This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2018.2874687, IEEE Transactions on Automatic Control

Stability of Stochastic Approximations with 'Controlled Markov' Noise and Temporal Difference Learning

Arunselvan Ramaswamy[†] arunr@mail.uni-paderborn.de , Shalabh Bhatnagar[‡] shalabh@iisc.ac.in

***Abstract*—We are interested in understanding stability (almost sure boundedness) of stochastic approximation algorithms (SAs) driven by a 'controlled Markov' process. Analyzing this class of algorithms is important, since many reinforcement learning (RL) algorithms can be cast as SAs driven by a 'controlled Markov' process. In this paper, we present easily verifiable sufficient conditions for stability and convergence of SAs driven by a 'controlled Markov' process. Many RL applications involve continuous state spaces. While our analysis readily ensures stability for such continuous state applications, traditional analyses do not. As compared to literature, our analysis presents a two-fold generalization: (a) the Markov process may evolve in a continuous state space and (b) the process need not be ergodic under any given stationary policy. Temporal difference learning (TD) is an important policy evaluation method in reinforcement learning. The theory developed herein, is used to analyze generalized $TD(0)$, an important variant of TD. Our theory is also used to analyze a TD formulation of supervised learning for forecasting problems.**

***Index Terms*—Stochastic approximation algorithms, 'controlled Markov' noise, stability, convergence, temporal difference learning, $TD(0)$, reinforcement learning, supervised learning.**

I. Introduction

Reinforcement learning (RL) algorithms such as Q learning, temporal difference learning and value iteration methods have seen a major resurgence in recent years as model-free, yet simple and effective, solutions to many important problems. RL is used to solve problems in fields ranging from health-care to transportation. As RL becomes ubiquitous in solving critical problems, there is a need to provide "behavioral guarantees" for RL. Stochastic approximation algorithms (SAs) are an important class of model-free algorithms, with associated analytical tools, that play an important role in providing such guarantees. The important foundational papers on SAs include [12], [3], [4], [5] and [11]. Recent results in this field include [14], [15] and [1].

SAs with 'controlled' Markov noise are an important subclass of algorithms, particularly since many RL algorithms can be cast in this setting. The groundwork for analyzing such algorithms was laid by Benveniste et. al. [7] and Borkar [9]. The analysis of [7] requires the Markov process to evolve in a finite state space and be ergodic. The analysis of [9] allows for continuous state spaces and the process may be governed by an additional control-valued sequence in addition to the parameter iterates, a setting that we also consider, and be non-ergodic, *i.e.,* it can have multiple stationary distributions. However, [9] requires that stability (almost sure boundedness of the algorithm) is ensured. Stability is a highly non-trivial assumption as there is no easy way to ensure compliance with this requirement.

Many RL applications involve Markov processes that evolve over a continuous state space. Here, ensuring stability is especially hard. The main contribution of this paper is the development of easily verifiable sufficient conditions for stability and convergence of SAs driven with an iterate-dependent Markov process that may depend on another control-valued sequence. *Our analysis presents a two-fold generalization over traditional ones (a) the Markov process may evolve in a continuous state space and (b) the process need not be ergodic under any given stationary policy.* Under our conditions, the algorithm is shown to be stable and it tracks a solution to a limiting differential inclusion (DI), defined in terms of the ergodic occupation measures of the Markov process. Further, the limiting set is internally chain transitive and invariant. Our stability assumptions are particularly interesting, since they can be readily used to ensure stability in many RL applications, and are compatible with traditional convergence analyses.

Temporal difference learning (TD) is an important RL algorithm that is popularly employed in 'policy evaluation' problems. $TD(0)$ is an important variant of TD that is effective, yet simple to implement. Our theory is used to provide a complete analysis of generalized $TD(0)$. Previously, Tsitsiklis and Van Roy [17] have analyzed TD. However, [17] assumes that the Markov process is ergodic and evolves in a finite state space. Further, the second moments of the single stage rewards are assumed to be bounded. Our analysis eliminates the need to impose such restrictions, see Section V-A for details.

As yet another application of our theory, we analyze a *TD formulation of supervised learning*, to solve the weather forecasting problem described in *Chapter 11* of Spall [16]. It may be noted that the analyses in [17] and [1] cannot be applied to analyze the aforementioned algorithm.

[†] Dept. of Electrical Engineering and Information Technology, Paderborn University, Paderborn - 33908, Germany. His position was funded by the German Research Foundation (DFG) - 315248657.

✠Part of this research was conducted when Ramaswamy was at Indian Institute of Science.

[‡] Department of Computer Science and Automation and the Robert Bosch Centre for Cyber Physical Systems, Indian Institute of Science, Bangalore - 560012, India.

A. Notations & Definitions

[Upper-semicontinuous map] We say that H is upper-semicontinuous, if given sequences $\{x_n\}_{n\geq 1}$ (in $\mathbb{R}^n$) and $\{y_n\}_{n\geq 1}$ (in $\mathbb{R}^m$) with $x_n \to x$, $y_n \to y$ and $y_n \in H(x_n)$, $n \geq 1$, then $y \in H(x)$.

[Marchaud Map] A set-valued map $H : \mathbb{R}^n \to \{subsets\ of\ \mathbb{R}^m\}$ is called *Marchaud* if it satisfies the following properties: **(i)** for each $x \in \mathbb{R}^n$, $H(x)$ is convex and compact; **(ii)** *(point-wise boundedness)* for each $x \in \mathbb{R}^n$, $\sup_{w\in H(x)} \|w\| < K(1+\|x\|)$ for some $K > 0$; **(iii)** H is *upper-semicontinuous*.

Let H be a Marchaud map on $\mathbb{R}^d$. The differential inclusion (DI) given by $\dot{x} \in H(x)$ is guaranteed to have at least one solution that is absolutely continuous. The reader is referred to [2] for more details.

[Open and closed neighborhoods of a set] Let $x \in \mathbb{R}^d$ and $A \subseteq \mathbb{R}^d$, then $d(x, A) := \inf\{\|a-y\| \mid y \in A\}$. We define the *$\delta$-open neighborhood* of A by $N^\delta(A) := \{x \mid d(x, A) < \delta\}$. The *$\delta$-closed neighborhood* of A is defined by $\overline{N^\delta}(A) := \{x \mid d(x, A) \leq \delta\}$. The open ball of radius r around the origin is represented by $B_r(0)$, while the closed ball is represented by $\overline{B}_r(0)$.

[Upper-limit of a sequence of sets, Limsup] Let $\{K_n\}_{n\geq 1}$ be a sequence of sets in $\mathbb{R}^d$. The *upper-limit* of $\{K_n\}_{n\geq 1}$ is given by, $Limsup_{n\to\infty} K_n := \{y \mid \underline{lim}_{n\to\infty}\, d(y, K_n) = 0\}$.

We may interpret that the upper-limit collects its accumulation points.

II. Assumptions

As stated earlier, we are motivated by the need to analyze RL algorithms. However, we present our analysis for a more general class of algorithms: SAs driven by an iterate-dependent 'controlled Markov process'. Later, we shall recast this analysis to understand generalized $TD(0)$ and a TD formulation of supervised learning with delayed rewards. It may be noted that the analysis of TD for supervised learning was only possible due to our consideration of the general class.

