Automatica 79 (2017) 108–114

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Brief paper

A stability criterion for two timescale stochastic approximation schemes☆

Chandrashekar Lakshminarayanan, Shalabh Bhatnagar

Department of Computer Science and Automation, Indian Institute of Science, Bangalore-560012, India

ARTICLE INFO

Article history:
Received 8 June 2015
Received in revised form
5 June 2016
Accepted 27 November 2016
Available online 6 March 2017

Keywords:
Simulation
Two-timescale stochastic approximation
Stability of iterates
Limiting ODE
Reinforcement learning

ABSTRACT

We present *the first* sufficient conditions that guarantee stability of two-timescale stochastic approximation schemes. Our analysis is based on the ordinary differential equation (ODE) method and is an extension of the results in Borkar and Meyn (2000) for single-timescale schemes. As an application of our result, we show the stability of iterates in a two-timescale stochastic approximation scheme arising in reinforcement learning.

1. Introduction

Many applications require stochastic computations in *nested* loops. For instance, consider parameter tuning in the case of a queuing system. Here, the performance of the system for a given parameter setting needs to be estimated first and then the parameters have to be tuned based on the estimation. Thus the estimation constitutes the *inner* loop and the parameter tuning, the *outer* loop. Another example of such nested computations arises in reinforcement learning (especially actor-critic algorithms Bhatnagar, Sutton, Ghavamzadeh, & Lee, 2009), wherein, the inner loop estimates the value of the agent's behavior while the outer loop tunes the agent's behavior. A large class of such nested stochastic computations fall under the category of two timescale stochastic approximation schemes (Bhatnagar, 2005; Bhatnagar & Borkar, 1998; Bhatnagar, Fu, Marcus, & Wang, 2003; Borkar, 1997) given as under.

$$x_{n+1} = x_n + a(n)[h(x_n, y_n) + M^{(1)}_{n+1}], \tag{1a}$$

$$y_{n+1} = y_n + b(n)[g(x_n, y_n) + M^{(2)}_{n+1}], \tag{1b}$$

where, $x_n \in \mathbf{R}^{d_1}$, $y_n \in \mathbf{R}^{d_2}$, $n \geq 0$ are the iterates, $h: \mathbf{R}^{d_1+d_2} \to \mathbf{R}^{d_1}$, $g: \mathbf{R}^{d_1+d_2} \to \mathbf{R}^{d_2}$ are Lipschitz continuous functions, $M^{(i)}_n$, $i = 1, 2$ are (martingale difference) noise terms and $a(n)$ and $b(n)$ are both diminishing step-size schedules such that $\frac{b(n)}{a(n)} \to 0$ as $n \to \infty$. The term *stochastic approximation* (SA) signifies that the quantities $h(\cdot)$ and $g(\cdot)$ are corrupted by noise. The term *two timescale* signifies the fact that there are two different step-size schedules namely $a(n)$ and $b(n)$, with x_n and y_n being the corresponding *faster* and *slower* timescale iterates respectively (due to the relative magnitude of the step-sizes). The condition $\frac{b(n)}{a(n)} \to 0$ ensures that, at any given instant, the slower timescale iterates do not change as much compared to the faster timescale iterates, thereby producing an analogous effect as a *nested*-loop.

Under certain suitable conditions (Borkar, 2008; Tadić, 2004) one can analyze the convergence of the iterates in (1). For instance, in the case of reinforcement learning (RL), y_n dictates the agent's behavior i.e., policy, and x_n estimates the value accumulated by the behavior. Thus for a given RL algorithm (Bhatnagar et al., 2009), it is of interest to show that $y_n \to y^*, x_n \to x^*$, where $y^* \in \mathbf{R}^{d_2}$ and $x^* \in \mathbf{R}^{d_1}$ are the best agent's behavior/policy and the value the agent accumulates under the same. An important condition required to show convergence (Borkar, 2008; Tadić, 2004) is the boundedness/stability of the iterates i.e., $\sup_n \|x_n\| < \infty$ and $\sup_n \|y_n\| < \infty$. It is always possible to project the iterates after each update onto a compact set $\mathcal{C} \subset \mathbf{R}^{d_1+d_2}$ to ensure their boundedness. However, a shortcoming of such an approach is that the desired solution (i.e. (x^*, y^*)) might lie outside of the compact set $\mathcal{C}$.

☆ The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Valery Ugrinovskii under the direction of Editor Ian R. Petersen.

E-mail addresses: chandrurec5@gmail.com (C. Lakshminarayanan), shalabh@csa.iisc.ernet.in (S. Bhatnagar).

http://dx.doi.org/10.1016/j.automatica.2016.12.014
0005-1098/

Thus it is of interest and importance to be able to analytically establish the stability of the iterates (without making use of projections).

In this paper, we present the conditions that imply stability of iterates in a two timescale SA scheme, and apply our analysis to establish the stability of iterates in a two-timescale SA scheme arising in an application in reinforcement learning (Bhatnagar et al., 2009). Our analysis is based on the ordinary differential equation (ODE) method (Borkar, 1997, 2008; Borkar & Meyn, 2000) for stochastic approximation schemes. In what follows, in Section 2 we first present an overview of the convergence and stability analysis (Borkar & Meyn, 2000) of single timescale SA schemes (a degenerate case of two timescale SA schemes). In Section 3, we state the assumptions as well as discuss our contributions. Sections 4, 5.1 and 5.2 contain the main body of our analysis (Theorems 7 and 10). In Section 6, we use our analysis to show the stability of iterates for a two timescale RL algorithm presented in Bhatnagar et al. (2009) and also present a numerical example. And finally we provide our concluding remarks in Section 7.

2. Overview of the ODE method

Consider the following (single timescale) SA scheme:

$$z_{n+1} = z_n + a(n)[f(z_n) + M_{n+1}], \quad n \geq 0, \tag{2}$$

where $z_n \in \mathbf{R}^d$ are the iterates, $f : \mathbf{R}^d \to \mathbf{R}^d$ is a Lipschitz continuous function, M_{n+1} are zero-mean noise terms and $a(n)$ are diminishing step-sizes. Notice that (2) is a degenerate version of (1), i.e., it has only one step-size schedule $a(n)$ as opposed to two step-sizes $a(n)$ and $b(n)$. The ODE method has been employed to analyze the convergence of single timescale SA schemes (Borkar & Meyn, 2000) as well as two timescale SA schemes (Borkar, 1997). While an ODE based stability analysis for single timescale SA schemes exists (Borkar & Meyn, 2000), there exists no ODE based stability analysis in the literature for two timescale SA schemes. Before we present our stability analysis for two timescale SA schemes, we present an overview of the ideas involved in the convergence and the stability analysis of single timescale SA schemes in this section. The presentation here is only aimed at motivating the ODE method and we refer the reader to Borkar (2008) for a detailed presentation.

Convergence analysis: The ODE method considers the iterates $\{z_n\}$ of (2) as a noisy-discretized version of the trajectory $\phi(t, z_0)$, $t \geq 0$ of the ODE $\dot{z}(t) = f(z(t))$ with initial condition $z_0 \in \mathbf{R}^d$. In order to study $\{z_n\}$, $n \geq 0$, which is a countable set, and the trajectory $\phi(t, z_0)$, that is a function of continuous time t (for a given initial condition z_0), we need the notions of *timescale* and *interpolated trajectory*. The time instants that define timescale are $t(n) \stackrel{\text{def}}{=} \sum_{i=0}^{n-1} a(i)$ and the interpolated trajectory is defined as $\bar{z}(t) = z_n + (z_{n+1} - z_n)\frac{t-t(n)}{t(n+1)-t(n)}$, $t \in [t(n), t(n+1)]$. We can then show that $\bar{z}(t)$ 'tracks' $\phi(t, z_0)$ (Lemma 1, Chapter 2, Borkar, 2008), i.e. the difference between them can be made arbitrarily small. Since, z_ns are embedded in $\bar{z}(t)$, it can be said that $\{z_n\}$ tracks the ODE $\dot{z}(t) = f(z(t))$. The convergence analysis can be captured formally (i.e., the statement of Theorem 2, Chapter 2, Borkar, 2008) as follows.

Theorem 1. *Almost surely, the sequence $\{z_n\}$ generated by (2) converges to a (possibly sample path dependent) compact connected internally chain transitive invariant set of the ODE $\dot{z}(t) = f(z(t))$.*

In many applications, we have $f(z) = \nabla J(z)$, for some performance measure $J : \mathbf{R}^d \to \mathbf{R}^d$, and in such cases Theorem 1 implies that the iterates will converge to the set $\{z : \nabla J(z) = 0\}$, i.e., the iterates converge to a local extremum or stay within a compact connected set of local extrema.

