2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

The Curse of Memory in Stochastic Approximation

Caio Kalil Lauand and Sean Meyn[1]

***Abstract*— Theory and application of stochastic approximation (SA) has grown within the control systems community since the earliest days of adaptive control. This paper takes a new look at the topic, motivated by recent results establishing remarkable performance of SA with (sufficiently small) constant step-size $\alpha > 0$. If averaging is implemented to obtain the final parameter estimate, then the estimates are asymptotically unbiased with nearly optimal asymptotic covariance. These results have been obtained for random linear SA recursions with i.i.d. coefficients.**

This paper obtains very different conclusions in the more common case of geometrically ergodic Markovian disturbance: (i) The *target bias* is identified, even in the case of non-linear SA, and is in general non-zero. The remaining results are established for linear SA recursions: (ii) the bivariate parameter-disturbance process is geometrically ergodic in a topological sense; (iii) the representation for bias has a simpler form in this case, and cannot be expected to be zero if there is multiplicative noise; (iv) the asymptotic covariance of the averaged parameters is within $O(\alpha)$ of optimal. The error term is identified, and may be massive if mean dynamics are not well conditioned. The theory is illustrated with application to TD-learning.

***Index Terms*— Stochastic Approximation, Stochastic Recursive Algorithms, TD learning.**

I. Introduction

The stochastic approximation (SA) algorithm of Robbins and Monro is designed to solve the root finding problem

$$\bar{f}(\theta^*) = 0, \textit{ in which } \quad \bar{f}(\theta) := \mathsf{E}[f(\theta, \Phi)]\,, \quad \theta \in \mathbb{R}^d\,, \quad (1)$$

$\mathbf{\Phi}$ a random variable taking values in a set X, and $f\colon \mathbb{R}^d \times \mathsf{X} \to \mathbb{R}^d$. The SA algorithm is the d-dimensional recursion,

$$\theta_{n+1} = \theta_n + \alpha_{n+1} f(\theta_n, \Phi_{n+1})\,, \quad n \ge 0 \qquad (2)$$

with initial condition $\theta_0 \in \mathbb{R}^d$, a non-negative step-size sequence $\{\alpha_n\}$, and $\Phi_n \xrightarrow{\text{d}} \Phi$ as $n \to \infty$ (convergence in distribution). The sequence $\mathbf{\Phi} = \{\Phi_k\}$ is often assumed to be Markovian, which is the setting of the present paper. Convergence theory is based on comparison of (2) with the *mean flow*,

$$\tfrac{d}{dt}\vartheta_t = \bar{f}(\vartheta_t)\,. \qquad (3)$$

Much of this theory is based on a vanishing step-size sequence, with $\alpha_n = n^{-\rho}$ a common choice. The constraint $\rho \in (1/2, 1]$ is imposed so that the step-size is square summable, but $\sum_n \alpha_n = \infty$. Almost sure convergence of $\{\theta_n\}$ to θ^* holds under minimal assumptions on $\bar{f}$ and $\mathbf{\Phi}$. The Markovian assumption is *not* required—see Chapter 2 of [3] for consistency results under minimal conditions.

Theory for convergence rates remains a research frontier. It is known that the mean-squared error (MSE) vanishes slowly: for vanishing gain algorithms, one expects the bound $\mathsf{E}[\|\theta_n - \theta^*\|^2] = O(\alpha_n)$. One approach to speed up convergence while also optimizing variance is to employ Polyak-Ruppert (PR) averaging [18], [17]:

$$\theta_N^{\mathsf{PR}} = \frac{1}{N - N_0} \sum_{k=N_0+1}^{N} \theta_k\,, \quad 0 < N_0 < N \qquad (4)$$

with $N_0 > 0$ introduced to discard transients.

Subject to conditions on f, $\mathbf{\Phi}$ and the step-size sequence, averaging achieves two benefits: the MSE decays at rate $O(N^{-1})$, and the *asymptotic covariance* is minimal. That is, $\Sigma^{\mathsf{PR}} = \Sigma^*_\theta$, with

$$\Sigma^{\mathsf{PR}} := \lim_{N\to\infty} N\mathsf{E}[(\theta_N^{\mathsf{PR}} - \theta^*)(\theta_N^{\mathsf{PR}} - \theta^*)^\intercal] \qquad (5a)$$

$$\Sigma^*_\theta = [A^*]^{-1}\Sigma_{\mathcal{W}^*}[A^{*\intercal}]^{-1} \qquad (5b)$$

in which $A^* = \partial_\theta \bar{f}\,(\theta^*)$, and $\Sigma_{\mathcal{W}^*}$ is the "noise covariance matrix" defined in (18). The matrix Σ^*_θ is minimal in the matricial sense.

Despite this attractive theory for SA with vanishing step-size, many practitioners advocate a fixed step-size, $\alpha_n \equiv \alpha > 0$. There is little hope for convergence of $\{\theta_n\}$ in this case, but bounds on bias and variance can be obtained once boundedness of the recursion is established. Similar to the vanishing step-size case, the MSE is determined by the step-size, $\limsup_{n\to\infty} \mathsf{E}[\|\theta_n - \theta^*\|^2] = O(\alpha)$ (see [4] for sufficient conditions).

Recent research provides a bridge between theory and practice, through the application of PR averaging in algorithms with fixed step-size [1], [16]. These papers obtain not only convergence of filtered estimates $\{\theta_N^{\mathsf{PR}}\}$ to θ^*, but also establish the optimal $O(N^{-1})$ convergence rate for the MSE. These positive results come with a large price: it is assumed that $\mathbf{\Phi}$ is an independent and identically distributed (i.i.d.) sequence. Such strong assumptions are rarely justified.

A major goal of the present paper is to show that averaging cannot in general eliminate bias for the Markovian SA recursion with constant step-size. We also find that the covariance matrix (5a) differs from the ideal.

TD-Learning: An illustration of the theory is provided in Section III-A using an instance of TD-learning in an ideal setting: the true value function lies within the span of the two-dimensional function class. The parameter sequence $\{\theta_n\}$ evolves in $\mathbb{R}^2$, and is convergent to the optimal θ^* for the standard setting with vanishing step-size.

*This work was supported by ARO award W911NF2010055 and NSF award EPCN 1935389

[1]Caio Kalil Lauand and Sean Meyn are with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL, USA. Emails: caio.kalillauand@ufl.edu and meyn@ece.ufl.edu

For the fixed step-size algorithm there is bias: $\theta_\infty^{\mathsf{PR}} = \lim_{n\to\infty} \theta_n^{\mathsf{PR}} \neq \theta^*$. The extensive literature on the central limit theorem for SA suggests that the PR averaged estimates admit the approximation,

$$\theta_N^{\mathsf{PR}} \overset{\text{dist}}{\approx} \sqrt{\frac{\alpha}{N}} W_\infty + \theta_\infty^{\mathsf{PR}}, \quad W_\infty \sim N(0, \Sigma^{\mathsf{PR}}) \tag{6}$$

where the asymptotic covariance Σ^{PR} is approximately equal to Σ_θ^* for small α (see (5b) for the definition).