Let us begin by describing a SA driven by a 'controlled Markov' process, following which we list the assumptions involved. Since we use results from [9] in our convergence analysis, we relate our assumptions with those of [9]. We have the following recursion in $\mathbb{R}^d$:

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

(A1)(i) $h : \mathbb{R}^d \times S \to \mathbb{R}^d$ is a jointly continuous map, and S a compact metric space. The map h is Lipschitz continuous in the first component, with Lipschitz constant L, which does not change with the second component.[1]

(A1)(ii) $\{y_n\}_{n\geq 0}$ is an S-valued 'controlled Markov' process controlled by (a) the iterate sequence $\{x_n\}$ and (b) an additional control-valued sequence. [2]

(A2) $\{M_n\}_{n\geq 1}$ is a square integrable martingale difference sequence (noise). Further, $E\left[\|M_{n+1}\|^2 \mid \mathcal{F}_n\right] \leq K(1+\|x_n\|^2)$, where $n \geq 0$ and $\mathcal{F}_n := \sigma\langle x_m, y_m, M_m; m \leq n\rangle$, $n \geq 0$. [3]

(A3) The step-size sequence $\{a(n)\}_{n\geq 0}$ satisfies the following: it is non-increasing, $a(n) > 0$ for all $n \geq 0$, $\sum_{n=0}^{\infty} a(n) = \infty$ and $\sum_{n=0}^{\infty} a(n)^2 < \infty$. Without loss of generality let $\sup_{n\geq 0} a(n) \leq 1$. [4]

Before proceeding, we define a family of rescaled functions as follows. For each $c \geq 1$, define $h_c : \mathbb{R}^d \times S \to \mathbb{R}^d$ by $h_c(x, y) := h(cx, y)/c$. Also define, $h_\infty : \mathbb{R}^d \times S \to \{\text{subsets of } \mathbb{R}^d\}$ by $h_\infty(x, y) := Limsup_{c\to\infty}\{h_c(x, y)\}$, where $Limsup$ is the upper-limit of a sequence of sets (see Section I-A). Finally, define the set-valued map, H, as $H(x) := \overline{co}\left(\bigcup_{y\in S} h_\infty(x, y)\right)$, where $x \in \mathbb{R}^d$. In Lemma 2 we show that H is Marchaud. Consequently, there exists a solution to the DI $\dot{x}(t) \in H(x(t))$, see [2] for details.

Below we present our key stability assumptions, $(S1)$ and $(S2)$. They are based on the limiting behavior of the objective function h.

(S1) If $c_n \uparrow \infty$, $y_n \to y$ and $\lim_{n\to\infty} h_{c_n}(x, y_n) = u$ for some $u \in \mathbb{R}^d$, then $u \in h_\infty(x, y)$.

(S2) There exists an attracting set, $\mathcal{A}$, associated with $\dot{x}(t) \in H(x(t))$ such that $\sup_{u\in\mathcal{A}}\|u\| < 1$ and $\overline{B}_1(0)$ is a fundamental neighborhood of $\mathcal{A}$.

It follows from $(S2)$ that we can find $\delta_1, \delta_2, \delta_3$ and δ_4 such that $\delta_1 := \sup_{u\in\mathcal{A}} \|u\|$ and $\delta_1 < \delta_2 < \delta_3 < \delta_4 < 1$.

OBSERVATION 1. *h_c is a jointly continuous function that is Lipschitz continuous in the first component. Further, the Lipschitz constant is independent of the second component and c. Without loss of generality we take this Lipschitz constant to be L from $(A1)$. Since S is compact, $\|h_c(0,\cdot)\|_\infty \leq \|h(0,\cdot)\|_\infty \leq \overline{M}$ for some $0 < \overline{M} < \infty$. It now follows from $\|h_c(x,y) - h_c(0,y)\| \leq L\|x\|$, that $\|h_c(x,y)\| \leq \|h_c(0,y)\| + L\|x\|$ and $\|h_c(x,y)\| \leq K(1+\|x\|)$ are satisfied, where $K := L \vee \overline{M}$. Again, without loss of generality this K is from $(A2)$ and does not change with c. In other words, $\sup_{c\geq 1}\|h_c(x,y)\| \leq K(1+\|x\|)$. Further, it follows from the definition of h_∞ and H that*

$$\sup_{u\in H(x)} \|u\| \leq K(1+\|x\|). \tag{2}$$

As stated earlier, we need to show that the set-valued map H is Marchaud. First, we prove a technical lemma that is needed for this purpose.

Lemma 1. *Suppose $x_n \to x$ in $\mathbb{R}^d$, $y_n \to y$ in S, $c_n \uparrow \infty$ and $\lim_{c_n\uparrow\infty} h_{c_n}(x_n, y_n) = u$. Then $u \in h_\infty(x, y)$.*

Proof. It follows from Observation 1 that $\|h_{c_n}(x, y_n) - h_{c_n}(x_n, y_n)\| \leq L\|x_n - x\|$. Since $x_n \to x$ the sequences $\{h_{c_n}(x, y_n)\}_n$ and $\{h_{c_n}(x_n, y_n)\}_n$ have the same limit as $n \to \infty$ *i.e.,* u. It now follows from assumption $(S1)$ that $u \in h_\infty(x, y)$. □

[1] assumption (1) in *Section 2*, [9].

[2] S is a compact metric space, and hence Polish.

[3] assumption (2) in *Section 2*, [9].

[4] assumption (3) in *Section 2*, [9].

We claim the following: if $x_n \to x$ in $\mathbb{R}^d$, $\{y_n\} \subset S$ and $c_n \to \infty$ then $d(h_{c_n}(x_n, y_n), H(x)) \to 0$. Suppose this claim were false, then, without loss of generality, $d(h_{c_n}(x_n, y_n), H(x)) > \epsilon$ for some $\epsilon > 0$, $n \geq 0$. Since S is compact, $\exists \{m(n)\} \subseteq \{n\}$ such that $c_{m(n)} \uparrow \infty$, $\lim_{m(n)\to\infty} y_{m(n)} = y$ and $h_{c_{m(n)}}(x_{m(n)}, y_{m(n)}) \to u$ for some $y \in S$ and $u \in \mathbb{R}^d$. Hence, $x_{m(n)} \to x$, $y_{m(n)} \to y$, $c_{m(n)} \uparrow \infty$, $h_{c_{m(n)}}(x_{m(n)}, y_{m(n)}) \to u$ and $u \notin h_\infty(x, y) \subseteq H(x)$. This contradicts Lemma 1.

Lemma 2. *The set-valued map H is Marchaud.*

Proof. Recall that $H(x) = \overline{co}\left(\cup_{y\in S} h_\infty(x, y)\right)$. As explained earlier (*cf.* (2)),

$$\sup_{u\in H(x)} \|u\| \leq K(1 + \|x\|).$$

Hence H is point-wise bounded. From the definition of H it follows that $H(x)$ is convex and compact for each $x \in \mathbb{R}^d$.

It is left to show that H is upper semi-continuous. Let $x_n \to x$, $u_n \to u$ and $u_n \in H(x_n)$, $n \geq 1$. We need to show that $u \in H(x)$. If this is not true, then there exists a linear functional on $\mathbb{R}^d$, say f, such that $\sup_{v\in H(x)} f(v) \leq \alpha - \epsilon$ and $f(u) \geq \alpha + \epsilon$, for some $\alpha \in \mathbb{R}$ and $\epsilon > 0$. Since $u_n \to u$, there exists N such that for each $n \geq N$ $f(u_n) \geq \alpha + \frac{\epsilon}{2}$, *i.e.*, $H(x_n) \cap [f \geq \alpha + \frac{\epsilon}{2}] \neq \phi$, here $[f \geq a]$ is used to denote the set $\{x \mid f(x) \geq a\}$. For the sake of notational convenience let us denote $\cup_{y\in S} h_\infty(x, y)$ by $A(x)$ for all $x \in \mathbb{R}^d$. We claim that $A(x_n) \cap [f \geq \alpha + \frac{\epsilon}{2}] \neq \phi$ for all $n \geq N$. We shall prove this claim later, for now we assume that the claim is true and proceed.