There are two sources of error that need to be contained while analyzing the iterates of an SA scheme via its corresponding ODE. These are, the discretization error, and the error due to the noise term. Convergence analysis based on the ODE method can be carried out under suitable assumptions ((A1)–(A4), Chapter 2 of Borkar, 2008) involving boundedness of the iterates, Lipschitz continuity of f, non-summability and square summability of the step-sizes and quadratic variation of the martingale noise terms to contain these errors. In particular, Lipschitz continuity helps in containing the discretization error and the conditions on the quadratic variation and step-sizes help to contain the error due to the noise.

Stability analysis: The *rescaling* technique (Chapter 3, Borkar, 2008) is an important method used to establish the stability of single timescale SA schemes. The aim of the stability analysis is to show that the iterates are bounded (the convergence analysis needs this as an assumption see (A4), Chapter 2 of Borkar, 2008, Theorems 1–8 of Tadić, 2004). Since the boundedness of iterates is not known, in the case when the iterates go out of the unit ball around the origin, we can rescale the iterates into the unit ball using an appropriate scaling factor $c \geq 1$. This rescaling can be done every $T > 0$ instants in a periodic manner. This periodic rescaling gives rise to scaled iterates $\hat{z}_n$, scaled functions $f_c(z) \stackrel{\text{def}}{=} f(cz)/c$, $c \geq 1$, scaled interpolated trajectory $\hat{z}(t)$ and a scaled ODE $\dot{\hat{z}}(t) = f_c(\hat{z}(t))$. The convergence analysis holds for the scaled iterates (since their boundedness is ensured by construction). Consider the case when the iterates are not bounded, then the iterates have to keep leaving the unit ball, which in turn implies that the scaling factor $c \to \infty$. Then the iterates have to track the limiting ODE given by $\dot{z}(t) = f_\infty(z(t))$, where $f_\infty(z) \stackrel{\text{def}}{=} \lim_{c\to\infty} f_c(z)$. Now, if the limiting ODE has the origin as its unique globally asymptotic stable equilibrium, then the iterates have to fall at an exponential rate into the unit ball. Thus, in order to escape to infinity, the sizes of the jumps from within the unit ball to its outside should be unbounded, but an application of Gronwall's inequality (see Lemma 6, Chapter 3, Borkar, 2008) shows that this cannot happen. This in turn implies that the iterates are stable and bounded. This assumption of existence of stable equilibrium on the limiting ODE is weaker than assuming stable equilibrium for the original ODE (consider the case when $f^1(z) = -z + \cos(z)$ for which the corresponding $f^1_\infty(z) = -z$).

The stability analysis (Chapter 3 of Borkar, 2008) requires assumptions ((A5), Chapter 3 of Borkar, 2008) on the existence of scaled and limiting functions and the corresponding ODEs. Since the iterates track a scaled ODE, the analysis also needs results (Lemmas 1, 2 and corollary 3, Chapter 3 of Borkar, 2008) that show that the scaled ODE tracks the limiting ODE for large enough c.

3. Assumptions and contributions

The stability analysis available for single timescale SA schemes is not directly applicable for two timescale SA schemes. In particular, the iterates evolve in both the timescales simultaneously and it is not enough to rescale in the faster or the slower timescale alone. As it turns out, in this paper, we rescale the iterates in both the timescales and the analysis involves scaled as well as limiting ODEs in the faster as well as the slower timescales. To this end, we assume the following conditions (A1–A5) and show that they imply the stability of the iterates in the two timescale SA scheme in (1).

A1 $h : \mathbf{R}^{d_1+d_2} \to \mathbf{R}^{d_1}$ and $g : \mathbf{R}^{d_1+d_2} \to \mathbf{R}^{d_2}$ are Lipschitz continuous functions.

A2 $\{M_n^{(1)}\}$, $\{M_n^{(2)}\}$ are martingale difference sequences w.r.t. the increasing sequence of σ-fields $\{\mathcal{F}_n\}$, where $\mathcal{F}_n \stackrel{\text{def}}{=} \sigma(x_m, y_m, M_m^{(1)}, M_m^{(2)}, m \leq n)$, $n \geq 0$, with $\mathbf{E}[\|M_{n+1}^{(i)}\|^2 | \mathcal{F}_n] \leq K(1 + \|x_n\|^2 + \|y_n\|^2)$, $i = 1, 2$, $n \geq 0$, for some constant $K > 0$.

A3 $\{a(n)\}$, $\{b(n)\}$ are two step-size schedules satisfying $a(n) > 0$, $b(n) > 0$, $\sum_n a(n) = \sum_n b(n) = \infty$, $\sum_n (a(n)^2 + b(n)^2) < \infty$, $\frac{b(n)}{a(n)} \to 0$ as $n \to \infty$.

A4 The functions $h_c(x, y) \stackrel{\text{def}}{=} \frac{h(cx,cy)}{c}$, $c \geq 1$, satisfy $h_c \to h_\infty$ as $c \to \infty$, uniformly on compacts for some h_∞. Also, the *limiting* ODE $\dot{x}(t) = h_\infty(x(t), y)$ has a unique globally asymptotically stable equilibrium (a.s.e.) $\lambda_\infty(y)$, where $\lambda_\infty : \mathbf{R}^{d_2} \to \mathbf{R}^{d_1}$, is a Lipschitz map. Further $\lambda_\infty(0) = 0$, i.e., the ODE $\dot{x}(t) = h_\infty(x(t), 0)$ has the origin in $\mathbf{R}^{d_1}$ as its unique globally a.s.e.

A5 The functions $g_c(y) \stackrel{\text{def}}{=} \frac{g(c\lambda_\infty(y),cy)}{c}$, $c \geq 1$, satisfy $g_c \to g_\infty$ as $c \to \infty$, uniformly on compacts for some g_∞. Also, the *limiting* ODE $\dot{y}(t) = g_\infty(y(t))$ has the origin in $\mathbf{R}^{d_2}$ as its unique globally asymptotically stable equilibrium.

While convergence of two timescale iterates have been shown in prior work (Borkar, 1997, 2008; Tadić, 2004), they nevertheless assume boundedness of the iterates (in this paper, however, we provide conditions that imply boundedness). Further, conditions (A4) and (A5) are *weaker* (see Section 2 or Chapter 3 of Borkar, 2008) than (A4), (A5), Chapter 6, Borkar (2008) or A3 and A4 of Tadić (2004). We present our arguments in two parts, one part each for the two timescales (Section 5.1 for the faster timescale and Section 5.2 for the slower timescale). Assumptions A1, A2, A3 are the same as those required for the convergence analysis of two timescale SA schemes (Chapter 6, Borkar, 2008). We require A1, A2, A3 to reuse the results from Chapter 6, Borkar (2008) and show that the scaled iterates are convergent (both in the faster as well as the slower timescale). A4 and A5 summarize the conditions on the scaled and limiting ODE in the faster and the slower timescales respectively. While the scaled ODE and the limiting ODE in the case of single timescale SA schemes are homogeneous ODEs (see (A5), Chapter 3, Borkar, 2008), the scaled ODE and the limiting ODE in A4 have an external input. For our arguments in Section 5.1, we would require the results on the scaled and the limiting ODE in A4. To this end, we extend the results known (Lemmas 1, 2 and Corollary 3, Chapter 3 of Borkar, 2008) for homogeneous ODEs to ODEs with external input in Section 4. However, the scaled and the limiting ODEs for the slower timescale are homogeneous ODEs and hence we reuse the results in Chapter 3, Borkar (2008).

4. Results on ODE with external input

The ODE method for analysis of single timescale SA schemes deals with homogeneous ODEs such as $\dot{z}(t) = f(z(t))$. However, in the case of two timescale iterates we need to study coupled ODEs i.e., those with external input such as in A4. In this section, we extend the results known for homogeneous ODEs in Chapter 3, Borkar (2008), to ODEs with external inputs. To this end, we define the following:

D1 Let $\eta^{y(t)}(t, x)$ denote the solution to the ODE $\dot{x}(t) = h(x(t), y(t))$, $t \geq 0$, with initial condition $x \in \mathbf{R}^{d_1}$ and the external input $y(t) \in \mathbf{R}^{d_2}$. The superscript $y(t)$ in $\eta^{y(t)}(t, x)$ is to be understood as a symbol that indicates the fact that the external input is $y(t)$.

D2 Let $B(x_0, r) = \{x : \|x - x_0\| \leq r\}$ denote the closed ball of radius r around x_0.

D3 Let $\eta_c^{y(t)}(t, x)$ and $\eta_\infty^{y(t)}(t, x)$ be solutions to the ODEs

$$\dot{x}(t) = h_c(x(t), y(t)), \tag{3}$$

$$\dot{x}(t) = h_\infty(x(t), y(t)), \tag{4}$$

respectively, with $x(0) = x$.