Fig. 1(a) shows the L_2 norm of the estimation error obtained from PR averaged estimates as a function of α, and a fixed value of $N = 5 \times 10^5$. The plot is consistent with the heuristic (6):

Small step-size: for $\alpha < 10^{-3}$ the L_2-error is approximated well by an affine function of $\sqrt{\alpha}$. This would be anticipated by (6) and the theory in this paper, establishing that the bias $\|\theta_\infty^{\mathsf{PR}} - \theta^*\|$ is of order $O(\alpha)$.

Large step-size: for $\alpha > 10^{-3}$ the L_2-error is approximately affine as a function α. It appears that the bias dominates variance in this regime.

The histograms shown in Fig. 1 (b) compare performance of the algorithm with two choices of step-size. The vanishing gain algorithm is best in terms of both bias and variance. To obtain comparable mean and variance with a constant step-size algorithm would require a very small value of $\alpha > 0$ and a larger value of N.

Contributions: This brings us to the main contributions of this paper. The first two concern the nonlinear SA recursion.

(i) Moment bounds are obtained for the fixed step-size algorithm: there is $\alpha_0 > 0$ and $b^{2.1} < \infty$, such that for $0 < \alpha \le \alpha_0$ and each $\theta_0 \in \mathbb{R}^d$,

$$\limsup_{n\to\infty} \mathsf{E}[\|\theta_n - \theta^*\|^4] \le b^{2.1}\alpha^2 \tag{7}$$

(ii) The *target bias* is of order $O(\alpha)$:

$$\limsup_{N\to\infty} \left\| \frac{1}{N} \sum_{k=0}^{N-1} \mathsf{E}[\bar{f}(\theta_k)] \right\| = \alpha \limsup_{N\to\infty} \left\| \frac{1}{N} \sum_{k=0}^{N-1} \mathsf{E}[\Upsilon_k] \right\| \tag{8}$$

in which the limit supremum on the right hand side is uniformly bounded in α. The d-dimensional sequence $\{\Upsilon_k\}$ plays a crucial role in this paper, appearing for the first time in the *disturbance decomposition* (12).

Contributions (i) and (ii) summarize Thm. 2.1, which concerns the SA recursion subject to standard assumptions on f, but strong assumptions on the Markov chain. Strong assumptions are in general necessary to obtain moment bounds—see [2] for a counter example based on linear SA with vanishing gain.

The pair process $\{\theta_n, \Phi_{n+1}\}$ is a time homogeneous Markov chain. Under the assumptions of the paper it is possible to establish the existence of an invariant measure, but we have not yet established ergodicity as would be required to improve the conclusions in (i) and (ii). Sharper results are obtained for the linear SA recursion,

$$\theta_{n+1} = \theta_n + \alpha[A_{n+1}\theta_n - b_{n+1}] \tag{9}$$

in which $(A_n; b_n)$ is a function of Φ_n for each n.

The remaining contributions are established for (9): there is a unique invariant measure for the pair process, which is geometrically ergodic in a topological sense. This provides a framework for analysis that brings us to the following conclusions:

(iii) The bias admits the representation,

$$\lim_{n\to\infty} \mathsf{E}[\theta_n^{\mathsf{PR}}] = \lim_{n\to\infty} \mathsf{E}[\theta_n] = \theta^* + \alpha[A^*]^{-1}\bar{\Upsilon}^* + O(\alpha^2) \tag{10}$$

where $A^* = \partial_\theta \bar{f}(\theta^*)$ and $\bar{\Upsilon}^* \in \mathbb{R}^d$ is identified in Thm. 2.5.

(iv) For a $d \times d$ matrix Z, and Σ_θ^* defined in (5b),

$$\lim_{N\to\infty} N\text{Cov}\,(\theta_N^{\mathsf{PR}}) = \Sigma_\theta^* - \alpha Z + O(\alpha^2) \tag{11}$$

Survey of relevant literature and approach to analysis. While the present paper was inspired by the recent articles [1], [16], the contents are more closely related to [6], [7] which also treat the fixed step-size SA algorithm with Markovian noise. The main conclusions are L_2-bounds on the parameter error, and geometric ergodicity in the same topological sense as in the present paper.

The conditions on $\mathbf{\Phi}$ in [6] are milder—only geometric ergodicity is imposed—but the assumptions on f are far stronger. There is no representation for bias.

A bias representation for SA with constant gain first appeared in the preprint [11] and the extended abstract [10]. Analysis of bias for linear recursions is considered in [7] subject to a stronger *uniform ergodicity* assumption on $\mathbf{\Phi}$. The stronger condition is imposed because they seek uniform bounds on the transient behavior of the algorithm.

Analysis of the SA recursion (2) commonly begins with its interpretation as a "noisy" Euler approximation of the mean flow, with "noise" or "disturbance" $\Delta_{n+1} := f(\theta_n, \Phi_{n+1}) - \bar{f}(\theta_n)$. The starting point of analysis in this paper is the decomposition of Métivier and Priouret [13],

$$\Delta_{n+1} = \mathcal{W}_{n+2} - \mathcal{T}_{n+2} + \mathcal{T}_{n+1} - \alpha\Upsilon_{n+2}, \tag{12}$$

in which $\{\mathcal{W}_{n+2}\}$ is a martingale difference sequence. This prior work considered the case of vanishing step-size, and based on this representation established convergence of the SA algorithm. The idea has been applied in many other papers, and in particular leads to a functional Central Limit Theorem under suitable conditions. The weakest conditions to-date are found in [2], and this recent prior work is a major foundation of the present paper. Key lemmas from [2] extend to the setting of this paper, which form components of the proof of eqs. (7) and (8).

Finer results, such as the final set of conclusions in eqs. (10) and (11), require multiple applications of the disturbance decomposition (12) to general functions of (θ_n, Φ_{n+1}). We are not aware of repeated application of these martingale approximations in prior work.

Organization. The remainder of the paper is organized in three sections. Section II summarizes the assumptions and notation imposed throughout the paper and presents contributions (i)–(iv), along with proof outlines for each result.