Pick $w_n \in A(x_n) \cap [f \geq \alpha + \frac{\epsilon}{2}]$ for each $n \geq N$. Let $w_n \in h_\infty(x_n, y_n)$ for some $y_n \in S$. Since $\{w_n\}_{n\geq N}$ is norm bounded it contains a convergent subsequence, say $\{w_{n(k)}\}_{k\geq 1} \subseteq \{w_n\}_{n\geq N}$. Let $\lim_{k\to\infty} w_{n(k)} = w$. Since $w_{n(k)} \in h_\infty(x_{n(k)}, y_{n(k)})$, $\exists\ c_{n(k)} \in \mathbb{N}$ such that $\|w_{n(k)} - h_{c_{n(k)}}(x_{n(k)}, y_{n(k)})\| < \frac{1}{n(k)}$. The sequence $\{c_{n(k)}\}_{k\geq 1}$ is chosen such that $c_{n(k+1)} > c_{n(k)}$ for each $k \geq 1$. Since $\{y_{n(k)}\}_{k\geq 1}$ is from a compact set, there exists a convergent subsequence. For the sake of notational convenience (without loss of generality) we assume that the sequence itself has a limit, *i.e.*, $y_{n(k)} \to y$ for some $y \in S$. We have the following: $c_{n(k)} \uparrow \infty$, $x_{n(k)} \to x$, $y_{n(k)} \to y$, $w_{n(k)} \to w$ and $w_{n(k)} \in h_{c_{n(k)}}(x_{n(k)}, y_{n(k)})$ for $k \geq 1$. It follows from Lemma 1 that $w \in h_\infty(x, y)$. Since $w_{n(k)} \to w$ and $f(w_{n(k)}) \geq \alpha + \frac{\epsilon}{2}$ for each $k \geq 1$, we have that $f(w) \geq \alpha + \frac{\epsilon}{2}$. This contradicts $\sup_{w\in H(x)} f(w) \leq \alpha - \epsilon$.

It remains to prove that $A(x_n) \cap [f \geq \alpha + \frac{\epsilon}{2}] \neq \phi$ for all $n \geq N$. If this were not true, then $\exists \{m(k)\}_{k\geq 1} \subseteq \{n \geq N\}$ such that $A(x_{m(k)}) \subseteq [f < \alpha + \frac{\epsilon}{2}]$ for all k. It follows that $H(x_{m(k)}) = \overline{co}(A(x_{m(k)})) \subseteq [f \leq \alpha + \frac{\epsilon}{2}]$ for each $k \geq 1$. Since $u_{n(k)} \to u$, $\exists N_1$ such that for all $n(k) \geq N_1$, $f(u_{n(k)}) \geq \alpha + \frac{3\epsilon}{4}$. This leads to a contradiction. □

In the following section, we show that (1) is stable, provided $(A1)$-$(A3)$, $(S1)$ and $(S2)$ are satisfied. Following this, in Section IV, we show that (1) tracks a solution to a limiting DI, defined in terms of the ergodic occupation measures.

III. Stability analysis

For our analysis, we need to define a continuous trajectory, $\overline{x}([0, \infty))$, such that the limit of this trajectory coincides with that of $\{x_n\}_{n\geq 0}$. For this, we divide time axis as follows: $t(0) := 0$ and $t(n) := \sum_{i=0}^{n-1} a(i)$, $\forall n \geq 1$. Then we let, $\overline{x}(t(n)) := x_n\ \forall n \geq 0$, and for $t \in (t(n), t(n+1))$ let

$$\overline{x}(t) := \left(\frac{t(n+1) - t}{t(n+1) - t(n)}\right) \overline{x}(t(n)) + \left(\frac{t - t(n)}{t(n+1) - t(n)}\right) \overline{x}(t(n+1)).$$

The time axis is further divided into intervals of approximate length T, where $T := T(\delta_2 - \delta_1) + 1$ and δ_1, δ_2 are defined in Section II. Now, we define $T(\epsilon)$ for $\epsilon > 0$ as follows: given a solution $x(\cdot)$ to $\dot{x}(t) \in H(x(t))$ with $\|x(0)\| \leq 1$, $x(t) \in N^\epsilon(\mathcal{A})$ for $t \geq T(\epsilon)$. Given $\epsilon > 0$, $\exists\ T(\epsilon)$ with the aforementioned property, since $\overline{B}_\epsilon(0)$ is a fundamental neighborhood of $\mathcal{A}$. Define, $T_0 := 0$ and $T_n := min\{t(m) : t(m) \geq T_{n-1} + T\}$ for $n \geq 1$. *In other words, $\exists \{m(n)\}_{n\geq 0} \subseteq \mathbb{N}$ such that $T_n = t(m(n))$ for all $n \geq 0$.*

For the purpose of analyzing stability, we need to define the following rescaled trajectory: $\hat{x}(t) := \frac{\overline{x}(t)}{r(n)}$, where $t \in [T_n, T_{n+1})$ and $r(n) = \|\overline{x}(T_n)\| \vee 1$. Also, let $\hat{x}(T_{n+1}^-) := \lim_{t\uparrow T_{n+1}} \hat{x}(t)$. We also define the rescaled martingale difference terms as follows: $\hat{M}_{k+1} := \frac{M_{k+1}}{r(n)}$, $t(k) \in [T_n, T_{n+1})$.

Finally, we define the following piece-wise constant trajectories. Define $\hat{z}(t) := h_{r(n)}(\hat{x}(t(m)), y_m)$, where $t \in [t(m), t(m+1))$ and $T_n \leq t(m) < t(m+1) \leq T_{n+1}$; also define $\overline{y}(t) := y_n$ for $t \in [t(n), t(n+1))$ and $n \geq 0$.

It can be readily verified that: $E\left[\|\hat{M}_{k+1}\|^2 | \mathcal{F}_k\right] \leq K\left(1 + \|\hat{x}(t(k))\|^2\right)$. The following lemma states that the Martingale noise convergences a.s. A proof can be found in Borkar & Meyn [11].

Lemma 3. $\sup_{t\geq 0} E\|\hat{x}(t)\|^2 < \infty$. *Further, the sequence $\hat{\zeta}_n$, $n \geq 0$, converges almost surely, where $\hat{\zeta}_n := \sum_{k=0}^{n-1} a(k)\hat{M}_{k+1}$ for all $n \geq 1$.*

Let $x^n([0, T])$ denote a solution up to time T for $\dot{x}^n(\cdot) = \hat{z}(T_n + \cdot)$ with $x^n(0) = \hat{x}(T_n)$. Then, $x^n(t) = \hat{x}(T_n) + \int_0^t \hat{z}(T_n + s)\, ds$. The following lemma states that the rescaled trajectories track the above defined solution trajectories.

Lemma 4. $\lim_{n\to\infty} \sup_{t\in[T_n, T_n+T]} \|x^n(t) - \hat{x}(t)\| = 0\ a.s.$

Proof. The proof proceeds along the lines of the proof of *Lemma 4* in [14]. □

Recall that $T = T(\delta_2 - \delta_1) + 1$. We show that $\{x^n([0, T]) \mid n \geq 0\}$ and $\{\hat{x}([T_n, T_n + T]) \mid n \geq 0\}$ are relatively compact subsets of $C([0, T], \mathbb{R}^d)$, endowed with the sup-norm, $\|\cdot\|_\infty$. To do this, we merely show that $\{x^n([0, T]) \mid n \geq 0\}$ is relatively compact. The relative compactness of $\{\hat{x}([T_n, T_n + T]) \mid n \geq 0\}$ follows from Lemma 4. In what follows, we show that $\{x^n([0, T]) \mid n \geq 0\}$ is equicontinuous and point-wise bounded. Relative compactness

of $\{x^n([0,T]) \mid n \geq 0\}$ then follows from the *Arzela-Ascoli Theorem*.

Recall that $\sup_{t\geq 0} E\|\hat{x}(t)\|^2 < \infty$ *a.s.* and $\|\hat{z}(t)\| \leq K(1 + \|\hat{x}([t])\|)$, where $[t] := \max\{t(m) \mid t(m) \leq t\}$. Hence $\sup_{t\geq 0} \|\hat{z}(t)\| < \infty$ *a.s.* For $\delta > 0$, we have:

$$\|x^n(t+\delta) - x^n(t)\| \leq \int_t^{t+\delta} \|\hat{z}(s)\|\, ds \leq M\delta,$$

where M is a, possibly sample path dependent, constant such that $\sup_{t\geq 0} \|\hat{z}(t)\| \leq M$. In other words, $\{x^n([0,T]) \mid n \geq 0\}$ is equicontinuous. Now, we need to show that the point-wise boundedness property is satisfied. Since $\|x^n(0)\| = \|\hat{x}(T_n)\| \leq 1$, it follows from Gronwall's inequality that $\sup_{n\geq 0} \|x^n\|_\infty < \infty$ a.s., where $\|x^n\|_\infty = \sup_{t\in[0,T]} \|x^n(t)\|$. In other words, $\{x^n([0,T]) \mid n \geq 0\}$ is point-wise bounded.