In the case when $y(t) = y, \forall t \geq 0$ for some $y \in \mathbf{R}^{d_2}$ we make use of the notations $\eta^y(t, x)$, $\eta_c^y(t, x)$ and $\eta_\infty^y(t, x)$ respectively. The lemma below shows that the trajectory of the ODE tends towards and stays inside a small neighborhood of its asymptotic stable equilibrium within a given time period.

Lemma 2. *Let $K \subset \mathbf{R}^{d_1}$ be a compact set and $y \in \mathbf{R}^{d_2}$ be a fixed external input. If the limiting ODE in A4 has a unique globally asymptotically stable equilibrium (a.s.e) $\lambda_\infty(y) \in \mathbf{R}^{d_1}$, then given any $\delta > 0$, there exists a $T_\delta > 0$ such that for all initial conditions $x \in K$, we have $\eta_\infty^y(t, x) \in B(\lambda_\infty(y), \delta), \forall t \geq T_\delta$.*

Proof. See Lemma 1, Chapter 3, Borkar (2008).

The following lemma shows that for large values of c, the trajectories of (3) and (4) are close enough.

Lemma 3. *Let $x \in \mathbf{R}^{d_1}$, $y \in \mathbf{R}^{d_2}$, $[0, T]$ be a given time interval and $r > 0$ be a small positive constant. Let $y'(t) \in B(y, r)$, $\forall t \in [0, T]$, then*

$$\|\eta_c^{y'(t)}(t, x) - \eta_\infty^y(t, x)\| \leq (\epsilon(c) + Lr)Te^{LT}, \quad \forall t \in [0, T],$$

where $\epsilon(c)$ is independent of x and y and $\epsilon(c) \to 0$, as $c \to \infty$.

Proof. See Lemma 2, Chapter 3 of Borkar (2008).

The following lemma uses the results of Lemmas 2 and 3 to show that for a fixed $y \in \mathbf{R}^{d_2}$ any trajectory of (3) starting within a δ-neighborhood of $\lambda_\infty(y)$ stays within an ϵ-neighborhood of $\lambda_\infty(y)$.

Lemma 4. *Let $y \in \mathbf{R}^{d_2}$, then given any $\epsilon > 0$ and $T > 0$, there exist $c_{\epsilon,T} > 0$, $\delta_{\epsilon,T} > 0$ and $r_{\epsilon,T} > 0$ such that $\forall t \in [0, T)$, $\forall x \in B(\lambda_\infty(y), \delta_{\epsilon,T})$, $\forall c > c_{\epsilon,T}$ and external input $y'(s)$ with $y'(s) \in B(y, r_{\epsilon,T}), s \in [0, T]$, we have $\eta_c^{y'(t)}(t, x) \in B(\lambda_\infty(y), 2\epsilon), \forall t \in [0, T]$.*

Proof. Given any $\epsilon > 0$, it follows from Lyapunov stability that there exists $\delta_{\epsilon,T} > 0$ such that any trajectory of (4) starting in $B(\lambda_\infty(y), \delta_{\epsilon,T})$ stays within the ball $B(\lambda_\infty(y), \epsilon)$. Without loss of generality we can assume $\delta_{\epsilon,T} < \epsilon$. From Lemma 3 we know that there exist $c_{\epsilon,T}$ and $r_{\epsilon,T}$ such that $\|\eta_c^{y'(t)}(t, x) - \eta_\infty^y(t, x)\| \leq \epsilon$. Hence, $\forall t \in [0, T]$, $\forall y'(t) \in B(y, r_{\epsilon,T})$, $\forall x \in B(\lambda_\infty(y), \delta_{\epsilon,T})$ and $\forall c > c_{\epsilon,T}$,

$$\begin{aligned}
&\|\eta_c^{y'(t)}(t, x) - \lambda_\infty(y)\| \\
&\quad \leq \|\eta_c^{y'(t)}(t, x) - \eta_\infty^y(t, x)\| + \|\eta_\infty^y(t, x) - \lambda_\infty(y)\| \\
&\quad \leq \delta/2 + \epsilon \\
&\quad \leq 2\epsilon.
\end{aligned}$$

The following lemma shows that for a fixed $y \in \mathbf{R}^{d_2}$ and large enough c the trajectory of (3) falls within a neighborhood of $\lambda_\infty(y)$ within a given time period and continues to stay in that neighborhood.

Lemma 5. *Let $x \in B(0, 1) \subset \mathbf{R}^{d_1}$, $y \in K' \subset \mathbf{R}^{d_2}$ and let $\lambda_\infty(y)$ be the unique globally asymptotically stable equilibrium of (4). Then, given $\epsilon > 0$, there exist $c_\epsilon \geq 1$, $r_\epsilon > 0$ and $T_\epsilon > 0$ such that for any external input $y'(s)$ satisfying*

$$y'(s) \in B(y, r_\epsilon), \quad \forall s \in [0, T], \tag{5}$$

we have $\|\eta_c^{y'(t)}(t, x) - \lambda_\infty(y)\| \leq 2\epsilon, \forall t \geq T_\epsilon, \forall c > c_\epsilon$.

Proof. Given any $\epsilon > 0$, it follows from Lyapunov stability that there exists $\delta > 0$ such that any trajectory of (4) starting in $B(\lambda_\infty(y), \delta)$ stays within the ball $B(\lambda_\infty(y), \epsilon)$. Without loss of generality we can assume $\delta < \epsilon$. Let $K = B(0, 1) \cup B(\lambda_\infty(y), \delta)$, then from Lemma 2, there exists a $T_{\delta/2} > 0$ such that $\eta_\infty^y(t, x) \in B(\lambda_\infty(y), \delta/2), \forall t \geq T_{\delta/2}$. Pick $T_\epsilon \triangleq T_{\delta/2}$ as given by Lemma 2 and divide the time-line into intervals T_ϵ apart, i.e., $t \in \cup_n [nT_\epsilon, (n + 1)T_\epsilon)$, $n \geq 0$. We know $\eta_\infty^y(t, x) \in B(\lambda_\infty(y), \delta/2), \forall t \geq T_\epsilon$. From Lemma 3, it follows that there exist $c^1_{\epsilon,T_\epsilon}$ and $r^1_{\epsilon,T_\epsilon}$ such

that $\|\eta_c^{y'(t)}(T_\epsilon, x) - \eta_\infty^{y}(T_\epsilon, x)\| \leq \delta/2, \forall c > c^1_{\epsilon,T_\epsilon}$ and $y'(t)$ satisfying (5). This implies that $\eta_c^{y'(t)}(T_\epsilon, x) \in B(\lambda_\infty(y), \delta) \subset B(\lambda_\infty(y), 2\epsilon), \forall c > c^1_{\epsilon,T_\epsilon}$ and with $y'(t)$ satisfying (5). Thus starting from x, the trajectory $\eta_c^{y'(t)}(t, x)$ falls into the ball $B(\lambda(\infty), \delta)$ for all $c > c^1_{\epsilon,T_\epsilon}$ and $y'(t)$ satisfying (5).

It is easy to note that the trajectory $\eta_c^{y'(t)}(t, x)$ of the ODE (3) in the time interval $[T_\epsilon, 2T_\epsilon)$ starting from x is the same as the trajectory of the ODE (3) in the time interval $[0, T_\epsilon]$ but starting from the initial condition $\eta_c^{y'(T_\epsilon)}(T_\epsilon, x)$. Now we know from Lemma 3 that $\eta_c^{y'(t)}(t, \eta_c^{y'(t)}(T_\epsilon, x))$ and $\eta_\infty^{y}(t, \eta_c^{y'(t)}(T_\epsilon, x))$ track each other closely in the time $t \in [0, T_\epsilon)$. Formally, $\forall t \in [T_\epsilon, 2T_\epsilon)$,

$$\begin{aligned}&\|\eta_c^{y'(t)}(t, x) - \eta_\infty^{y}(t - T_\epsilon, \eta_c^{y'(t)}(T_\epsilon, x))\| \\ &= \|\eta_c^{y'(t)}(t - T_\epsilon, \eta_c^{y'(t)}(T_\epsilon, x)) - \eta_\infty^{y}(t - T_\epsilon, \eta_c^{y'(t)}(T_\epsilon, x))\| \\ &\leq \delta/2. \end{aligned} \tag{6}$$

Since $\eta_\infty^{y}(T_\epsilon, \eta_c^{y'(t)}(T_\epsilon, x)) \in B(\lambda_\infty(y), \delta/2)$ and from (6) we can conclude that $\eta_c^{y'(t)}(2T_\epsilon, x) \in B(\lambda_\infty(y), \delta)$. Also, since $\eta_c^{y'(t)}(T_\epsilon, x) \in B(\lambda_\infty(y), \delta)$, we know from Lemma 4 that there exist $c^2_{\epsilon,T_\epsilon}$ and $r^2_{\epsilon,T_\epsilon}$ such that $\eta_c^{y'(t)}(t, x) \in B(\lambda_\infty(y), 2\epsilon), \forall t \in [T_\epsilon, 2T_\epsilon)$.