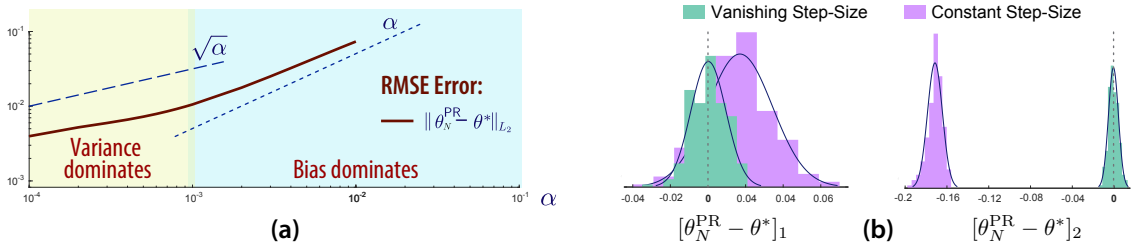

Fig. 1: Performance of TD(λ) Learning with $\lambda = 0$. (a) L_2 norm of estimation error for the fixed step-size algorithm as a function of α. (b) Histogram of estimation error for vanishing and constant steps-size algorithms for each dimension of θ^{PR}.

Section III contains numerical experiments that illustrate the theory in Section II, including details on the experiments supporting Fig. 1. Conclusions and open paths for future research are contained in Section IV. Selected technical results supporting the main conclusions of the paper are contained in the Appendix. Full proofs for each of the main results can be found in the preprint version of this article [12].

II. Main Results

A. Preliminaries

It is assumed that $\mathbf{\Phi} := \{\Phi_n\}$ in (2) is a geometrically ergodic Markov chain, whose state space X is a locally compact and separable metric space. Its transition kernel is denoted P, and unique invariant measure π, so that $\bar{f}(\theta) = \mathsf{E}_\pi[f(\theta, \Phi_k)]$. The subscript in the expectation defining $\bar{f}$ indicates that it is taken in steady-state: $\Phi_k \sim \pi$.

The following assumptions are in place throughout:

(A1) The SA recursion (2) is considered with $\alpha_n \equiv \alpha > 0$.

(A2i) There is a function $L\colon \mathsf{X} \to \mathbb{R}$ satisfying $\|f(\theta, x) - f(\theta', x)\| \le L(x)\|\theta - \theta'\|$ and $\|f(0, x)\| \le L(x)$ for all $x \in \mathsf{X}$ and $\theta, \theta' \in \mathbb{R}^d$.

(A2ii) $\mathbf{\Phi}$ is an aperiodic Markov chain satisfying **(DV3)**: For functions $V\colon \mathsf{X} \to \mathbb{R}_+$, $W\colon \mathsf{X} \to [1, \infty)$, a small set C, $b > 0$ and all $x \in \mathsf{X}$,

$$\mathsf{E}\big[\exp\big(V(\Phi_{k+1})\big) \mid \Phi_k = x\big] \le \exp\big(V(x) - W(x) + b\mathbb{I}_C(x)\big)$$

In addition, $S_W(r) := \{x : W(x) \le r\}$ is either small or empty and $\sup\{V(x) : x \in S_W(r)\} < \infty$.

See [15] for the definition of a small set.

(A3) The scaled vector field $\bar{f}_\infty(\theta)$ exists for each $\theta \in \mathbb{R}^d$: $\frac{1}{c} f(c\theta) \to \bar{f}_\infty(\theta)$ as $c \to \infty$. Moreover, the ODE@∞, $\frac{d}{dt}\vartheta_t^\infty = \bar{f}_\infty(\vartheta_t^\infty)$ is globally asymptotically stable.

(A4) $\lim_{n\to\infty} \Big(\sup\Big\{\dfrac{L(x)}{W(x)} : W(x) \ge n\Big\}\Big) = 0$.[1]

(A5) $\bar{f}\colon \mathbb{R}^d \to \mathbb{R}^d$ is continuously differentiable in θ, and the Jacobian matrix $\bar{A} = \partial \bar{f}$ is uniformly bounded and uniformly Lipschitz continuous. Moreover, $A^* := \bar{A}(\theta^*)$ is Hurwitz.

The mean flow (3) is exponentially asymptotical stable under (A3) and (A5) [11] (further results are obtained in [20] under stronger conditions). Geometric ergodicity of $\mathbf{\Phi}$ follows from (A2)—see [9] or [15, Chapter 20.1].

[1] A relaxation of (A4) may be found in [2].

More notation: For a sequence $\{\gamma_n\}$ and a non-negative sequence $\{\kappa_n\}$, we write $\gamma_n = O(\kappa_n)$ if there is a constant $b < \infty$ such that $\|\gamma_n\| \le b\kappa_n$ for all n (the first sequence may be vector valued).

The joint process $\mathbf{\Psi} := \{\Psi_n = (\theta_n, \Phi_{n+1}) : n \ge 0\}$ is Markovian. When an invariant measure ϖ exists for $\mathbf{\Psi}$, its second marginal is the invariant measure π for $\mathbf{\Phi}$. Any functions are assumed to be measurable with respect to the Borel sigma-algebra over their domain.

For functions $g, h : \mathbb{R}^d \times \mathsf{X} \to \mathbb{R}$, we denote

$$\bar{g}(\theta) = \int g(\theta, x)\pi(dx)\,, \qquad \tilde{g}(z) = g(z) - \varpi(g)\,,$$

with $\varpi(g) = \int g(z)\varpi(dz)$. We adopt the following L_2 notation, and also notation for asymptotic covariances of vector-valued functions:

$$\langle g, h\rangle_{L_2} = \mathsf{E}_\varpi[g(\Psi_0)h(\Psi_0)^\intercal]\,, \quad \Sigma_g = \langle g, g\rangle_{L_2} \tag{13}$$

$$\langle g, h\rangle_{\mathsf{CLT}} = \sum_{k=-\infty}^{\infty} \mathsf{E}_\varpi[\tilde{g}(\Psi_0)\tilde{h}(\Psi_k)^\intercal]\,, \quad \Sigma^g_{\mathsf{CLT}} = \langle g, g\rangle_{\mathsf{CLT}}\,.$$

The notation is extended to stationary realizations of vector-valued stochastic processes $\{\mathcal{G}_k, \mathcal{H}_k\}$. In particular, $\langle \mathcal{G}, \mathcal{H}\rangle_{\mathsf{CLT}} = \sum_{k=-\infty}^{\infty} \mathsf{E}_\varpi[\widetilde{\mathcal{G}}_0\widetilde{\mathcal{H}}_k^\intercal]$, with $\{\widetilde{\mathcal{G}}_k, \widetilde{\mathcal{H}}_k\}$ the zero mean centered processes.