We are interested in showing that $\sup_{n\geq 0}\|x_n\| < \infty$ a.s. We present a proof by contradiction. Suppose the iterates are unstable, then with positive probability we have $\sup_{n\geq 0} r(n) = \infty$ and there exists $\{l\} \subseteq \{n\}$ such that $r(l) \uparrow \infty$. Note that this sub-sequence may be sample path dependent. The following lemma shows that $\{\hat{x}([T_l, T_l + T]) \mid \{l\} \subseteq \{n\} \ \&\ r(l) \uparrow \infty\}$ tracks a solution to $\dot{x}(t) \in H(x(t))$ as $r(l) \uparrow \infty$.

Lemma 5. *Let $\{l\} \subseteq \{n\}$ such that $r(l) \uparrow \infty$. Any limit of $\{\hat{x}([T_l, T_l + T]) \mid \{l\} \subseteq \{n\}\ \&\ r(l) \uparrow \infty\}$ in $C([0,T], \mathbb{R}^d)$ is of the form $x(t) = x(0) + \int_0^t z(s)\, ds$, where $x(0) \in \overline{B}_1(0)$ and $z : [0,T] \to \mathbb{R}^d$ is a measurable function such that $z(t) \in H(x(t))$, $t \in [0,T]$.*

Proof. Define the notation $[t] := \max\{t(m) \mid t(m) \leq t\}$. For a fixed $n \geq 0$ and $t_0 \in [T_n, T_{n+1})$, we have $\hat{z}(t_0) = h_{r(n)}(\hat{x}([t_0]), \overline{y}([t_0]))$ and $\|\hat{z}(t_0)\| \leq K(1 + \|\hat{x}([t_0])\|)$. It follows from Lemma 3 that $\|\hat{z}(t)\| < \infty$ *a.s.* Recall that $\{\hat{x}(T_l + \cdot) \mid \{l\} \subseteq \{n\}\}$ is relatively compact in $C([0,T], \mathbb{R}^d)$. Without loss of generality we may assume that: **(a)** $\hat{x}(T_l + \cdot) \to x(\cdot)$ in $C([0,T], \mathbb{R}^d)$, for some $x(\cdot) \in C([0,T], \mathbb{R}^d)$; **(b)** $\hat{z}(T_l + \cdot) \to z(\cdot)$ weakly in $L^2([0,T], \mathbb{R}^d)$, for some $z(\cdot) \in L^2([0,T], \mathbb{R}^d)$.

From Lemma 4 we get $x^l(\cdot) \to x(\cdot)$ in $C([0,T], \mathbb{R}^d)$. Letting $r(l) \to \infty$ in

$$x^l(t) = x^l(0) + \int_0^t \hat{z}(T_l + s)\, ds, \text{ we get}$$
$$x(t) = x(0) = \int_0^t z(s) ds.$$

Since $\|x^l(0)\| = \|\hat{x}(T_l)\| \leq 1\ \forall l$, we get $\|x(0)\| \leq 1$. $\hat{z}(T_l + \cdot) \to z(\cdot)$ weakly in $L^2([0,T], \mathbb{R}^d)$, hence it follows from *Banach-Saks Theorem* that

$$\exists\ \{k(l)\} \subseteq \{l\} \text{ such that } \frac{1}{N}\sum_{l=1}^{N} \hat{z}(T_{k(l)} + \cdot) \to z(\cdot) \text{ strongly in } L^2([0,T], \mathbb{R}^d).$$

Further,

$$\exists\ \{m(N)\} \subseteq \{N\} \text{ such that } \frac{1}{m(N)}\sum_{l=1}^{m(N)} \hat{z}(T_{k(l)} + \cdot) \to z(\cdot)\ a.e. \text{ on } [0,T]. \quad (3)$$

Fix $t_0 \in [0,T]$ such that (3) holds, *i.e.*,

$$\lim_{m(N)\to\infty} \frac{1}{m(N)}\sum_{l=1}^{m(N)} \hat{z}(T_{k(l)} + t_0) = z(t_0). \quad (4)$$

We know that $\hat{z}(T_{k(l)} + t_0) = h_{r(k(l))}(\hat{x}([T_{k(l)} + t_0]), \overline{y}([T_{k(l)} + t_0]))$. Note that $\overline{y}([T_{k(l)} + t_0]) = \overline{y}(T_{k(l)} + t_0)$.

We claim the following: For any $\epsilon > 0$ there exists N such that for all $n \geq N$ $\|\hat{x}(t(m)) - \hat{x}(t(m+1))\| < \epsilon$, where $T_n \leq t(m) < t(m+1) < T_{n+1}$. If $t(m+1) = T_{n+1}$ then we claim that $\|\hat{x}(t(m)) - \hat{x}(T_{n+1}^-)\| < \epsilon$. We shall prove this later, for now we assume it is true and proceed.

Since $\hat{x}(T_{k(l)} + t_0) \to x(t_0)$ it follows from the above claim that $\hat{x}([T_{k(l)} + t_0]) \to x(t_0)$. Since $r(k(l)) \uparrow \infty$ it follows from Lemma 1 that

$$\lim_{r(k(l))\uparrow\infty} d\left(h_{r(k(l))}(\hat{x}([T_{k(l)} + t_0]), \overline{y}([T_{k(l)} + t_0])), H(x(t_0))\right) = 0 \text{ i.e., } \lim_{r(k(l))\uparrow\infty} d\left(\hat{z}(T_{k(l)} + t_0), H(x(t_0))\right) = 0.$$

Further, since $H(x(t_0))$ is convex and compact, it follows from equation (4) that $z(t_0) \in H(x(t_0))$. On the measure zero subset of [0,T] where (3) does not hold, the value of $z(\cdot)$ can be modified to ensure that $z(t) \in H(x(t))$ for all $t \in [0,T]$.

It is left to prove the claim that was made earlier. We first show that given any $\epsilon > 0$ there exists N such that $n \geq N$ implies that $\|\hat{x}(t(m)) - \hat{x}(t(m+1))\| < \epsilon$, where $T_n \leq t(m) < t(m+1) < T_{n+1}$. We know that

$$\hat{x}(t(m+1)) = \hat{x}(t(m)) + a(n)\left(h_{r(n)}(\hat{x}(t(m)), \overline{y}(t(m))) + \hat{M}_{n+1}\right).$$

Hence,

$$\|\hat{x}(t(m)) - \hat{x}(t(m+1))\| \leq a(n)\|h_{r(n)}(\hat{x}(t(m)), \overline{y}(t(m)))\| + \|\zeta_{n+1} - \zeta_n\|.$$

From (2), the above inequality becomes

$$\|\hat{x}(t(m)) - \hat{x}(t(m+1))\| \leq a(n)K(1 + \|\hat{x}(t(m))\|) + \|\zeta_{n+1} - \zeta_n\|.$$

It follows from Lemma 3 that $a(n)K(1 + \|\hat{x}(t(m))\|) \to 0$ and $\|\zeta_{n+1} - \zeta_n\| \to 0$ respectively in the 'almost sure' sense. In other words, there exists N (possibly sample path dependent) such that the claim holds. The second part of the unproven claim considers the situation when $t(m+1) = T_{n+1}$, the proof of which follows in a similar manner. □

The following is an immediate corollary to the above lemma.

Corollary 1. *$\exists\ 1 < R_0 < \infty$ such that $\forall\ r(l) > R_0$ $\|\hat{x}(T_l + \cdot) - x(\cdot)\| < \delta_3 - \delta_2$, where $\{l\} \subseteq \mathbb{N}$ and $x(\cdot)$ is a solution (up to time T) of $\dot{x}(t) \in H(x(t))$ such that $\|x(0)\| \leq 1$. The form of $x(\cdot)$ is given by Lemma 5.*

Proof. Assume to the contrary that $\exists\ r(l) \uparrow \infty$ such that $\hat{x}(T_l + \cdot)$ is at least $\delta_3 - \delta_2$ away from any solution to the DI. It follows from Lemma 5 that there exists a subsequence of $\{\hat{x}(T_l + t), 0 \leq t \leq T\ :\ l \subseteq \mathbb{N}\}$ guaranteed to converge, in $C([0,T], \mathbb{R}^d)$, to a solution of $\dot{x}(t) \in H(x(t))$ such that $\|x(0)\| \leq 1$. This is a contradiction. □

It is worth noting that R_0 may be sample path dependent. Since $T = T(\delta_2 - \delta_1) + 1$ we get $\|\hat{x}([T_l + T])\| < \delta_3$ for all T_l such that $\|\overline{x}(T_l)\| (= r(l)) > R_0$. We are now ready to state our stability theorem.