Notice that by arguing in a similar manner one can show that $\eta_c^{y'(t)}(nT_\epsilon, x) \in B(\lambda_\infty(y), \delta), \forall n \geq 0$ and $\eta_c^{y'(t)}(t, x) \in B(\lambda_\infty(y), 2\epsilon), \forall t \in [nT_\epsilon, (n+1)T_\epsilon)$. The proof is complete by choosing $c_\epsilon = \max(c^1_{\epsilon,T_\epsilon}, c^2_{\epsilon,T_\epsilon})$ and $r_\epsilon = \min(r^1_{\epsilon,T_\epsilon}, r^2_{\epsilon,T_\epsilon})$.

5. Main results

The idea of rescaling will be used twice, first for the faster timescale and then for the slower timescale recursions.

5.1. Faster timescale analysis

D4 Define the timescale according to $t(n) \stackrel{\text{def}}{=} \sum_{i=0}^{n-1} a(i)$. Let $z = (x, y)$ and $\bar{z}(t) = z_n + (z_{n+1} - z_n)\frac{t - t(n)}{t(n+1) - t(n)}, t \in [t(n), t(n+1)]$.

D5 Given a timescale $t(n), n \geq 0$ and a positive constant $T > 0$, define sampling instants $T_n, n \geq 0$ as $T_0 = 0$ and $T_n = \min\{t(m) : t(m) \geq T_{n-1} + T\}$. Note that $T_n = t(m(n))$ for some $m(n) \uparrow \infty$ as $n \uparrow \infty$.

D6 Define the scaling sequence $r(n) \stackrel{\text{def}}{=} \max(r(n-1), \|\bar{z}(T_n)\|, 1)$.

D7 The scaled iterates for $m(n) \leq k \leq m(n+1) - 1$ are given by $\hat{x}_{m(n)} = \frac{x_{m(n)}}{r(n)}, \hat{y}_{m(n)} = \frac{y_{m(n)}}{r(n)}$,

$$\begin{aligned}\hat{x}_{k+1} &= \hat{x}_k + a(k)[h_c(\hat{x}_k, \hat{y}_k) + \hat{M}^{(1)}_{k+1}], \\ \hat{y}_{k+1} &= \hat{y}_k + a(k)[\epsilon_k + \hat{M}^{(2)}_{k+1}], \end{aligned} \tag{7}$$

where $c = r(n), \epsilon_k = \frac{b(k)}{a(k)}\left(\frac{g(c\hat{x}_k, c\hat{y}_k)}{c}\right), \hat{M}^{(1)}_{k+1} = \frac{M^{(1)}_{k+1}}{r(n)}, \hat{M}^{(2)}_{k+1} = \frac{M^{(2)}_{k+1}}{r(n)}$. Here h_c is as defined in A4.

D8 $\hat{z}(t) = \hat{z}_n + (\hat{z}_{n+1} - \hat{z}_n)\frac{t - t(n)}{t(n+1) - t(n)}, t \in [t(n), t(n+1)]$.

D9 Let $z_n(t) = (x_n(t), y_n(t)), t \in [T_n, T_{n+1}]$, denote the trajectory of the ODEs $\dot{x}(t) = h_{r(n)}(x(t), y(t)), \dot{y}(t) = 0$, with initial conditions $x_n(T_n) = \hat{x}(T_n)$ and $y_n(T_n) = \hat{y}(T_n)$ respectively.

Here T_ns are instants that are roughly $T > 0$ apart i.e., $T_{n+1} - T_n \approx T$. These time instants in the faster timescale are used to monitor the growth of the iterates. Also, note that the scaled iterates in D6 are bounded (by the very nature of their definition). The following lemma (Lemma 6) uses the convergence results for two timescale SA iterates (Borkar, 1997, 2008) to show the convergence of the scaled iterates in D6.

Lemma 6. *Under* A1–A3, *we have*

(i) *The sequence* $\hat{\zeta}^{(1)}_n = \sum_{m=0}^{n-1} a(m)\hat{M}^{(1)}_{m+1}, n \geq 1$ *is a.s. convergent.*
(ii) *For* $0 < k \leq m(n+1) - m(n)$, $\|\hat{z}(t(m(n)+k))\| \leq K_2$, *a.s. for some* $K_2 > 0$.
(iii) $\lim_{n\to\infty} \|\hat{z}(t) - z_n(t)\| = 0$, *a.s.*, $\forall t \in [T_n, T_{n+1})$.

Proof. See Chapter 3 and Chapter 6, Borkar (2008).

Lemma 6(i) shows that the martingale difference sequence $\hat{M}^{(1)}_{k+1}$ participating in the scaled recursion in (7) is convergent. Lemma 6(ii) shows that in the faster timescale, between instants T_n and T_{n+1}, starting from within a unit ball around origin, the norm of the iterates can grow only by a factor $K_2 > 0$. Lemma 6(iii) shows that scaled iterates track a corresponding scaled ODE.

Theorem 7. *Under* A1–A4, *we have the following:*

(i) *For* n *large and* $T \stackrel{\text{def}}{=} T_{1/4}$ *(where* T *is the sampling period used in* D5 *and* $T_{1/4}$ *is* T_ϵ *as in Lemma 5 with* $\epsilon = 1/4$*), if* $\|\bar{x}(T_n)\| > C_1(1 + \|\bar{y}(T_n)\|)$, *then* $\|\bar{x}(T_{n+1})\| < \frac{3}{4}\|\bar{x}(T_n)\|$.
(ii) $\|\bar{x}(T_n)\| \leq C^*(1 + \|\bar{y}(T_n)\|)$ *almost surely, for some* $C^* > 0$.
(iii) $\|x_n\| \leq K^*(1 + \|y_n\|)$ *almost surely for some* $K^* > 0$.

Proof. (i) $\|\bar{x}(T_n)\| > C_1(1 + \|\bar{y}(T_n)\|)$ implies that $r(n) \geq C_1$, $\|\hat{y}(T_n)\| < \frac{1}{C_1}$ and $\|\hat{x}(T_n)\| > \frac{1}{1+1/C_1}$. Let $\eta_\infty^{y(t)}(t, x)$ and $\eta_c^{y(t)}(t, x)$ be the solutions to the ODEs $\dot{x}(t) = h_\infty(x(t), y(t))$ and $\dot{x}(t) = h_c(x(t), y(t))$, respectively, with initial condition x in both. Let $y'(t - T_n) = y_n(t), \forall t \in [T_n, T_{n+1}]$. It is easy to see that $x_n(t) = \eta^{y'(t)}(t - T_n, \hat{x}(T_n)), \forall t \in [T_n, T_{n+1}]$, where $x_n(t)$ and $y_n(t)$ are as in D9.

We also know from Lemma 5 that there exist $r_{1/4}, c_{1/4}$, and $T_{1/4}$ such that $\|\eta_c^{y'(t)}(t, \hat{x}(T_n))\| \leq \frac{1}{4}, \forall t \geq T_{1/4}, \forall c \geq c_{1/4}$, whenever $y'(t) \in B(0, r_{1/4}), \forall t \in [0, T]$.

Now, let us pick $C_1 > \max(c_{1/4}, \frac{2}{r_{1/4}})$ and the positive constant $T > 0$ as $T \stackrel{\text{def}}{=} T_{1/4}$. Since $y'(t - T_n) = y_n(t) = \hat{y}(T_n), \forall t \in [T_n, T_{n+1}]$, we have for our choice of C_1, $y'(s) \in B(0, r_{1/4}), \forall s \in [0, T]$. We also know from Lemma 6(iii) that $\|\hat{x}(T^-_{n+1}) - x_n(T_{n+1})\| < \frac{1}{4}$ for sufficiently large n. Since $\|x_n(T_{n+1})\| = \|\eta^{y'(t)}(T_{n+1} - T_n, \hat{x}(T_n))\| \leq 1/4$, we have $\|\hat{x}(T^-_{n+1})\| \leq \|\hat{x}(T^-_{n+1}) - x_n(T_{n+1})\| + \|x_n(T_{n+1})\| \leq \frac{1}{2}$. Since $\frac{\bar{x}(T_{n+1})}{\bar{x}(T_n)} = \frac{\hat{x}(T^-_{n+1})}{\hat{x}(T_n)}$, it follows that $\|\bar{x}(T_{n+1})\| < \frac{1+1/C_1}{2}\|\bar{x}(T_n)\|$, and the result holds by assuming without loss of generality that $C_1 > \max(c_{1/4}, \frac{2}{r_{1/4}}) > 2$.