B. Moment Bounds

Moment bounds are obtained in [2] for vanishing step-size algorithms by establishing a Lyapunov drift condition for the function $\mathcal{V}(\theta, x) = (1 + \beta\|\theta\|^4)v_+(x)$, with $\beta > 0$,

$$v_+(x) = \mathsf{E}[\exp\big(V(\Phi_{k+1}) + \varepsilon^\circ W(\Phi_k)\big) \mid \Phi_k = x] \tag{14}$$

and $\varepsilon^\circ < 1$. It is shown that similar Lyapunov bounds hold in the setting of this paper, leading to the following conclusions:

Theorem 2.1: Suppose that *(A1)-(A5)* hold. Then there exists $b^{2.1} < \infty$ and $\alpha_0 > 0$ such that for $0 < \alpha \le \alpha_0$,

(i) $\sup_{k,z} \dfrac{1}{\mathcal{V}(z)} \mathsf{E}\big[\mathcal{V}(\Psi_k) \mid \Psi_0 = z\big] \le b^{2.1}$

(ii) $\limsup_{n\to\infty} \mathsf{E}\big[\|\tilde{\theta}_n\|^4\big] \le b^{2.1}\alpha^2$

(iii) The bound (8) holds for each Ψ_0.

Proof (outline): Introducing the suggestive notation

$$\theta_{n+1} = \theta_n + \alpha[\bar{f}(\theta_n) + \Delta_{n+1}] \tag{15}$$

we define "sampling times" to define the ODE approximation: for each k, let $\tau_k = \alpha k$. For a given $T > 0$, let $T_0 = 0$ and $T_{n+1} = \min\{\tau_k : \tau_k \geq T_n + T\}$. Consequently, the sequence $\{T_n\}$ satisfies $T \leq T_{n+1} - T_n \leq T + \alpha$ for each n. Let $m_0 := 0$ and m_n the integer satisfying $\tau_{m_n} = T_n$ for each $n \geq 1$.

No assumptions in [2] preclude $\{\alpha_k\}$ from being constant on any of the finite intervals $\{k : m_n \leq k < m_{n+1}\}$. Consequently, any of the finite-interval bounds in [2] are valid here, on choosing $\alpha_k \equiv \alpha$ on any such interval. Part (i) is a direct corollary to [2, Lemma A.18].

[2, Proposition A.22] obtains a uniform L_4 bound on $z_k^{(n)} = \frac{1}{\sqrt{\alpha_k}}[\theta_k - \vartheta_t^{(n)}]$ where $\{\vartheta_t^{(n)} : t \geq \tau_{m_n}\}$ denotes the solution to the mean flow initialized with $\vartheta_{\tau_{m_n}}^{(n)} = \theta_{m_n}$. This bound can be extended to constant gain, and we then obtain (ii) from this and exponential asymptotic stability of the mean flow (3) following the proof of [2, Lemma A.23 (ii)].

Part (iii) follows from (15) and the decomposition (12). ☐

The disturbance decomposition (12) was introduced in [13], beginning with the solution $\hat{f}$ to *Poisson's equation* with forcing function f: for $\theta \in \mathbb{R}^d$, $x \in \mathsf{X}$,

$$\mathsf{E}[\hat{f}(\theta, \Phi_{n+1}) - \hat{f}(\theta, \Phi_n) \mid \Phi_n = x] = -f(\theta, x) + \bar{f}(\theta) \quad (16)$$

The solution $\hat{f}$ is unique up to an additive constant under the assumptions imposed on f and $\mathbf{\Phi}$.

Proposition 2.2: If (A2) holds then $\hat{f} \colon \mathbb{R}^d \times \mathsf{X} \to \mathbb{R}^d$ exists solving (16), with $\mathsf{E}_\pi[\hat{f}(\theta, \Phi_n)] = 0$ for each $\theta \in \mathbb{R}^d$, and for a constant b_f and all θ, θ', x:

$\|\hat{f}(\theta, x)\| \leq b_f\big(1 + V(x)\big)\big[1 + \|\theta\|\big]$

$\|\hat{f}(\theta, x) - \hat{f}(\theta', x)\| \leq b_f\big(1 + V(x)\big)\|\theta - \theta'\|$

Proof: (DV3) together with Jensen's inequality gives

$$\mathsf{E}\big[V(\Phi_{k+1}) \mid \Phi_k = x\big] \leq V(x) - W(x) + b\mathbb{I}_C(x)$$

The result then follows from [15, Theorem 17.4.2]. ☐

The terms in (12) admit representations in terms of $\hat{f}$:

Lemma 2.3: eq. (12) holds under (A2) with

$$\mathcal{W}_{n+2} := \hat{f}(\theta_n, \Phi_{n+2}) - \mathsf{E}[\hat{f}(\theta_n, \Phi_{n+2}) \mid \mathcal{F}_{n+1}], \quad (17a)$$

$$\mathcal{T}_{n+1} := \hat{f}(\theta_n, \Phi_{n+1}), \quad (17b)$$

$$\Upsilon_{n+2} := -\frac{1}{\alpha}\big[\hat{f}(\theta_{n+1}, \Phi_{n+2}) - \hat{f}(\theta_n, \Phi_{n+2})\big] \quad (17c)$$

The sequence $\{\Upsilon_{n+2}\}$ defined in (17c) is most important in finer analysis of bias and variance in linear stochastic approximation. The martingale difference sequence dominates the variance. We denote

$$\Sigma_{\mathcal{W}} = \langle \mathcal{W}, \mathcal{W}^{\intercal} \rangle_{\mathsf{CLT}}, \quad \Sigma_{\mathcal{W}^*} = \langle \mathcal{W}^*, \mathcal{W}^{*\intercal} \rangle_{\mathsf{CLT}},$$

$$\textit{with } \mathcal{W}^*_{n+1} := \hat{f}(\theta^*, \Phi_{n+1}) - \mathsf{E}[\hat{f}(\theta^*, \Phi_n) \mid \mathcal{F}_n] \quad (18)$$

$$\textit{and } \Upsilon^*_n = -\widehat{A}_n A_{n-1}\theta^* + \widehat{A}_n b_{n-1}$$

where $A_n := A(\Phi_n) = \partial_\theta f(\theta, \Phi_n)$ and $\widehat{A}$ denotes the solution to Poisson's equation with forcing function A. In the special case considered in Thm. 2.5 where f is affine in θ, its Jacobian A is independent of θ.

We also require solutions to Poisson's equation for the full process $\mathbf{\Phi}$. Under conditions on the forcing function $g \colon \mathbb{R}^d \times \mathsf{X} \to \mathbb{R}^m$, a zero-mean solution is obtained:

$$\check{g}(\theta, x) = \sum_{k=0}^{\infty} \mathsf{E}[\tilde{g}(\theta_k, \Phi_{k+1}) | \theta_0 = \theta, \Phi_1 = x] \quad (19)$$

where the sum exists for all $\theta \in \mathbb{R}^d, x \in \mathsf{X}$. With (19) in hand, an alternative representation is obtained for the asymptotic covariances (13) of vector-valued functions:

$$\langle g, h \rangle_{\mathsf{CLT}} = \langle \check{g}, h \rangle_{L_2} + \langle g, \check{h} \rangle_{L_2} - \langle g, h \rangle_{L_2} \quad (20a)$$

$$\|\langle h, h \rangle_{\mathsf{CLT}}\|_F \leq 2\|\tilde{h}\|_{L_2}\|\check{h}\|_{L_2} \quad (20b)$$

C. Sensitivity

The theory of Lyapunov exponents is a promising approach to establish a form of ergodicity for the bivariate process $\mathbf{\Phi}$.