Theorem 1 (The Stability Theorem)**.** *Under assumptions $(A1) - (A3)$, $(S1)$ & $(S2)$, $\sup_{n \geq 0} \|x_n\| < \infty$ a.s.*

Proof. Define $\mathcal{B} := \{\hat{\zeta}_n \text{ converges}\}$. It is enough to show that $\sup_{n \geq 0} \|x_n\| < \infty$ on $\mathcal{B}$. Suppose not, $\exists\ \mathcal{D} \subseteq \mathcal{B}$ such that $P(\mathcal{D}) > 0$ and $\sup_{n \geq 0} \|x_n\| = \infty$ on $\mathcal{D}$. In other words, $\exists \{l\} \subseteq \{n\}$ such that $r(l) \uparrow \infty$. Recall the notation $[t] = \max\{t(k) \mid t(k) \leq t\}$ and the sequence $\{m(l)\}_{l \geq 0}$ such that $T_l = t(m(l))$. There exists $N > 0$ such that the following hold simultaneously.

1) If $m(l) \geq N$ then $r(l) > R_0$ (see Corollary 1).
2) If $m(l) \geq N$ then $\|\hat{x}(T_l + T)\| < \delta_3$ (since $T = T(\delta_2 - \delta_1) + 1$ and from Corollary 1).
3) If $n > m \geq N$ then $\|\hat{\zeta}_n - \hat{\zeta}_m\| \leq \epsilon_m$ ($\epsilon_m \to 0$ as $N \to \infty$).
4) If $n \geq N$ then $a(n) < \frac{\delta_4 - \delta_3}{[K(1+C) + \epsilon_m]}$. Here C is a sample path dependent constant such that $\sup_{t \geq 0} \|\hat{x}(t)\| \leq C$.

Given T_l there exists k such that $T_{l+1} = t(m(l) + k + 1)$. Consider the following:

$$\hat{x}(T_{l+1}^-) = \hat{x}(t(m(l)+k)) + a(m(l)+k) \left(h_{r(l)}(\hat{x}(t(m(l)+k)), y_{m(l)+k}) + \hat{M}_{m(l)+k+1} \right).$$

Taking norms on both sides we get the following inequality,

$$\|\hat{x}(T_{l+1}^-)\| \leq \|\hat{x}(t(m(l)+k))\| + a(m(l)+k) \left(K(1 + \|\hat{x}(t(m(l)+k))\|) + \|\hat{M}_{m(l)+k+1}\| \right).$$

From the observations made earlier we get,

$$\|\hat{x}(T_{l+1}^-)\| \leq \delta_3 + \frac{\delta_4 - \delta_3}{[K(1+C) + \epsilon_m]} \left(K(1+C) + \epsilon_m \right).$$

Since $\|\overline{x}(T_l)\| = 1$ and $\|\hat{x}(T_{l+1}^-)\| < \delta_4$, we get

$$\frac{\|\overline{x}(T_{l+1}^-)\|}{\|\overline{x}(T_l)\|} = \frac{\|\hat{x}(T_{l+1}^-)\|}{\|\hat{x}(T_l)\|} < \delta_4 < 1. \tag{5}$$

Since $\|\overline{x}(T_l)\| \uparrow \infty$ and $\|\overline{x}(T_{l+1}^-)\| < \delta_4$. It follows that the iterates have to make larger and larger jumps within a single interval of length T. For all $m(l) > N$ the trajectory falls exponentially till it enters the ball of radius R_0. This implies that $x(T_l) \in \overline{B}_{R_0}(0)$ when $r(l-1) < r(l)$. Further, the trajectory made a jump of at least $r(l) - R_0$ within time interval T. This violates Gronwalls inequality. □

IV. Convergence analysis

In the previous section, we showed that the iterates given by (1) are bounded almost surely (stable), provided the conditions in Section II are satisfied. In this section, we show that our stability assumptions can be combined with additional assumptions, based on those found in Borkar [9], to provide a complete analysis of (1). In other words, the main result, Theorem 2, of this section states the following: Recursion (1) is bounded almost surely and converges to an internally chain transitive invariant set associated with a DI that is defined in terms of the ergodic occupation measures associated with the 'Markov process'. Below, we list the additional assumptions involved in our analysis.

(B1) $\{y_n\}_{n \geq 0}$ is an S-valued Markov process with two associated control processes: the iterate sequence $\{x_n\}_{n \geq 0}$ and another random process $\{z_n\}_{n \geq 0}$ taking values in a compact metric space U. Thus, for $n \geq 0$,

$$P\left(y_{n+1} \in A \mid y_m, z_m, x_m,\ m \leq n\right) = \int_A p(dy | y_n, z_n, x_n),$$

with A Borel in S. The map

$$(y, z, x) \in S \times U \times \mathbb{R}^d \to p(dw | y, z, x) \in \mathcal{P}(S)$$

is continuous, and further it is uniformly continuous in the x variable on compacts with respect to the other variables. $\mathcal{P}(S)$ is used to denote the space of probability measures on S.

Let $\varphi : S \to \varphi(S)$ with $\varphi(y, dz) \in \mathcal{P}(U)$ for each $y \in S$, be a measurable map. Suppose the Markov process has a (possibly non-unique) invariant probability measure $\eta_{x,\varphi}(dy) \in \mathcal{P}(S)$, we can define the corresponding *ergodic occupation measure*

$$\Psi_{x,\varphi}(dy, dz) := \eta_{x,\varphi}(dy)\varphi(dy, dz) \ \in \mathcal{P}(S \times U). \tag{6}$$

Let $D(x)$ be the set of all such ergodic occupation measures for a prescribed x. It can be shown that $D(x)$ is closed and convex for each $x \in \mathbb{R}^d$. Further, the map $x \mapsto D(x)$ is upper-semicontinuous. For a proof of the aforementioned results the reader is referred to *Chapter 6.2* of [10].

(B2) $D(x)$ is compact.

Let us define a $\mathcal{P}(S \times U)$-valued random process $\mu(t) = \mu(t, dy,\ dz)$, $t \geq 0$, by $\mu(t) := \delta_{y_n, z_n}, t \in [t(n), t(n+1))$, for $n \geq 0$. For $t > s \geq 0$, define $\mu_s^t \in \mathcal{P}(S \times U \times [s,t])$ by $\mu_s^t(A \times B) := \frac{1}{t-s} \int_B \mu(y, A)\, dy$ for A, B Borel in $S \times U$, $[s,t]$ respectively.

(B3) Almost surely, for $t > 0$, the set $\{\mu_s^{s+t}, s \geq 0\}$ remains tight.

Define $\widetilde{h}(x, \nu) := \int h(x, y)\nu(dy, U)$ for $\nu \in \mathcal{P}(S \times U)$. We use this to define the following *DI*.

$$\dot{x}(t) \in \hat{h}(x(t)),\ \text{where } \hat{h}(x) := \{\widetilde{h}(x, \nu) \mid \nu \in D(x)\}. \tag{7}$$

We are now ready to state the main result of this paper.

Theorem 2 (Stability & Convergence)**.** *Under assumptions $(A1) - (A3)$, $(S1)$, $(S2)$ and $(B1) - (B3)$, almost surely the iterates given by (1) are stable and converge to an internally chain transitive invariant set associated with $\dot{x}(t) \in \hat{h}(x(t))$.*

Proof. Under assumptions $(A1) - (A3)$, $(S1)$ and $(S2)$, (1) is stable as a consequence of Theorem 1. It now follows from

Theorem 3.1 of Borkar [9] that the iterates converge to an internally chain transitive invariant set associated with $\dot{x}(t) \in \hat{h}(x(t))$. □

V. Applications

In this section, we present two applications of the theory developed in Sections II - IV. First, we analyze a generalized form of $TD(0)$, a simple yet effective TD algorithm. Second, we analyze an online-TD formulation for supervised learning with delayed feedbacks.