(ii) On a set of positive probability, let us assume on the contrary that there exists a monotonically increasing sequence $\{n_k\}$ for which $C_{n_k} \uparrow \infty$ as $k \to \infty$ and $\|\bar{x}(T_{n_k})\| \geq C_{n_k}(1 + \|\bar{y}(T_{n_k})\|)$. Now from Theorem 7(i), we know that if $\|\bar{x}(T_n)\| > C_1(1 + \|\bar{y}(T_n)\|)$, then $\|\bar{x}(T_k)\|$ for $k \geq n$ falls at an exponential rate until it is within the ball of radius $C_1(1 + \|\bar{y}(T_k)\|)$. Thus corresponding to the sequence $\{n_k\}$, there must exist another sequence $\{n'_k\}$ such that $n_{k-1} \leq n'_k \leq n_k$ and $\|\bar{x}(T_{n'_k - 1})\|$ is within the ball of radius $C_1(1 + \|\bar{y}(T_{n'_k - 1})\|)$ but $\|\bar{x}(T_{n'_k})\|$ is greater than $C_{n_k}(1 + \|\bar{y}(T_{n'_k})\|)$. However, from Lemma 6(ii) we know that the iterates can grow only by a factor of K_2 between $m(n'_k - 1)$ and $m(n'_k)$. This leads to a contradiction. So we conclude that $\|\bar{x}(T_n)\| \leq C^*(1 + \|\bar{y}(T_n)\|)$ for some $C^* > 0$.

(iii) We showed that $\|\bar{x}(T_n)\| \leq C^*(1 + \|\bar{y}(T_n)\|)$. From Lemma 6(ii), we know that $\|\bar{z}(t)\| \leq K_2\|\bar{z}(T_n)\|, \forall t \in [T_n, T_{n+1})$. The result follows by choosing $K^* = K_2C^*$.

Remark. In the above proof of Theorem 7, (i) follows from the fact that (see Lemma 6(iii)) the scaled iterates in (7) track the scaled ODE in D9, which in turn falls within an ϵ ball around the stable equilibrium (see Lemma 5) of the limiting ODE in A4. (ii) follows from (i) and Lemma 6(ii). (iii) follows from (ii) and Lemma 6(ii). We also observe that the stability result (Theorem 7, Chapter 3, Borkar, 2008) for single timescale SA schemes can be recovered by letting $y_n = 0, \forall n \geq 0$ in Theorem 7.

D8 Given any $\epsilon > 0$ and $y \in \mathbf{R}^{d_2}$, define the set $A^\epsilon(y) \subset \mathbf{R}^{d_1}$ as $A^\epsilon(y) \stackrel{\text{def}}{=} \{x\colon \|x - \lambda_\infty(y)\| < \epsilon\}$.

Theorem 8. *For any given $\epsilon > 0$, there exists $c_\epsilon > 0$ such that, if $r(n) > c_\epsilon$ for some n, then it follows that $(\hat{x}_k, \hat{y}_k) \in (A^\epsilon(\hat{y}_k), \hat{y}_k), \forall k \geq m(n)$.*

Proof. The proof follows by a repeated application of Lemmas 6 and 5 to the intervals $[T_w, T_{w+1}], \forall k \geq n$ and using the fact that $r(w) \geq r(n), \forall w \geq n$.

Theorem 8 shows that for large n and large scaling factor $c \gg 1$, the faster timescale iterates are within an ϵ neighborhood of the stable equilibrium of the limiting ODE in A4.

5.2. Slower timescale analysis

The following arguments are with respect to the slower timescale. Thus, the notions of timescale, the corresponding sampling instants T_n and the scaled iterates are re-defined with respect to the slower timescale as below:

D10 Define now the timescale to be $t(n) \stackrel{\text{def}}{=} \sum_{i=0}^{n-1} b(i)$. Let $z = (x, y)$ and $\bar{z}(t) = z_n + (z_{n+1} - z_n)\frac{t - t(n)}{t(n+1) - t(n)}$, $t \in [t(n), t(n+1)]$.

D11 Given a timescale $t(n), n \geq 0$ and a positive constant $T > 0$, define sampling instants $T_n, n \geq 0$ as $T_0 = 0$ and $T_n = \min\{t(m) : t(m) \geq T_{n-1} + T\}$. Note that $T_n = t(m(n))$ for some $m(n) \uparrow \infty$ as $n \uparrow \infty$.

D12 Define scaling sequence $r(n) \stackrel{\text{def}}{=} \max(r(n-1), \|\bar{z}(T_n)\|, 1)$.

D13 The scaled iterates for $m(n) \leq k \leq m(n+1) - 1$ are given by $\hat{x}_{m(n)} = \frac{x_{m(n)}}{r(n)}, \hat{y}_{m(n)} = \frac{y_{m(n)}}{r(n)}$,

$$\begin{aligned} \hat{x}_{k+1} &= \hat{x}_k + a(k)[h_c(\hat{x}_k, \hat{y}_k) + \hat{M}^{(1)}_{k+1}], \\ \hat{y}_{k+1} &= \hat{y}_k + b(k)[g_c(\hat{x}_k, \hat{y}_k) + \hat{M}^{(2)}_{k+1}], \end{aligned} \tag{8}$$

where $c = r(n)$, $\hat{M}^{(1)}_{k+1} = \frac{M^{(1)}_{k+1}}{r(n)}$, $\hat{M}^{(2)}_{k+1} = \frac{M^{(2)}_{k+1}}{r(n)}$. Here h_c and g_c are as defined in A4 and A5 respectively.

D14 $\hat{z}(t) = \hat{z}_n + (\hat{z}_{n+1} - \hat{z}_n)\frac{t - t(n)}{t(n+1) - t(n)}$, $t \in [t(n), t(n+1)]$.

D15 We let $y_n(t), t \in [T_n, T_{n+1}]$, denote the trajectory of the ODEs $\dot{y}(t) = g_c(y(t))$ with initial conditions $y_n(T_n) = \hat{y}(T_n)$.

Lemma 9. *Under assumptions* A1–A3, *we have the following:*

(i) *The sequence $\hat{\zeta}^{(2)}_n = \sum_{m=0}^{n-1} a(m)\hat{M}^{(2)}_{m+1}$, $n \geq 1$ is a.s. convergent.*
(ii) *For $0 < k \leq m(n+1) - m(n)$, we have a.s. $\|\hat{z}(t(m(n)+k))\| \leq K_3$, for some $K_3 > 0$.*
(iii) *For sufficiently large n we have $\sup_{t \in [T_n, T_{n+1})} \|\hat{y}(t) - y_n(t)\| \leq \epsilon(c) L T e^{L(L+1)T}$, a.s, where $\epsilon(c) \to 0$ as $c \to \infty$.*

Proof. See Chapter 3 and Chapter 6, Borkar (2008).

Lemma 9 is similar to Lemma 6 and uses the result of Theorem 7 i.e., in the slower timescale the scaled iterates in (8) obey $\|\hat{x}_k\| \leq K^*(1 + \|\hat{y}_k\|)$, $\forall k \geq 0$. Let $g_c\colon \mathbf{R}^{d_2} \to \mathbf{R}^{d_2}$ and $g_\infty\colon \mathbf{R}^{d_2} \to \mathbf{R}^{d_2}$ be functions as defined in A5, and let $\chi_\infty(t, y)$ denote the solution to the ODE $\dot{y}(t) = g_\infty(y(t))$, with initial condition y, and let $\chi_c(t, y)$ denote the solution to the ODE $\dot{y}(t) = g_c(y(t))$, with initial condition y.

Theorem 10. *Under* A1–A5, *we have the following:*

(i) *Let K^* be as in Theorem 7*(iii), *then it follows that $\|\hat{y}(T_n)\| \geq \frac{1}{K^*+2}$ for sufficiently large $\|\bar{y}(T_n)\|$.*
(ii) *For n large, $T \stackrel{\text{def}}{=} T_{1/4}$ (with T as in* D11 *and $T_{1/4}$ is T_ϵ as in Lemma 5 with $\epsilon = 1/4$), if $\|\bar{y}(T_n)\| > C$, then $\|\bar{y}(T_{n+1})\| < \frac{1}{2}\|\bar{y}(T_n)\|$.*
(iii) *$\|\bar{y}(T_n)\| \leq C'$ a.s., for some $C' > 0$.*
(iv) *$\sup_n \|y_n\| < \infty$, a.s. and $\sup_n \|x_n\| < \infty$, a.s.*

Proof. (i) From Theorem 7(iii), we know that $\|r(n)\| \leq \|\bar{y}(T_n)\| + K^*(1 + \|\bar{y}(T_n)\|)$. Also, $\|\hat{y}(T_n)\| = \frac{\|\bar{y}(T_n)\|}{r(n)} \geq \frac{\|\bar{y}(T_n)\|}{\|\bar{y}(T_n)\| + K^*(1 + \|\bar{y}(T_n)\|)} = \frac{1}{1 + \frac{K^*}{\|\bar{y}(T_n)\|} + K^*}$. The claim follows for any $\|\bar{y}(T_n)\| > K^*$.