The *sensitivity process* is defined by $\zeta_n^0 = \partial_{\theta_0}\theta_n$, in which θ_n is viewed as a smooth function of the initial condition; this requires common randomness for each initial condition in (2). It evolves according to the recursion,

$$\zeta_{n+1}^0 = \zeta_n^0 + \alpha A_{n+1}\zeta_n^0, \quad A_{n+1} := \partial_\theta f(\theta_n, \Phi_{n+1}) \quad (21)$$

The L_p-Lyapunov exponent λ_p is the growth rate,

$$\log(\lambda_p) := \lim_{n\to\infty} \frac{1}{n} \log(\mathsf{E}[\|\zeta_n^0\|_F^p]^{1/p})$$

where $\|\cdot\|_F$ denotes the Frobenious norm. If this limit exists and is negative, then parameter sequences from distinct initial conditions converge to a steady-state in a topological sense.

Consider the scaled sensitivity process defined by $\zeta_n = \exp(n\delta_s\alpha)\zeta_n^0$, with $\delta_s > 0$. Multiplying each side of (21) by $\exp(\delta_s\alpha)$ results in the recursion,

$$\zeta_{n+1} = \zeta_n + \alpha\big[M_\delta + A_{n+1}\big]\zeta_n \quad (22)$$

where $M_\delta := \alpha^{-1}[\exp(\delta_s\alpha) - 1]I$. Associated with the pair of recursions (2, 22) is the $2d$ dimensional mean-flow,

$$\tfrac{d}{dt}\vartheta_t = \bar{f}(\vartheta_t), \quad \tfrac{d}{dt}z_t = \big[M_\delta I + \bar{A}(\vartheta_t)\big]z_t$$

Under the assumptions of Thm. 2.1, the first ODE is globally exponentially stable with unique equilibrium θ^*. Under (A5) the matrix $A^* := \bar{A}(\theta^*)$ is Hurwitz, which tells us how we should choose δ_s: sufficiently small so that $M_\delta + A^*$ is Hurwitz for all $\alpha \in (0, \alpha_0]$. The bivariate mean flow is then globally asymptotically stable.

Unfortunately we cannot apply Thm. 2.1 to the joint SA recursion generating $(\theta_n; \zeta_n)$ because the right hand side of (22) is not jointly Lipschitz continuous in $(\theta_n; \zeta_n)$.

We turn next to the linear setting in which $\bar{A}(\theta) \equiv A^*$, so that the required Lipschitz conditions hold.

D. Linear Stochastic Approximation

The linear SA recursion and scaled sensitivity process are viewed as a single SA recursion: for $(\theta_0; \zeta_0) \in \mathbb{R}^{2d}$,

$$\theta_{n+1} = \theta_n + \alpha[A_{n+1}\theta_n - b_{n+1}] \quad (23a)$$

$$\zeta_{n+1} = \zeta_n + \alpha\big[M_\delta + A_{n+1}\big]\zeta_n. \quad (23b)$$

The second recursion differs from (22) in two respects. First, ζ_n is a d-dimensional vector and not a matrix; if $\zeta_0 = e^i$ (the ith basis vector), then $\zeta_n = \exp(n\delta_s\alpha)\partial_{\theta_0^i}\theta_n$. Second, $A_{n+1} = A(\Phi_{n+1})$ and $b_{n+1} = b(\Phi_{n+1})$ do not depend upon θ_n, giving $\zeta_n = \exp(n\delta_s\alpha)A_n \cdots A_1\zeta_0$.

The disturbance decomposition also simplifies:

$$\begin{aligned}\hat{f}(\theta_n, \Phi_{n+1}) &= \widehat{A}_{n+1}\theta_n - \widehat{b}_{n+1} \\ \Upsilon_{n+2} &= -\widehat{A}_{n+2}A_{n+1}\theta_n + \widehat{A}_{n+2}b_{n+1}\end{aligned} \tag{24}$$

Using $\bar{f}(\theta) = A^*(\theta - \theta^*)$, with $\theta^* = [A^*]^{-1}\bar{b}$, the mean flow for the pair process is linear:

$$\tfrac{d}{dt}\vartheta_t = A^*(\vartheta_t - \theta^*)\,, \quad \tfrac{d}{dt}z_t = \big[M_\delta + A^*\big]z_t$$

Consequently, (23) satisfies the assumptions of Thm. 2.1.

Theorem 2.4: Suppose that *(A1)-(A5)* hold, and that $\delta_s > 0$ is chosen so that $M_\delta + A^*$ is Hurwitz for all $\alpha \in (0, \alpha_0]$. Then, there is $b^{2.4} < \infty$ such that for $0 < \alpha \le \alpha_0$,

(i) $\displaystyle\sup_{k,z,\zeta} \frac{\mathsf{E}\big[(1 + \|\zeta_k\|^4)\mathcal{V}(\Psi_k) \mid \Psi_0 = z, \zeta_0 = \zeta\big]}{(1 + \|\zeta\|^4)\mathcal{V}(z)} \le b^{2.4}$

(ii) There is a unique invariant measure ϖ for $\mathbf{\Psi}$, and constant $\varrho_{2.4} < 1$ for which the following conclusions hold: if $G\colon \mathbb{R}^d \times \mathsf{X} \to \mathbb{R}$ satisfies the bounds,

$$\begin{aligned}|G(\theta, x)| &\le b_G\mathcal{V}^{1/4}(\theta, x) \\ |G(\theta, x) - G(\theta', x)| &\le b_G v_+^{1/4}(x)\|\theta - \theta'\|\end{aligned}$$

for some $b_G < \infty$ and all $z = (x, u) \in \mathbb{R}^d \times \mathsf{X}$, then

$$\mathsf{E}_\varpi[|\widetilde{G}(\theta_0, \Phi_1)|^2] \le b^{2.4}b_G \tag{25}$$

where $\widetilde{G} = G - \varpi(G)$, and for each initial condition,

$$|\mathsf{E}[\widetilde{G}(\Psi_k) \mid \Psi_0 = z]| \le b_G b^{2.4}\varrho_{2.4}^k\mathcal{V}^{1/2}(z)\,, \quad k \ge 0 \tag{26}$$

Proof (outline): Part (i) follows from Thm. 2.1. The proof of (ii) begins with establishing $\lambda_p \le \exp(-\delta_s\alpha)$ with $p = 4$, which in particular implies the existence of ϖ [6]. □

The Lyapunov exponent bound $\lambda_4 \le \exp(-\delta_s\alpha)$ in Thm. 2.4 is a consequence of the following.