A. Analysis of generalized $TD(0)$

Temporal difference (TD) learning is a popular prediction algorithm. In this section, we consider a generalized version of online $TD(0)$, a simple yet effective TD algorithm. Given a policy π, $TD(0)$ iteratively updates it's estimate V of V^π, where V^π is the expected total reward associated with policy π. When a non-terminal state s_n is visited at time n, the algorithm updates the estimate, $V(s_n)$, based on what happens 'after the visit'. The $TD(0)$ algorithm is given by,

$$V(s_n) := V(s_n) + a(n)(r_{n+1} + \gamma V(s_{n+1}) - V(s_n)). \quad (8)$$

In the above, s_n is the current state, s_{n+1} is the next state, r_{n+1} is the observed reward and γ is the discount factor. Note that the starting state, s_0, is arbitrarily chosen. Define $\mathcal{F}_n := \sigma(s_0, \ldots, s_n, r_1, \ldots, r_n)$ for $n \geq 1$ and $\mathcal{F}_0 := \sigma(s_0)$. Also, define $M_{n+1} := [r_{n+1} + \gamma V(s_{n+1})] - E[r_{n+1} + \gamma V(s_{n+1})|\mathcal{F}_n]$ and $V^\pi(s_n) := E[r_{n+1} + \gamma V(s_{n+1})|\mathcal{F}_n]$ for $n \geq 0$.

In this section we assume that s_n belongs to a compact state space, S. Further, without loss of generality we may also assume that

$$V^\pi(s_n) - V(s_n) = \Psi(s_n)V(s_n) + \psi(s_n), \quad (9)$$

where $\Psi : S \to \mathbb{R}^{d\times d}$ and $\psi : S \to \mathbb{R}^d$. In other words, $TD(0)$ given by (8) can be rewritten as:

$$V(s_n) := V(s_n) + a(n)(\Psi(s_n)V(s_n) + \psi(s_n) + M_{n+1}). \quad (10)$$

For a detailed exposition on temporal difference algorithms the reader is referred to Tsitsiklis and Van Roy [17]. In this section, we impose conditions on maps Ψ and ψ that guarantee the 'stability and convergence' of the iterates given by (8).

REMARK 1. *It is important to note that our $TD(0)$ algorithm (cf. 8) is more general than the regular $TD(0)$ update with function approximation, as in (say) Tsitsiklis and Van Roy [17]. In particular, the regular $TD(0)$ with function approximation can be written (see [17]) as in (8). Note also that unlike the usual analyses of $TD(0)$, we do not assume that the Markov process $\{s_n\}_{n\geq 0}$ **(a)** has a finite state and **(b)** is ergodic under the given stationary policy.*

We state the first of our two assumptions below.

(T1) $\Psi : S \to \mathbb{R}^{d\times d}$ and $\psi : S \to \mathbb{R}^d$ are continuous maps.

If we define a map $h : \mathbb{R}^d \times S \to \mathbb{R}^d$ as $h(v, s) \mapsto \Psi(s)v + \psi(s)$, then we have the following lemma.

Lemma 6. *The map h defined above is Lipschitz continuous in the first component. Further, the Lipschitz constant does not vary with the second component i.e., h satisfies assumption $(A1)$.*

Proof. Since Ψ is continuous, it's range, $\Psi(S) \subset \mathbb{R}^{d\times d}$, is compact. It follows that

$$\|h(v_1, s) - h(v_2, s)\| \leq \|\Psi(s)\| \times \|v_1 - v_2\| \leq L\|v_1 - v_2\|,$$

where $L := \sup_{M\in A(S)} \|M\| (< \infty)$. Clearly, L is independent of the second component. □

For the purpose of this section we assume that the martingale sequence defined earlier in this section, $\{M_n\}_{n\geq 0}$, satisfies assumption $(A2)$ and the step-size sequence, $\{a(n)\}_{n\geq 0}$ satisfies assumption $(A3)$. We are now ready to define the rescaled family of functions. For $c \geq 1$, define $h_c(v, s) := \Psi(s)v + \psi(s)/c$, then the upper-limit of $\{h_c\}_{c\geq 1}$ is given by $h_\infty(v, s) = \{\Psi(s)v\}$.

Lemma 7. *(8) satisfies assumption $(S1)$.*

Proof. Let $c_n \uparrow \infty$, $s_n \to s$ and $\lim_{n\to\infty} h_{c_n}(v, s_n) = u$. We need to show that $u = h_\infty(v, s)$. Since Ψ is continuous, $\lim_{n\to\infty} \Psi(s_n)v = \Psi(s)v$. Since ψ is a bounded function, $\lim_{n\to\infty} \psi(s_n)/c_n = 0$. Hence, we get $u = \lim_{n\to\infty} h_{c_n}(v, s_n) = \Psi(s)v \in h_\infty(v, s)$. □

The following technical result is needed to state our second and final assumption.

Lemma 8. *Let $H : \mathbb{R}^d \to \{$subsets of $\mathbb{R}^d\}$ be a Marchaud map. Let $\mathcal{A}$ be an associated attracting set that is also Lyapunov stable. Let $\mathcal{B}$ be a compact subset of the basin of attraction of $\mathcal{A}$. Then for all $\epsilon > 0$ there exists $T(\epsilon)$ such that $\Phi_{[T(\epsilon),\infty)}(\mathcal{B}) \subseteq N^\epsilon(\mathcal{A})$.*

Proof. Since $\mathcal{A}$ is Lyapunov stable, corresponding to $N^\epsilon(\mathcal{A})$ there exists $N^\delta(\mathcal{A})$ such that $\Phi_{[0,+\infty)}(N^\delta(\mathcal{A})) \subseteq N^\epsilon(\mathcal{A})$. Fix $x_0 \in \mathcal{B}$. Since $\mathcal{B}$ is contained in the basin of attraction of $\mathcal{A}$, $\exists t(x_0) > 0$ such that $\Phi_{t(x_0)}(x_0) \subseteq N^{\delta/4}(\mathcal{A})$. Further, from the upper semi-continuity of flow it follows that, for all $x \in N^{\delta(x_0)}(x_0)$, $\Phi_{t(x_0)}(x) \subseteq N^{\delta/4}(\Phi_{t(x_0)}(x_0))$ for some $\delta(x_0) > 0$, see *Chapter 2* of Aubin and Cellina [2]. Hence $\Phi_{t(x_0)}(x) \subseteq N^\delta(\mathcal{A})$ for all $x \in N^{\delta(x_0)}(x_0)$. Since $\mathcal{A}$ is Lyapunov stable, we get $\Phi_{(t(x_0),+\infty]}(x) \subseteq N^\epsilon(A)$. In this manner for each $x \in \mathcal{B}$ we calculate $t(x)$ and $\delta(x)$, the collection $\left\{N^{\delta(x)}(x) : x \in \mathcal{B}\right\}$ is an open cover for $\mathcal{B}$. Let $\left\{N^{\delta(x_i)}(x_i) \mid 1 \leq i \leq m\right\}$ be a finite sub-cover. If we define $T(\epsilon) := \max\{t(x_i) \mid 1 \leq i \leq m\}$ then $\Phi_{[T(\epsilon),+\infty)}(\mathcal{B}) \subseteq N^\epsilon(\mathcal{A})$. □

If we define $H(v) := \overline{co}(\{\Psi(s)v \mid s \in S\})$ for all $v \in \mathbb{R}^d$, then it follows from Lemma 2 that H is Marchaud. We state our second assumption below.