(ii) By A5 we have $\mathbf{0} \in \mathbf{R}^{d_2}$ is the unique globally asymptotically stable equilibrium of the ODE $\dot{y}(t) = g_\infty(y(t))$, and as a consequence of Lemma 5 there exist $c_{1/4}$ and $T_{1/4}$ such that $\|\chi_c(t, y)\| < \frac{1}{4(K^*+2)}$, $\forall t \geq T_{1/4}, c > c_{1/4}$.

Also, if $\|\bar{y}(T_n)\| > K^*$, it follows from (i) above that $\|\hat{y}(T_n)\| \geq \frac{1}{K^*+2}$. We know from Lemma 9(iii) that for sufficiently large n, there exists $C_1 > 0$ such that $\|\hat{y}(T^-_{n+1}) - y_n(T_{n+1})\| < \frac{1}{4(K^*+2)}$, for $r(n) > C_1$. Now, let us pick $C = \max(c_{1/4}, C_1, K^*)$ and $T = T_{1/4}$. Then for sufficiently large n it follows that $\|\hat{y}(T^-_{n+1})\| \leq \|\hat{y}(T^-_{n+1}) - y_n(T_{n+1})\| + \|y_n(T_{n+1})\| \leq \frac{1}{2(K^*+2)}$. Since $\frac{\bar{y}(T_{n+1})}{\bar{y}(T_n)} = \frac{\hat{y}(T^-_{n+1})}{\hat{y}(T_n)}$, it follows that $\|\bar{y}(T_{n+1})\| < \frac{1}{2}\|\bar{y}(T_n)\|$.

(iii) Let us assume on the contrary that on a set of positive probability, there exists a sequence $\{n_k\}$ such that $C_{n_k} \uparrow \infty$ as $k \to \infty$ and $\|\bar{y}(T_{n_k})\| \geq C_{n_k}$. From Theorem 10(ii) we know that if $\|\bar{y}(T_n)\| > C$, then $\|\bar{y}(T_k)\|$ for $k \geq n$ falls at an exponential rate to the ball of radius C. Thus corresponding to the sequence $\{n_k\}$, there exists another sequence $\{n'_k\}$ such that $n_{k-1} \leq n'_k \leq n_k$ and $\|\bar{y}(T_{n'_k - 1})\|$ is within the ball of radius C but jumps outside this ball of radius C (i.e., $\|\bar{y}(T_{n'_k})\| > C$) to points which are at a distance greater than C_{n_k} from the origin. However, from Lemma 9(ii) we know that the iterates can grow only by a factor of K_3 between $m(n'_k - 1)$ and $m(n'_k)$. This leads to a contradiction and the claim follows.

(iv) We know that $\|\bar{y}(T_n)\| \leq C'$. From Lemma 9(ii), we know that $\|\bar{y}(t)\| \leq K_3\|\bar{y}(T_n)\|$, $\forall t \in [T_n, T_{n+1})$. The result follows by noting that $\|y_n\| \leq K_3 C'$ almost surely. The fact that $\sup_n \|x_n\| < \infty$ follows from Theorem 7.

Remark. The proof above for Theorem 10 is along the lines of the proof of Theorem 7. Here, (ii) follows from the fact that the scaled iterates in (8) track the scaled ODE in D15, which in turn follows the limiting ODE in A5. Now, since the limiting ODE is stable to the origin, the iterates fall at an exponential rate into the unit ball around the origin. Together with the fact that the iterates can grow only by a constant factor (Lemma 9(ii)) within a given time period (iii) follows. Then, (iv) follows from (iii), Lemma 9(ii) and Theorem 7.

To summarize, we first showed in Section 5.1 that the faster timescale iterates are bounded by the growth of slower timescale iterates, and in Section 5.2 we showed that the slower timescale iterates themselves are bounded, thus completing our stability analysis of two timescale stochastic approximation schemes.

6. An application in reinforcement learning

Reinforcement Learning (RL) algorithms are sample trajectory based methods for solving Markov Decision Processes (MDP). In the example presented in this paper, we consider an MDP whose state space is denoted by $S = \{s_1, s_2, \ldots, s_n\}$, and the action space by $A = \{a_1, a_2, \ldots, a_m\}$. In state s and under action a, let $c_a(s)$ denote the single-stage cost incurred and $p_a(s, s')$ be the probability of transition to s'. By a policy, we mean a sequence $\pi = \{\pi_1, \pi_2, \ldots, \pi_k, \ldots\}$, where each π_k, $k = 1, 2, \ldots$, specifies a way by which states are mapped to actions. In this paper,

we consider the class of time stationary policies with the *Gibbs* parameterization given as under.

$$\pi^\theta(s,a) = \frac{e^{\phi_{sa}^\top \theta}}{\sum_{a'} e^{\phi_{sa'}^\top \theta}}, \tag{9}$$

where $\phi_{sa} \in \mathbf{R}^{d_2}$ is a d_2-dimensional feature vector for the state–action tuple (s,a), and $\theta \in \mathbf{R}^{d_2}$ is the parameter. We assume that under a policy π^θ the MDP is an irreducible and aperiodic Markov chain with probability transition matrix given by P_{π^θ}. The *average*-cost associated with a policy π^θ is given by $J(\theta) = \lim_{N\to\infty} \frac{1}{N}\mathbf{E}[\sum_{n=0}^{N-1} c_{a_n}(s_n)|\pi^\theta]$, where a_n is sampled from the distribution $\pi(s_n,\cdot)$, $\forall n \geq 0$ and given s_n, s_{n+1} is distributed as $p_{a_n}(s_n,\cdot)$. Other important quantities of interest are the differential state–action cost given by $Q^\theta(s) = \mathbf{E}[\sum_{n=0}^{\infty}(c_{a_n}(s_n) - J(\theta))|\pi^\theta]$, the differential cost $V^\theta(s) = \sum_{a\in A}\pi^\theta(s,a)Q^\theta(s)$ and the advantage function $A^\theta(s,a) = Q^\theta(s,a) - V^\theta(s)$. The average and differential costs are connected by the Bellman equation

$$Q^\theta(s,a) = \left[c_a(s) - J(\pi) + \sum_a p_a(s,s')V^\theta(s')\right]. \tag{10}$$

We also assume that under any given policy π^θ the MDP is an aperiodic irreducible Markov chain with unique stationary distribution d^π. In what follows, since π^θ is fixed for a fixed θ, we use π^θ, θ and π interchangeably to denote π^θ.

Actor-Critic RL algorithms are a sub-class of policy gradient RL algorithms, where the aim is to find the best policy amongst the class of parameterized policies (such as the Gibbs parameterization in (9)). AC algorithms are two timescale SA schemes and have two components namely, the actor which is responsible for control and updates the policy parameters (i.e. θ in (9)) in the slower timescale, and the critic which is responsible for prediction and evaluates the policy in the faster timescale. We now verify our stability conditions for the iterates in Algorithm 1 of Bhatnagar et al. (2009), which is an actor-critic algorithm (repeated here as Algorithm 1). However, there are certain minor differences between Algorithm 1 of Bhatnagar et al. (2009) and the Algorithm 1 presented here. Algorithm 1 of Bhatnagar et al. (2009) corresponds to the average reward setting while the algorithm here is for the setting of average cost. Also, the update in equation (23) of Algorithm 1 of Bhatnagar et al. (2009) makes use of a projection operator Γ in order to ensure boundedness of the iterates. However, in the actor update in line 5 of Algorithm 1 (here), we do not make use of projection and instead we introduce an additional term $\epsilon\theta_n$ (where $\epsilon > 0$ is a small positive constant). This is equivalent to adding a quadratic penalty term $\epsilon\theta^2$ to the objective $J(\theta)$. We now write the iterates

