X,A_n}\delta_{X,R_n,X_{n+1}}(\theta)\phi(X)]\|^2$ is continuously differentiable. Additionally, $\lambda(\theta)=C^{-1}E[\rho_{X,A_n}\delta_{X,R_n,X_{n+1}}(\theta)\phi(X)]$ and it is Lipschitz continuous in θ verifying **(A6)′**. For the slower o.d.e., the global attractor is $A^{-1}E[\rho_{X,A_n}R_n\phi(X)]$ verifying the additional assumption in Corollary 1. The attractor set here is a singleton. Also, **(A7)** is (v) in the statement of Theorem 3. Therefore the assumptions **(A1)–(A5)**, **(A6)′**, **(A7)** are verified. The proof would then follow from Corollary 1. □

Remark 11. The reason for using two time-scale framework for the TDC algorithm is to make sure that the o.d.e.'s have globally asymptotically stable equilibrium.

Remark 12. Because of the fact that the gradient is a product of two expectations, the scheme is a pseudogradient descent, which helps to find the global minimum here.

Remark 13. Here, we assume the stability of the iterates (23). Certain sufficient conditions have been sketched for showing stability of single time-scale stochastic recursions with controlled Markov noise Borkar [8, p. 75, Theorem 9]. This subsequently needs to be extended to the case of two time-scale recursions.

Another way to ensure boundedness of the iterates is to use a projection operator. However, projection may introduce spurious fixed points on the boundary of the projection region and finding globally asymptotically stable equilibrium of a projected o.d.e. is hard. Therefore we do not use projection in our algorithm.

Remark 14. Convergence analysis for TDC with importance weighting along with eligibility traces cf. Maei [13, p. 74] where it is called GTD(λ) can be done similarly using our results. The main advantage is that it works for $\lambda<1/(L\gamma)$ ($\lambda\in[0,1]$ being the eligibility function), whereas the analysis in Yu [21] is shown only for λ very close to 1.

Remark 15. One can analyze this algorithm when the state space is infinite by imposing assumptions on ϕ as well as the target and behavior policies.

6. Conclusion

We presented a general framework for two time-scale stochastic approximation with controlled Markov noise. Moreover, using a special case of our results, i.e., when the random process is a finite-state irreducible time-homogeneous Markov chain (hence has a unique stationary distribution) and uncontrolled (i.e., does not depend on iterates), we provided a rigorous proof of convergence for off-policy temporal-difference learning algorithm that is also extendible to eligibility traces (for a sufficiently large range of λ) with linear function approximation under the assumption that the "on-policy" trajectory for a behavior policy is only available. This has previously not been done to our knowledge.

Acknowledgments
The authors thank Csaba Szepesvári for some useful discussion on the literature of off-policy learning.

References
[1] Aubin J, Cellina A (1984) *Differential Inclusions: Set-Valued Maps and Viability Theory* (Springer, Berlin).
[2] Benaïm M (1999) Dynamics of stochastic approximation algorithms. Azéma J, Émery M, Ledoux M, Yor M, eds. *Séminaire de probabilités XXXIII* (Springer, Berlin), 1–68.
[3] Benaïm M, Hofbauer J, Sorin S (2005) Stochastic approximations and differential inclusions. *SIAM J. Control Optim.* 44(1):328–348.
[4] Benveniste A, Metivier M, Priouret P (1990) *Adaptive Algorithms and Stochastic Approximation* (Springer, New York).
[5] Borkar VS (1995) *Probability Theory: An Advanced Course* (Springer, New York).
[6] Borkar VS (1997) Stochastic approximation with two time scales. *Systems Control Lett.* 29(5):291–294.
[7] Borkar VS (2006) Stochastic approximation with "controlled Markov noise." *Systems Control Lett.* 55(2):139–145.
[8] Borkar VS (2008) *Stochastic Approximation: A Dynamic Systems Viewpoint* (Cambridge University Press, Cambridge, UK).
[9] Degris T, White M, Sutton RS (2012) Linear off-policy actor-critic. *Proc. 29th Internat. Conf. Machine Learning, ICML, '12* (Omnipress, Madison, WI).
[10] Konda VR, Tsitsiklis JN (2003) Linear stochastic approximation driven by slowly varying Markov chains. *Systems Control Lett.* 50(2): 95–102.
[11] Konda VR, Tsitsiklis JN (2003) On actor-critic algorithms. *SIAM J. Control Optim.* 42(4):1143–1166.
[12] Ma DJ, Makowski AM, Shwartz A (1990) Stochastic approximations for finite state Markov chains. *Stochastic Processes Their Appl.* 35(1):27–45.
[13] Maei HR (2011) Gradient temporal-difference learning algorithms. PhD thesis, University of Alberta, Alberta, Canada.
[14] Menache I, Mannor S, Shimkin N (2005) Basis function adaptation in temporal difference reinforcement learning. *Ann. Oper. Res.* 134(1):215–238.
[15] Metivier M, Priouret P (1984) Applications of a Kushner and Clark lemma to general classes of stochastic algorithms. *IEEE Trans. Inform. Theory* 30(2):140–151.
[16] Rudin W (1976) *Principles of Mathematical Analysis*, 3rd ed. (McGraw-Hill, New York).
[17] Sutton RS, Maei RS, Szepesvári C (2008) A convergent $O(n)$ algorithm for off-policy temporal-difference learning with linear function approximation. Koller D, Schuurmans D, Bengio Y, Bottou L, eds. *Adv. Neural Inform. Processing Systems 21, NIPS '08*.
[18] Sutton RS, Maei HR, Precup D, Bhatnagar S, Silver D, Wiewiora E (2009) Fast gradient-descent methods for temporal-difference learning with linear function approximation. Pohoreckyj Danyluk A, Bottou L, Littman ML eds. *Proc. 26th Internat. Conf. Machine Learning, ICML '10* (ACM, New York), 993–1000.
[19] Tadić VB (2004) Almost sure convergence of two time-scale stochastic approximation algorithms. *Proc. 2004 Amer. Control Conf.* (IEEE, Piscataway, NJ).
[20] Tadić VB (2015) Convergence and convergence rate of stochastic gradient search in the case of multiple and non-isolated extrema. *Stochastic Processes their Appl.* 125(5):1715–1755.
[21] Yu H (2012) Least squares temporal difference methods: An analysis under general conditions. *SIAM J. Control Optim.* 50(6):3310–3343.
[22] Yu H (2016) Weak convergence properties of constrained emphatic temporal-difference learning with constant and slowly diminishing stepsize. *J. Machine Learning Res.* 17(220):1–58.