(T2) Let $\epsilon > 0$ and $\mathcal{V} : \overline{B}_{1+\epsilon}(0) \to [0, \infty)$. Let Λ be a compact subset of $B_1(0)$, clearly $\sup_{u\in\Lambda} \|u\| < 1$. The following hold: **(i)** For all $t \geq 0$, $\Phi_t(B_{1+\epsilon}(0)) \subseteq B_{1+\epsilon}(0)$, where $\Phi_t(\cdot)$ is a solution to the DI $\dot{x}(t) \in H(x(t))$; **(ii)** $\mathcal{V}^{-1}(0) = \Lambda$; **(iii)**

$\mathcal{V}$ is a continuous map. For all $x \in \overline{B}_{1+\epsilon}(0) \setminus \Lambda$, $y \in \Phi_t(x)$ and $t > 0$ we have $\mathcal{V}(y) < \mathcal{V}(x)$.

Proposition 3.25 from Benaïm et. al. [6] : Under $(T2)$, Λ *is a Lyapunov stable attracting set, further there exists an attractor,* $\mathcal{A}$*, contained in* Λ *whose basin of attraction contains* $B_{1+\epsilon}(0)$.

Lemma 9. *(8) satisfies assumption* $(S2)$.

Proof. Since $\mathcal{A} \subset \Lambda$ and $B_{1+\epsilon}(0)$ is contained in the basin of attraction of $\mathcal{A}$, it follows that $\overline{B}_1(0) \subset B_{1+\epsilon}(0)$ is contained in the basin of attraction of Λ. Since $\overline{B}_1(0)$ is compact it follows from Lemma 8 that it is contained in some fundamental neighborhood of Λ. In this section, the attracting set associated with $\dot{x}(t) \in \overline{co}\left(\{A(y)x(t) \mid y \in S\}\right)$ in $(S2)$ is Λ. □

The following theorem is immediate.

Theorem 3. *Under assumptions* $(A1)$*-*$(A3)$*,* $(T1)$*,* $(T2)$ *&* $(B1)$*-*$(B3)$*, almost surely the iterates given by (8) are stable and converge to an internally chain transitive invariant set associated with* $\dot{x}(t) \in \hat{h}(x(t))$ *(*$\hat{h}$ *is defined in Section IV).*

Let us analyze the special case when Ψ is a constant map *i.e.,* $\Psi(s) = M$ for some fixed $M \in \mathbb{R}^{d \times d}$, $s \in S$. The recursion given by (8) becomes:

$$V(s_n) := V(s_n) + a(n)\left[MV(s_n) + \psi(s_n) + M_{n+1}\right], \quad (11)$$

and like before, $\psi : S \to \mathbb{R}^d$ is a continuous map. Clearly, (11) satisfies $(T1)$, $(A1)-(A3)$ and $(S1)$. The rescaled family of functions are $h_c(v,s) = Mv + \psi(s)/c$ and $h_\infty(v,s) = Mv$; $H(v) = \overline{co}\left(\bigcup_{s \in S} h_\infty(v,s)\right) = \{Mv\}$. Hence, the *DI* $\dot{v}(t) \in H(v(t))$ is really the *o.d.e.* $\dot{v}(t) = Mv$, here.

If we assume that all eigenvalues of M have strictly negative real parts, then origin is the unique globally asymptotic stable equilibrium point (a globally attracting set that is also Lyapunov stable) of $\dot{x}(t) = Mx(t)$ (see *(11.2.3)* of Borkar [10]).

We define the following: (a) $\epsilon := 1$; (b) $\mathcal{V}(v) : \overline{B}_2(0) \to [0,\infty)$ as $\mathcal{V}(v) := \|v\|$; (c) $\Lambda := \{0\}$.

Solving the *o.d.e.* $\dot{x}(t) = Mx(t)$, we get $\Phi_t(x(0)) = e^{Mt}x(0)$ for $t \geq 0$. Since all the eigenvalues of M have strictly negative real parts, we have $\|\Phi_t(x(0))\| < \|x(0)\|$, where $t > 0$ and $x(0) \in \mathbb{R}^d \setminus \{0\}$. Hence $\Phi_t(B_2(0)) \subseteq B_2(0)$ *((T2)(i) is satisfied)*. It follows from the definition of $\mathcal{V}$ (see (b)) that $\mathcal{V}^{-1}(0) = \Lambda$ *((T2)(ii) is satisfied)*. Finally, fix $v_0 \in B_2(0) \setminus \{0\}$ and $t > 0$, we have $\|\Phi_t(v_0)\| < \|v_0\|$, hence $\mathcal{V}(\Phi_t(v_0)) < \mathcal{V}(v_0)$ *((T2)(iii) is satisfied)*. Now, it follows from Theorem 3 that the iterates given by (11) are 'stable and convergent'.

B. A quick note on our assumptions

In this section, we compare our framework with that of [17]. Traditional analysis of TD by Tsitsiklis and Van Roy [17] requires the following assumptions, among others: **(i)** The state space is finite. **(ii)** The Markov chain, associated with state evolution, is ergodic. **(iii)** The second moment of the single stage reward function is bounded, *i.e.,* $\mathbb{E}r_{n+1}^2 < \infty$ for $n \geq 0$. These conditions are very restrictive in real-world applications. Further, in [17], the cost-to-go function is assumed to have the following form: $J(i_0,\theta) \approx \sum_{k=1}^{K} \theta(k)\phi_k(i_0)$, where $\{\phi_k\}_{1 \leq k \leq K}$ are the basis functions.

Below we state the additional stability assumptions in [17], following which we briefly discuss how our assumptions differ from [17]. Note that the state space S in [17] is finite.

TB1 There exists a function $f : S \mapsto \mathbb{R}^+$ satisfying the following requirements: for all i_0, $1 \leq k \leq K$ and $m \geq 0$, we have $\sum_{\tau=0}^{\infty} \|\mathbb{E}\left[\phi_k(i_\tau)\phi_k'(i_{\tau+m})\right] - \mathbb{E}\left[\phi_k(i_t)\phi_k'(i_{t+m})\right]\| \leq f(i_0)$ and $\sum_{\tau=0}^{\infty} \|\mathbb{E}\left[\phi_k(i_\tau)r_{\tau+m+1}\right] - \mathbb{E}\left[\phi_k(i_t)r_{t+m+1}\right]\| \leq f(i_0)$, where r_{n+1} is the reward at time n.

TB2 Given any $q > 0$, there exists μ_q such that for all i_0 and t: $\mathbb{E}\left[f^q(i_t) \mid i_0\right] \leq \mu^q f^q(i_0)$.

The DI from $(S2)$ of Section II is given by, $H(x) := \overline{co}\left(\bigcup_{y \in S} h_\infty(x,y)\right)$ and $h_\infty(x,y) = Limsup_{c \to \infty}\{h_c(x,y)\}$.

As stated in [17], $TB1$ and $TB2$ can be verified when the state space is finite. Further, $TB1$ and $TB2$ imply stability of TD when combined with previously stated assumptions, such as bounded second moments of the basis and reward functions, ergodicity of the Markov chain, etc., see [17] for details. On the other hand, our stability assumptions, $(S1)$ and $(S2)$, do not distinguish between finite and infinite state spaces. If S is a compact metric space or if the associated Markov chain is not ergodic, our framework is readily applicable as opposed to [17].

REMARK 2. *In Section V-C, our theory is used to provide an analysis of TD for supervised learning with delayed feedbacks. For the problem considered therein, the controlled Markov process evolves in a compact state space. In other words, the framework of [17] cannot be used to provide an analysis of this algorithm.*

C. Stability of online TD for supervised learning

In this section, we consider the following weather forecasting problem described in *Chapter 11* of Spall [16]. On every day of the week, we are interested in issuing a forecast for the weather on the following Saturday, using current and past weather conditions and meteorological indicators. Unlike traditional supervised learning, in a TD formulation, the predictor is trained using successive predictions.