Theorem 2.5: The following hold under the assumptions of Thm. 2.4, for each $0 < \alpha \le \alpha_0$:

(i) The bias approximation (10) holds with $\bar{\Upsilon}^* = \bar{\Upsilon}(\theta^*)$, where $\bar{\Upsilon}(\theta) := \mathsf{E}_\pi[\Upsilon_{n+1}] = -\mathsf{E}_\pi[\widehat{A}_{n+1}(A_n\theta + b_n)]$.

(ii) The representation (11) for the covariance of the PR averaged estimates holds, in which $Z = [A^*]^{-1}Z^0[A^*]^{-1\intercal}$,

$$Z^0 = \Sigma_{\mathsf{CLT}}^{\Upsilon^*,\mathcal{W}^*} + \big(\Sigma_{\mathsf{CLT}}^{\Upsilon^*,\mathcal{W}^*}\big)^\intercal + \frac{1}{\alpha}\big[\Sigma_{\mathcal{W}^*} - \Sigma_{\mathcal{W}}\big] \tag{27}$$

The right hand side is uniformly bounded in α. The cross covariance term is $\Sigma_{\mathsf{CLT}}^{\Upsilon^*,\mathcal{W}^*} = \displaystyle\sum_{k=-\infty}^{\infty}\mathsf{E}_\varpi[\Upsilon_0^*\mathcal{W}_k^{*\intercal}]$, in which $\{\Upsilon_k^*, \mathcal{W}_k^*\}$ is defined in (18).

Proof (outline): Key to each step is the combination of (23a) and Lemma 2.3, giving

$$\theta_n = \theta^* + [A^*]^{-1}\big[-\Delta_{n+1} + \tfrac{1}{\alpha}[\theta_{n+1} - \theta_n]\big] \tag{28}$$

The proof of (i) begins by taking expectations of each side of (28) in steady state. Applying Lemma 2.3 and Prop. A.2 gives the result in (10).

The proof of (ii) is based on the following steps:

1. $\Sigma_{\mathsf{CLT}}^{\mathcal{G}} = \langle \mathcal{G}, \mathcal{G}\rangle_{\mathsf{CLT}}$ is finite, and uniformly bounded over $0 < \alpha \le \alpha_0$, for stochastic processes of the form $\mathcal{G}_{k+\ell} = \mathcal{M}_{k+\ell}\theta_k + \gamma_{k+\ell}$, with $\ell \ge 1$, and $\mathcal{M}_{k+\ell}, \gamma_{k+\ell}$ functions of $(\Phi_{k+1}; \ldots; \Phi_{k+\ell})$ satisfying growth conditions that hold for the functions of interest—see Prop. A.3.

2. $\lim_{N\to\infty} N\mathrm{Cov}\,(\theta_N^{\mathsf{PR}}) = \Sigma_{\mathsf{CLT}}^{\mathcal{G}^0}$, with $\mathcal{G}_k = \mathcal{G}_k^0 := \theta_k$.

3. A telescoping sequence has zero asymptotic covariance. Hence, $\Sigma_{\mathsf{CLT}}^{\mathcal{G}^0} = \Sigma_{\mathsf{CLT}}^{\mathcal{G}^1}$, with $\mathcal{G}_k^1 = \Delta_k$ and $\mathcal{G}_k^0 = \theta_k$ as above.

4. The final step is to show that $\Sigma_{\mathsf{CLT}}^{\mathcal{G}^1} = \Sigma_{\mathcal{W}^*} + \alpha Z^0 + O(\alpha^2)$, with Z^0 given in (27). Lemma 2.3 gives $\Sigma_{\mathsf{CLT}}^{\mathcal{G}^1} = \langle \mathcal{W} - \alpha\Upsilon_k, \mathcal{W} - \alpha\Upsilon_k\rangle_{\mathsf{CLT}}$, implying something similar to the desired approximation. The final result follows from this combined with Propositions A.2 and A.3. □

III. Numerical Experiments

A. TD learning

We now return to TD learning to explain how the theory in this paper may be applied to inform algorithm design.

Consider the scalar linear system with i.i.d. Gaussian disturbance,

$$X_{n+1} = FX_n + W_{n+1}\,,\ |F| < 1\,,\ W_n \sim N(0, \sigma_W^2) \tag{29}$$

along with the quadratic cost function $c(x) = x^2$. The discounted-cost value function with $\gamma \in (0, 1)$ is finite valued,

$$J(x) = \sum_{k=0}^{\infty}\gamma^k\mathsf{E}[c(X_k)|X_0 = x]\,, \quad x \in \mathbb{R}\,, \tag{30}$$

The goal of TD-learning is to estimate this value function within a specified function class.

With linear function approximation $J^\theta = \theta^\intercal\psi$, and basis $\psi : \mathsf{X} \to \mathbb{R}^d$, the TD($\lambda$)-learning recursion is

$$\begin{aligned}\theta_{n+1} &= \theta_n + \alpha_{n+1}\mathcal{D}_{n+1}\zeta_{n+1} \\ \mathcal{D}_{n+1} &= -J^{\theta_n}(X_n) + c(X_n) + \gamma J^{\theta_n}(X_{n+1})\end{aligned}$$

in which the "eligibility vectors" $\{\zeta_n\}$ are defined by passing $\{\psi(X_n)\}$ through a first-order low-pass filter:

$$\zeta_{n+1} = \lambda\gamma\zeta_n + \psi(X_{n+1})\,, \quad n \ge 0, \textit{ with } \lambda \in [0, 1]. \tag{31}$$

The value function (30) is quadratic in the state:

$$J(x) = \theta_1^* + \theta_2^*x^2\,, \quad \theta^* := \begin{bmatrix}\theta_1^* & \theta_2^*\end{bmatrix}^\intercal \in \mathbb{R}^2\,, \tag{32}$$

which motivates the choice, $\psi(x) = (x^2, 1)^\intercal$.

The discounted objective (30) was estimated using TD(λ) for $M = 200$ independent runs using two different choices of step-size: $\alpha_n \equiv \alpha = 10^{-2}$ and $\alpha_n = \min\{\alpha, n^{-0.8}\}$. Histograms for the estimation error corresponding to both choices of α_n are shown for each component of θ^{PR}. The histograms corresponding to the algorithm with vanishing step-size are centered at 0 and show less spread than the histograms for the algorithm with constant step-size.

The impact of the bias formula (10) is evident. In particular, the histogram for estimates of θ_2^* is centered far from 0. See the extended abstract [10] for other examples showing how to calculate Υ, and showing how bias increases dramatically with Markovian memory.