Algorithm 1 The Actor-Critic Algorithm

1: **for** $n = 0, 1, 2, \ldots$ **do**
2: Average Cost Update: $\hat{J}_{n+1} = \hat{J}_n + a(n)(c_{a_n}(s_n) - \hat{J}_n)$
3: TD Error: $\delta_n = c_{a_n}(s_n) - \hat{J}_n + v_n^\top \gamma_{s_{n+1}} - v_n^\top \gamma_{s_n}$
4: Critic Update: $v_{n+1} = v_n + a(n)\delta_n\gamma_{s_n}$
5: Actor Update: $\theta_{n+1} = \theta_n - b(n)(\delta_n\psi_{s_na_n} + \epsilon\theta_n)$
6: **end for**

in the standard form as below:

$$\hat{J}_{n+1} = \hat{J}_n + a(n)[h^1(\hat{J}_n, v_n, \theta_n) + M_{n+1}^{(1)}], \tag{11}$$

$$v_{n+1} = v_n + a(n)[h^2(\hat{J}_n, v_n, \theta_n) + M_{n+1}^{(2)}], \tag{12}$$

$$\theta_{n+1} = \theta_n + b(n)[g(\hat{J}_n, v_n, \theta_n) + M_{n+1}^{(3)}], \tag{13}$$

where

- $h^1(\hat{J}_n, v_n, \theta_n) = \mathbf{E}[c_{a_n}(s_n)|\mathcal{F}_n] - \hat{J}_n = \sum_{s,a} d^\pi(s)\pi(s,a)\big(c_a(s) - \hat{J}_n\big)$,
- $h^2(\hat{J}_n, v_n, \theta_n) = \mathbf{E}[\delta_n\gamma_{s_n}|\mathcal{F}_n] = \sum_{s,a} d^\pi(s)\pi(s,a)\big(c_a(s) - \hat{J}_n + \sum_{s'} p_a(s,s')v_n^\top\gamma_{s'} - v_n^\top\gamma_s\big)\gamma_s$,
- $g(\hat{J}_n, v_n, \theta_n) = \mathbf{E}[\delta_n\psi_{s_na_n}|\mathcal{F}_n] = -\sum_s d^\pi(s)\sum_a \nabla\pi(s,a)\hat{A}^\theta(s,a) - \epsilon\theta_n$.

Note that, in the above equations, π stands for π^{θ_n} and $\mathcal{F}_n \overset{\text{def}}{=} \sigma(\theta_m, v_m, \hat{J}_m, M_m^{(1)}, M_m^{(2)}, M_m^{(3)}, 0 \leq m \leq n)$. We assume for simplicity here that states s_n are sampled independently according to the stationary distribution $d^{\pi^\theta}(s)$ of the underlying Markov chain (see Chapter 6 of Bertsekas & Tsitsiklis, 1996). Further, for the feature matrix F with sth row being f_s we assume that $Fe \neq e$, where $e = (1, \ldots, 1)$. A similar assumption has also been made in Bhatnagar et al. (2009) and Tsitsiklis and Van Roy (1999). Here, the differential cost $V^\theta(s)$ is approximated as $V^\theta(s) \approx v^{\pi\top}\gamma_s$ (where $\gamma_s \in \mathbf{R}^{d_1}$ is the feature vector of state s, $v^\theta \in \mathbf{R}^k$ is a learnt weight vector) and $\hat{A}^\theta(s,a) = c_a(s) - \hat{J}_n + \sum_{s'} p_a(s,s')v_n^\top\gamma_{s'} - v_n^\top\gamma_s$ is the approximate *advantage*-function (see Bhatnagar et al., 2009). The martingale terms are given by $M_{n+1}^{(1)} = c_{a_n}(s_n) - \mathbf{E}[c_{a_n}(s_n)|\mathcal{F}_n]$, $M_{n+1}^{(2)} = \delta_n\gamma_{s_n} - \mathbf{E}[\delta_n\gamma_{s_n}|\mathcal{F}_n]$, $M_{n+1}^{(3)} = \delta_n\psi_{s_na_n} - \mathbf{E}[\delta_n\psi_{s_na_n}|\mathcal{F}_n]$. It is easy to see that there exist C_i, $i = 1, 2, 3$ such that $\mathbf{E}[\|M_{n+1}^{(i)}\|^2|\mathcal{F}_n] \leq C_i(1 + \|\theta_n\|^2 + \|v_n\|^2 + \|\hat{J}_n\|^2)$. We show that A1–A5 hold for the iterates in Algorithm 1 in the proposition below.

Proposition 11. (1) *The functions* $h^1 : \mathbf{R}^{1+d_1+d_2} \to \mathbf{R}$, $h^2 : \mathbf{R}^{1+d_1+d_2} \to \mathbf{R}^{d_1}$, $g : \mathbf{R}^{1+d_1+d_2} \to \mathbf{R}^{d_2}$ *are Lipschitz continuous.*

(2) *The functions* $h_c^1(\hat{J}, v, \theta) \overset{\text{def}}{=} \frac{h^1(c\hat{J}, cv, c\theta)}{c}$, $c \geq 1$ *are Lipschitz continuous, and satisfy* $h_c^1 \to h_\infty^1$, *as* $c \to \infty$, *uniformly on compacts.*

(3) *The functions* $h_c^2(\hat{J}, v, \theta) \overset{\text{def}}{=} \frac{h^2(c\hat{J}, cv, c\theta)}{c}$ *are Lipschitz continuous, and satisfy* $h_c^2 \to h_\infty^2$ *as* $c \to \infty$, *uniformly on compacts.*

(4) *The ODE* $(\dot{\hat{J}}(t), \dot{v}(t)) = h_\infty(\hat{J}(t), v(t), \theta)$ *has a unique globally asymptotically stable equilibrium* $\lambda_\infty(\theta)$, *where* $\lambda_\infty : \mathbf{R}^{d_2} \to \mathbf{R}^{1+d_1}$ *is a Lipschitz map that satisfies* $\lambda_\infty(\mathbf{0}) = \mathbf{0}$.

(5) *The functions* $g_c(\theta) \overset{\text{def}}{=} \frac{g(c\theta, c\lambda_\infty(\theta))}{c}$ *are Lipschitz continuous and satisfy* $g_c \to g_\infty$, *as* $c \to \infty$, *uniformly on compacts and the ODE* $\dot{\theta} = g_\infty(\theta(t))$ *has the origin in* $\mathbf{R}^{d_2}$ *as its unique globally asymptotically stable equilibrium.*

Proof. (1) Recall that $h^1(\hat{J}, v, \theta) = \sum_s d^\pi(s)\sum_a \pi(s,a)c_a(s) - \hat{J}$. Then h^1 is clearly Lipschitz (in fact linear) in $\hat{J}$. We now show that the derivatives of h^1 are bounded w.r.t. θ. $\nabla_\theta h^1(\hat{J}, v, \theta) = \sum_s \nabla_\theta d^\pi(s)\sum_a \pi(s,a)c_a(s) + \sum_s d^\pi(s)\sum_a \nabla_\theta\pi(s,a)c_a(s)$. We know from Schweitzer (1968) that $d^\pi(s)$ is continuously differentiable in θ and has a bounded derivative. Now since $d^\pi(s) = \sum_{s'} d^\theta(s')\sum_a \pi(s',a)\sum_s p_a(s',s)$, and $\nabla\pi_\theta(s,a) = \pi(s,a)(\phi_{sa} - \sum_{a'}\phi(s,a')\pi(s,a'))$, it follows that $d^\pi(s)$ also has a bounded derivative w.r.t. θ. Using similar arguments on the boundedness of the derivatives of $d^\pi(s)$ and $\pi(s,a)$ w.r.t. θ we can show that h^2 and g are Lipschitz as well.

Given θ, define quantity $\pi_c(\theta) \overset{\text{def}}{=} \pi^{c\theta}$, $c \geq 1$. For any s, let $a^* = \arg\max_a \phi_{sa}^\top\theta$, and let $\pi_\infty^\theta(s,a) \overset{\text{def}}{=} \mathbf{1}_{\{a=a^*\}}$, where $\mathbf{1}$ is the indicator function. We show that $\pi_c^\theta(s,a) \to \pi_\infty^\theta(s,a)$, uniformly on compacts. For any $a \neq a^*$, $\pi_c^\theta(s,a) = \frac{e^{c\phi_{sa}^\top\theta}}{\sum_{a'} e^{c\phi_{sa'}^\top\theta}} \leq \frac{e^{c\phi_{sa}^\top\theta}}{e^{c\phi_{sa^*}^\top\theta}} = e^{c(\phi_{sa}^\top - \phi_{sa^*}^\top)\theta}$. Now since $a \neq a^*$ and $\pi_c^\theta(s,a^*) = 1 - \sum_{a\neq a^*}\pi_c^\theta(s,a)$, we have $\pi_c^\theta(s,a) \to 0$, as $c \to \infty$, $\forall s \in S, a \in A$, uniformly on compacts. We now drop the superscript θ and simply use π_∞ and π_c for notational simplicity.