We consider the online implementation of TD for supervised learning with delayed feedback, presented in Chapter 11 of Spall [16]. For this purpose we may use a predictor $f(\theta,\cdot)$, that is parameterized by θ. In other words, f uses a sequence of inputs $\{x_0, x_1, \ldots, x_N\}$ to predict some (outcome) Z. The predictor f is trained to minimize the following expected mean-squared error:

$$\frac{1}{2}\mathbb{E}\left[Z - f(\theta, x_n)\right]^2. \quad (12)$$

Within the context of our weather forecasting problem, Z is Saturday's weather outcome. The x_is represent the current and past weather conditions, among others, that are used as input

for prediction. We consider the following $TD(\lambda)$ approach to training the predictor f, in an online manner, see [16]:

$$\theta_{n+1} = \theta_n + a(n)\left[f(\theta_n, x_{n+1}) - f(\theta_n, x_n)\right] \sum_{i=0}^{n} \lambda^{n-i} \left[\nabla_\theta f(\theta, x_i)\right]_{\theta=\theta_n}, \tag{13}$$

where **(i)** $0 \leq \lambda \leq 1$ is a fading factor which gives lower importance to past observations, **(ii)** $\{a(n)\}_{n\geq 0}$ is the standard step-size sequence.

When $\lambda = 0$, (13) becomes $TD(0)$. Note that $TD(0)$ does not utilize older predictions in the current training step. Rewriting (13) with $\lambda = 0$, we get:

$$\theta_{n+1} = \theta_n + a(n)\left[f(\theta_n, x_{n+1}) - f(\theta_n, x_n)\right] \left[\nabla_\theta f(\theta, x_n)\right]_{\theta=\theta_n}. \tag{14}$$

Let us suppose that the predictor is a linear regression function, *i.e.,* $f(\theta, x) = \theta^T x$. For this to be effective, it is imperative to embed the input variables in a higher dimensional feature space. This embedding or feature extraction is often done using deep neural networks. Let $\phi(\cdot)$ be the given feature function, then we may rewrite (14) as:

$$\theta_{n+1} = \theta_n + a(n)\theta_n^T \left(\phi(x_{n+1}) - \phi(x_n)\right)\phi(x_n). \tag{15}$$

There are numerous results in literature which discuss the convergence of (15). The reader is referred to [17] or [8] for details. However, all these results prove convergence of the algorithm under strong assumptions. Further it is hard to ensure stability, especially since the state space is continuous. Below, we present a complete analysis using our theory. Note that we consider the state space to be continuous, unlike previous analyses.

First, let us define the following objective function, h:

$$h(\theta, z_1, z_2) := \left(\theta^T z_1\right) z_2, \tag{16}$$

where $z_1 = \phi(y_0) - \phi(y_1)$ and $z_2 = \phi(y_2)$ for some $\phi(y_0), \phi(y_1)$ and $\phi(y_2)$ in $\mathcal{K}$. Let us also define $S := (\mathcal{K} - \mathcal{K}) \times \mathcal{K}$, where $\mathcal{K} - \mathcal{K} = \{x - y \mid x, y \in \mathcal{K}\}$. In other words, h is a function with domain $\mathbb{R}^d \times S$ and range $\mathbb{R}^d$. Now, (15) can be rewritten as:

$$\theta_{n+1} = \theta_n + a(n)h(\theta_n, y_n), \tag{17}$$

where $y_n = ((\phi(x_{n+1}) - \phi(x_n)), \phi(x_n))$ and, so from (16), $h(\theta_n, y_n) = \theta_n^T \left(\phi(x_{n+1}) - \phi(x_n)\right)\phi(x_n)$. *Note that the $\{y_n\}_{n\geq 0}$ process is the controlled Markov process of $(A1)(ii)$ in Section II.* We make the following assumptions on (15):

(A) The feature extracted input variables $\phi(x_n)$ belong to $\mathcal{K}$, which is a compact subset of $\mathbb{R}^d$.

(B) The differential inclusion $\dot{\theta}(t) \in \underset{y\in S}{\cup} h(\theta, y)$ has an attractor inside the unit ball centered at the origin, such that its fundamental neighborhood is the closed unit ball itself.

If we consider the previously mentioned weather forecasting problem, the input vector consists of bounded quantities such as atmospheric pressure, temperature, etc. Hence assumption (A) is satisfied here as with many problems. It is also worth observing that assumption (B) is a recast of $(S2)$ for (15).

In Section II, we have presented easily verifiable sufficient conditions for the stability of general recursions such as (15)/(17). It follows from the above discussion that (15) satisfies $(A1) - (A3)$. In other words, to ensure stability of (15), it is sufficient to show that $(S1)$ and $(S2)$ are satisfied.

It follows from the above definition of h, that the rescaled family of functions $\{h_c \mid c \geq 1\}$, are such that $h_c(\theta, y) = h(\theta, y)$ for $\theta \in \mathbb{R}^d$, $y \in S$ and $c \geq 1$. Hence $h_\infty = h$. Since h is continuous in the y-coordinate, (15) satisfies $(S1)$. Suppose one is able to show that (15) also satisfies $(S2)$, then it would follow from Theorem 1, that (15) is stable. As stated earlier, it may be observed that $(S2)$ and (B) are equivalent. One way to show that (15) satisfies $(B)/(S2)$, is to construct an associated Lyapunov function and use Proposition 3.25 of Benaïm, Hofbauer and Sorin [6]. Constructing a Lyapunov function is often problem dependent. In summary, to ensure stability of (15), one can check that $\dot{\theta}(t) \in \underset{y\in S}{\cup} h(\theta, y)$ has an attractor inside the unit ball centered at the origin, such that its fundamental neighborhood is the unit ball itself.

Once stability is assured, we then proceed to prove convergence of the stochastic iterates in Section IV. *The above presented analysis can further be readily extended to $TD(\lambda)$ implementations of supervised learning with non-linear predictors, i.e., for the case of $\lambda \neq 0$.* Finally, our theory can be readily used to analyze online $TD(0)$ with linear function approximator and finite state space, see *Section 5.3* of [13].

REFERENCES

[1] C. Andrieu, V.B. Tadić, and M. Vihola. On the stability of some controlled Markov chains and its applications to stochastic approximation with markovian dynamic. *The Annals of Applied Probability*, 25(1):1–45, 2015.

[2] J. Aubin and A. Cellina. *Differential Inclusions: Set-Valued Maps and Viability Theory*. Springer, 1984.

[3] M. Benaïm. A dynamical system approach to stochastic approximations. *SIAM J. Control Optim.*, 34(2):437–472, 1996.

[4] M. Benaïm. Dynamics of stochastic approximation algorithms. *Seminaire de probabilites XXXIII*, pages 1–68, 1999.

[5] M. Benaïm and M. W. Hirsch. Asymptotic pseudotrajectories and chain recurrent flows, with applications. *J. Dynam. Differential Equations*, 8:141–176, 1996.

[6] M. Benaïm, J. Hofbauer, and S. Sorin. Stochastic approximations and differential inclusions. *SIAM Journal on Control and Optimization*, pages 328–348, 2005.

[7] A. Benveniste, M. Metivier, and P. Priouret. *Adaptive Algorithms and Stochastic Approximations*. Springer Publishing Company, Incorporated, 1st edition, 2012.

[8] D.P. Bertsekas and J.N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1st edition, 1996.

[9] V. S. Borkar. Stochastic approximation with 'controlled Markov' noise. *Systems & Control Letters*, 55(2):139–145, 2006.

[10] V. S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008.

[11] V. S. Borkar and S.P. Meyn. The O.D.E. method for convergence of stochastic approximation and reinforcement learning. *SIAM J. Control Optim*, 38:447–469, 1999.

[12] L. Ljung. Analysis of recursive stochastic algorithms. *Automatic Control, IEEE Transactions on*, 22(4):551–575, 1977.

[13] A. Ramaswamy and S. Bhatnagar. arxiv:1504.06043.

[14] A. Ramaswamy and S. Bhatnagar. A generalization of the Borkar-Meyn theorem for stochastic recursive inclusions. *Mathematics of Operations Research*, 42(3):648–661, 2016.

[15] A. Ramaswamy and S. Bhatnagar. Analysis of gradient descent methods with nondiminishing bounded errors. *IEEE Transactions on Automatic Control*, 63(5):1465–1471, 2018.

[16] J.C. Spall. *Introduction to stochastic search and optimization: estimation, simulation, and control*, volume 65. John Wiley & Sons, 2005.

[17] J. N. Tsitsiklis and B. Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 1997.