B. Impact of memory

The next experiment will investigate the impact of long memory in linear SA through the recursion,

$$\begin{aligned} f(\theta_n, \Phi_{n+1}) &= A_{n+1}\theta_n - b + \mathcal{W}_{n+1} \\ \textit{where } A_{n+1} &= -1 + \mathcal{W}_{n+1}\,, \\ \mathcal{W}_{n+1} &= \beta\mathcal{W}_n + \sqrt{1-\beta^2}N_{n+1} \end{aligned} \tag{33}$$

and $\{N_n\}$ is i.i.d. and Gaussian $N(0,1)$. It follows that $A^* = -1$ and $\theta^* = -b$.

The sequence $\{\mathcal{W}_n\}$ resembles the eligibility vector appearing in the TD-algorithms of reinforcement learning (31) [19], [14].

The scaling by $\sqrt{1-\beta^2}$ in (33) is introduced to ensure that the steady-state variance of $\mathcal{W}_n$ is unity, but the asymptotic variance is large when $\beta \sim 1$: $\Sigma^W_{\text{CLT}} = (1+\beta)/(1-\beta)$, using

$$\Sigma^W_{\text{CLT}} = \sum_{n=-\infty}^{\infty} \mathsf{E}[\mathcal{W}_0\mathcal{W}_n] = -\mathsf{E}[\mathcal{W}_0^2] + 2\sum_{n=0}^{\infty}\mathsf{E}[\mathcal{W}_0\mathcal{W}_n]$$

A proof of the following conclusions can be found in the extended version of this article [12].

Proposition 3.1: Consider the linear SA recursion in which $\{\mathcal{W}_n\}$ evolves according to the linear recursion (33), with $0 \le \beta < 1$, and $\boldsymbol{N}$ a standard i.i.d. Gaussian sequence, $N_n \sim N(0,1)$ for each n. The following conclusions hold:

(i) Provided an invariant probability measure $\varpi \sim (\theta_n, \Phi_n)$ exists, the bias is

$$\mathsf{E}_\varpi[\theta_0] = \theta^* - \alpha\mathsf{E}_\varpi[\Upsilon_2] \tag{34a}$$

$$\textit{with} \quad \mathsf{E}_\varpi\big[\Upsilon_2\big] = -\frac{\beta}{1-\beta}[1+\theta^*] \tag{34b}$$

$$-\frac{1}{1-\beta}\mathsf{E}_\varpi\big[\mathcal{W}_2\big[(\mathcal{W}_1-1)[\theta_0-\theta^*]\big] \tag{34c}$$

(ii) The optimal asymptotic covariance (5b) is the scalar

$$\Sigma^*_\theta = [1+\theta^*]^2\Sigma^W_{\text{CLT}} \tag{34d}$$

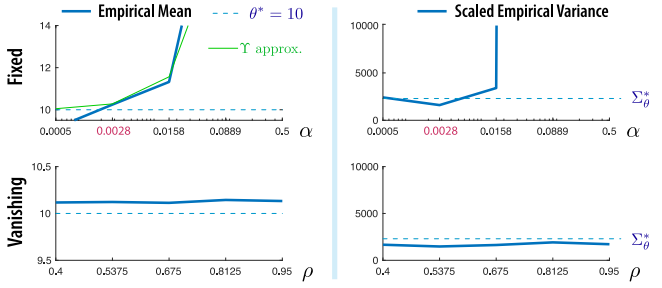

Fig. 2: Comparison of empirical bias and variance obtained from PR-averaging as functions of α for the recursion (33).

The SA recursion (33) was implemented for $\beta = 0.9$, $\theta^* = 10$ and several choices of step-size: $\alpha_n \equiv \alpha$ for constant step-size and $\alpha_n = \frac{1}{2}n^{-\rho}$ for vanishing step-size. Five values of α were tested for the fixed step-size algorithm, and five values of ρ for the vanishing step-size case:

$$\alpha \in \{5\times10^{-4}, 2.8\times10^{-3}, 1.58\times10^{-2}, 8.89\times10^{-2}, 0.5\}$$
$$\rho \in \{0.4000, 0.5375, 0.6750, 0.8125, 0.9\}$$

The estimates for the fixed step-size algorithm remained bounded in n for the range of α tested. The same was true for the vanishing step-size algorithm, as predicted by theory [2]. In application of PR-averaging (4), the value $N_0 = 0.2N$ was chosen in all ten cases. With the given numerical values, applying (34d) gives the approximation for the vanishing gain algorithm, $(N-N_0)\mathsf{E}[(\theta^{\text{PR}}_N - \theta^*)^2] \approx \Sigma^*_\theta \approx 2.3\times10^3$

Fig. 2 shows the estimates of mean and variance obtained in each case. The plot does not reveal much information for the fixed step-size algorithms because most values of α gave poor results. The singular winner over all fixed step-size gains was $\alpha^\star = 2.8\times10^{-3}$, resulting in $\mathsf{E}_\varpi[\theta_0] \approx 10.29$ and $(N-N_0)\text{Cov}_\varpi[\theta_0] \approx 0.7 * \Sigma^*_\theta$. The other four performed far worse.

Each of the experiments using a vanishing gain resulted in variance of approximately equal to what was obtained using $\alpha^\star$ and with smaller bias.

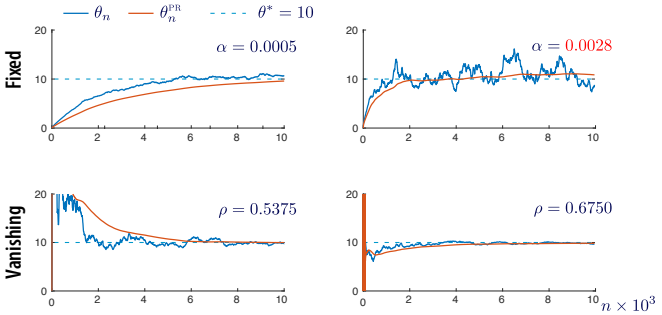

Fig. 3: Evolution of estimates with and without PR averaging from four experiments.

Large bias can be anticipated for the fixed step-size algorithms by consideration of (34c). For small α we obtain an approximation by ignoring the second term in this expression:

$$\mathsf{E}_\varpi[\theta_0] = \theta^* - \alpha\mathsf{E}_\varpi\big[\Upsilon_2\big] \approx \theta^* + \frac{\beta}{1-\beta}[1+\theta^*]\alpha = \theta^* + 99\alpha$$

For $\alpha^\star$ we have $\theta^* + 99\alpha^\star \approx 10.28$, so this approximation nearly matches the approximation $\mathsf{E}_\varpi[\theta_0] \approx 10.29$ obtained through simulation.