(2) $h_c^1(\hat{J}, v, \theta) = \frac{\sum_s d^{\pi_c}(s)\sum_a \pi_c(s,a)c_a(s) - c\hat{J}}{c}$, thus we have $h_\infty^1(\hat{J}, v, \theta) = -\hat{J}$.

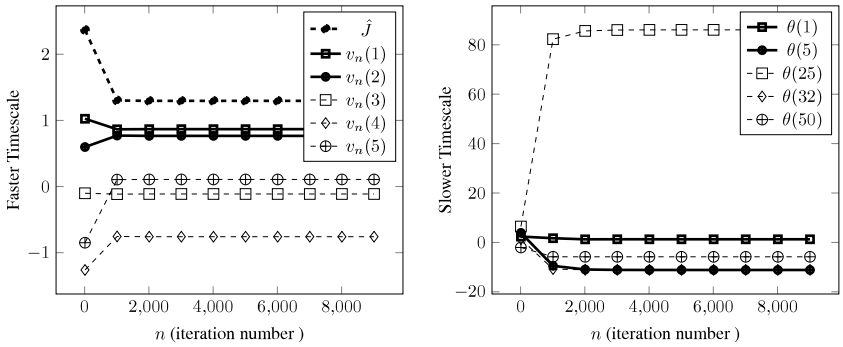

Fig. 1. Stability and convergence of faster timescale iterates (on the left) and slower timescale iterates (on the right). Algorithm 1 was run for 10,000 iterations and the plots show the iterates sampled every 1000 iterations.

(3) In a similar fashion, $h_c^2(\hat{J}, v, \theta) \stackrel{\text{def}}{=} \Big(\sum_s d^{\pi_c}(s) \sum_a \pi_c(s,a) [c_a(s) - c\hat{J} + \sum_{s'} p_a(s,s') v_n^\top f_{s'} - v_n^\top f_s]\Big)/c$, and $h_\infty^2(\hat{J}, v, \theta) = F^\top D^{\pi_\infty}(-I + P_\pi)Fv - \hat{J}$.

(4) Consider the ODE $(\dot{\hat{J}}(t), \dot{v}(t)) = h_\infty(\hat{J}, v, \theta)$, where $h_\infty = (h_\infty^1, h_\infty^2)$. It is straightforward to see that $\dot{\hat{J}}(t) = -\hat{J}(t)$ is stable to the origin in $\mathbf{R}$. Now, we know from Bhatnagar et al. (2009) (provided A3 of Bhatnagar et al., 2009 holds) that $\dot{v}(t) = F^\top D^{\pi_\infty}(-I + P_\pi)Fv$ has the origin in $\mathbf{R}^{d_1}$ as its unique asymptotically stable equilibrium. Thus, $\lambda(\theta) = (0, \mathbf{0}) \in \mathbf{R}^{1+d_1}, \forall \theta \in \mathbf{R}^{d_2}$, where $0 \in \mathbf{R}$ and $\mathbf{0} \in \mathbf{R}^{d_1}$.

(5) Now, using the fact that $\lambda_\infty(\theta) = (0, \mathbf{0})$, we have $g_c(\theta) = \Big(\sum_s d^{\pi_c}(s) \sum_a \nabla \pi_c(s,a)[c_a(s) - c \times 0 + \sum_{s'} p_a(s,s') c\, \mathbf{0}^\top f_{s'} - c\, \mathbf{0}^\top f_s]\Big)/c - \frac{\epsilon c \theta}{c}$, and $g_\infty(\theta) = \lim_{c\to\infty} \frac{g(c\theta)}{c}$. Also, it is easy to see that all the quantities in the numerator of the first term on the right hand side of $g_c(\theta)$ are bounded and therefore, $g_\infty(\theta) = -\epsilon\theta$. Then the ODE $\dot{\theta}(t) = -\theta(t)$, has the origin in $\mathbf{R}^{d_2}$ as its unique globally asymptotically stable equilibrium.

The claim follows.

We have thus shown that the iterates θ_n, v_n and $\hat{J}_n$ are stable. It is important to note that for the stability analysis we have not explicitly projected the iterates θ_n as in Bhatnagar et al. (2009).

We now present a numerical implementation of Algorithm 1. We considered an MDP with 50 states and 10 actions. For each state $s \in S$ and action $a \in A$, the transition probabilities $p_a(s, \cdot)$ were chosen at random and then normalized so that $\sum_{s' \in S} p_a(s,s') = 1$. For each state–action pair (s,a), the cost $c_a(s)$ was chosen uniformly from integers 0 to 10. The feature $\gamma_s \in \mathbf{R}^{d_2}$ for state $s \in S$, was chosen such that each co-ordinate was 0 or 1 w.p. $\frac{1}{2}$. The features ϕ_{sa_i} were then made to be $\phi_{sa_i} = (\underbrace{0, \ldots, 0}_{k\times(i-1)}, \gamma_s, \underbrace{0, \ldots, 0}_{k\times(m-i)})^\top \in \mathbf{R}^{d_2}$. We chose $d_1 = 5$ and by the manner of constructing ϕ_{sa}, we have $d_2 = d_1 \times m = 50$. The plot above (see Fig. 1) shows stability (and convergence) of all the faster timescale iterates and some of the slower timescale iterates (since θ has 50 co-ordinates we only show plots for some of the co-ordinates).

7. Conclusions

In this paper, we derived verifiable sufficient conditions that guarantee the stability of two-timescale stochastic approximation algorithms. This problem had not been addressed previously in the literature. The sufficient conditions turn out to be weaker than those needed to establish convergence (Chapter 6 of Borkar, 2008). Our stability analysis of two-timescale stochastic approximation is quite general and may easily be extended to the case of multi-timescale stochastic approximation, where the number of timescales is more than two. We also presented an application of our results to an actor-critic algorithm in reinforcement learning and also presented a numerical example using this algorithm.

Acknowledgment

Work for this paper was partially supported by the Robert Bosch Centre for Cyber-Physical Systems, Indian Institute of Science, Bangalore.

References

Bertsekas, D. P., & Tsitsiklis, J. N. (1996). *Neuro-dynamic programming* (1st ed.). Athena Scientific.
Bhatnagar, S. (2005). Adaptive multivariate three-timescale stochastic approximation algorithms for simulation based optimization. *ACM Transactions on Modeling and Computer Simulation (TOMACS), 15*(1), 74–107.
Bhatnagar, S., & Borkar, V. S. (1998). A two timescale stochastic approximation scheme for simulation-based parametric optimization. *Probability in the Engineering and Informational Sciences, 12*, 519–531.
Bhatnagar, S., Fu, M. C., Marcus, S. I., & Wang, I. (2003). Two-timescale simultaneous perturbation stochastic approximation using deterministic perturbation sequences. *ACM Transactions on Modeling and Computer Simulation, 13*, 180–209.
Bhatnagar, S., Sutton, R. S., Ghavamzadeh, M., & Lee, M. (2009). Natural actor–critic algorithms. *Automatica, 45*(11), 2471–2482.
Borkar, V. S. (1997). Stochastic approximation with two time scales. *Systems & Control Letters, 29*(5), 291–294.
Borkar, V. S. (2008). *Stochastic approximation: a dynamical systems viewpoint*. Cambridge Univ. Press.
Borkar, V. S., & Meyn, S. P. (2000). The ODE method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization, 38*(2), 447–469.
Schweitzer, Paul J. (1968). Perturbation theory and finite Markov chains. *Journal of Applied Probability, 5*(2), 401–413.
Tadić, Vladislav B. (2004). Almost sure convergence of two time-scale stochastic approximation algorithms. In *American control conference, 2004. Proceedings of the 2004, Vol. 4* (pp. 3802–3807). IEEE.
Tsitsiklis, J. N., & Van Roy, B. (1999). Average cost temporal-difference learning. *Automatica, 35*(11), 1799–1808.

Chandrashekar Lakshminarayanan received a Bachelors in Instrumentation and Control Engineering from National Institute of Technology, Tiruchirappalli in 2005, Masters in Systems Science and Automation, and Ph.D.degree in Computer Science, from the Indian Institute of Science, Bangalore in 2010 and 2016, respectively. He is a post-doctoral researcher at the Department of Computing Sciences at the University of Alberta, Canada. His research interests are in reinforcement learning, stochastic approximation algorithms, stochastic optimization and control, as well as applications such as crowd sourcing.

Shalabh Bhatnagar received a Bachelors in Physics (Hons) from the University of Delhi in 1988 and the Masters and Ph.D.degrees in Electrical Engineering from the Indian Institute of Science, Bangalore in 1992 and 1998, respectively. He is a Professor at the Department of Computer Science and Automation at the Indian Institute of Science, Bangalore. His research interests are in stochastic approximation algorithms, stochastic optimization and control, reinforcement learning as well as applications in communication, wireless and vehicular traffic networks.