See the plot on the upper left hand side of Fig. 2 for a comparison of this approximation with the empirical mean. For the smallest value of α tested, the parameter estimates are far from steady-state by the end of the run. In this case we typically observe *negative* bias.

The cause of the negative bias for $\alpha = 5\times10^{-4}$ is explained by the fact that θ^i_0 is drawn from $N(0,25)$ (so zero mean, while $\theta^* = 10$).

Fig. 3 shows sample paths with and without averaging for two selected values of fixed step-size, and two values of ρ for vanishing step-size, with initialization $\theta_0 = 0$ in each case. The plots using PR-averaging were obtained via (4) with $N_0 = \lfloor 0.8N \rfloor$ for each N. It is clear why $\alpha = 5\times10^{-4}$ fails, and $\alpha = 2.8\times10^{-3}$ performs much better.

IV. Conclusions

There are many other open paths for research:

◇ How far does theory extend to nonlinear SA, subject to (A1)–(A5)? It is hoped that the moment bounds in Thm. 2.1 will imply some variant of the Lyapunov exponent bound to enable a Markovian analysis similar to our treatment of the linear case.

◇ In prior work methods were developed to ensure $\bar{\Upsilon} = 0$ for *quasi-stochastic approximation* [8], [11]. It may be possible to obtain the same conclusion under the assumptions of this paper, but only through algorithm design. See [7] for an approach to eliminate $\bar{\Upsilon}$ based on running two algorithms with differing step-sizes.

◇ It is known that L_p bounds on parameter estimates can be obtain under milder assumptions on $\mathbf{\Phi}$, provided the algorithm is re-started when the parameter estimate exits a prescribed set [5]. A parallel theory for algorithms with constant step-size may allow us to obtain the bias bounds of this paper subject to geometric ergodicity as in [6].

References

[1] F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $o(1/n)$. In *Proc. Advances in Neural Information Processing Systems*, vol. 26, pp 773–781, 2013.

[2] V. Borkar, S. Chen, A. Devraj, I. Kontoyiannis, and S. Meyn. The ODE method for asymptotic statistics in stochastic approximation and reinforcement learning. *arXiv e-prints:2110.14427*, pp 1–50, 2021.

[3] V. S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Hindustan Book Agency, Delhi, India, 2nd edition, 2021.

[4] V. S. Borkar and S. P. Meyn. The ODE method for convergence of stochastic approximation and reinforcement learning. *SIAM J. Control Optim.*, 38(2):447–469, 2000.

[5] L. Gerencsér. A representation theorem for the error of recursive estimators. *SIAM Journal on Control and Optimization*, 44(6):2123–2188, 2006.

[6] J. Huang, I. Kontoyiannis, and S. P. Meyn. The ODE method and spectral theory of Markov operators. In *Lecture Notes in Control and Information Sciences*, T. E. Duncan and B. Pasik-Duncan, editors, pp 205–222, Berlin, 2002. Springer-Verlag.

[7] D. Huo, Y. Chen, and Q. Xie. Bias and extrapolation in Markovian linear stochastic approximation with constant stepsizes. *arXiv.2210.00953 (Abstract in Proc. of ACM SIGMETRICS, pp 81–82, 2023)*, 2022.

[8] C. Kalil Lauand and S. Meyn. Approaching quartic convergence rates for quasi-stochastic approximation with application to gradient-free optimization. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, vol. 35, pp 15743–15756. Curran Associates, Inc., 2022.

[9] I. Kontoyiannis and S. P. Meyn. Large deviations asymptotics and the spectral theory of multiplicatively regular Markov processes. *Electron. J. Probab.*, 10(3):61–123 (electronic), 2005.

[10] C. K. Lauand and S. Meyn. Bias in stochastic approximation cannot be eliminated with averaging. In *Allerton Conference on Communication, Control, and Computing*, pp 1–4, Sep. 2022.

[11] C. K. Lauand and S. Meyn. Markovian foundations for quasi stochastic approximation with applications to extremum seeking control. *arXiv 2207.06371*, 2022.

[12] C. K. Lauand and S. Meyn. The curse of memory in stochastic approximation: Extended version. *arXiv 2309.02944*, 2023.

[13] M. Metivier and P. Priouret. Theoremes de convergence presque sure pour une classe d'algorithmes stochastiques a pas decroissants. *Prob. Theory Related Fields*, 74:403–428, 1987.

[14] S. Meyn. *Control Systems and Reinforcement Learning*. Cambridge University Press, Cambridge, 2022.

[15] S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Cambridge University Press, Cambridge, second edition, 2009. Published in the Cambridge Mathematical Library.

[16] W. Mou, C. Junchi Li, M. J. Wainwright, P. L. Bartlett, and M. I. Jordan. On linear stochastic approximation: Fine-grained Polyak-Ruppert and non-asymptotic concentration. *Conference on Learning Theory and arXiv:2004.04719*, pp 2947–2997, 2020.

[17] B. T. Polyak. A new method of stochastic approximation type. *Avtomatika i telemekhanika (in Russian). translated in Automat. Remote Control, 51 (1991)*, pp 98–107, 1990.

[18] D. Ruppert. Efficient estimators from a slowly convergent Robbins-Monro processes. Technical Report Tech. Rept. No. 781, Cornell University, School of Operations Research and Industrial Engineering, Ithaca, NY, 1988.

[19] R. Sutton and A. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 2nd edition, 2018.

[20] M. Vidyasagar. A new converse Lyapunov theorem for global exponential stability and applications to stochastic approximation. In *IEEE Trans. Automat. Control*, pp 2319–2321. IEEE, 2022. Extended version on arXiv:2205.01303.

Appendix

To prove Thm. 2.5, we are interested in general functions of the joint state process $G : \mathbb{R}^d \times \mathsf{X} \to \mathbb{R}^d$ that are affine in θ, and of the form

$$G(\Psi_k) = M^G(\Phi_{k+1})\theta_k + u^G(\Phi_{k+1})\,, \tag{35}$$

$$\bar{G}(\theta_k) = \overline{M}^G\theta_k + \bar{u}^G \tag{36}$$

Unfortunately, simple covariance bounds for such functions are ill-behaved for small α:

Lemma A.1: If the assumptions of Thm. 2.4 hold for the entries of the vector-valued function G, then $\|\Sigma^G_{\mathsf{CLT}}\|_F \le b^{\text{A.1}}/\alpha$.

Proof: The upper bound $\|\Sigma^G_{\mathsf{CLT}}\|_F \le 2\|\widetilde{G}\|_{L_2}\|\check{G}\|_{L2}$ holds, precisely as in the scalar case (20b). Part (ii) of Thm. 2.4 gives $\|\widetilde{G}\|_{L_2} \le b^{2.4} b_G$. The bound (26) combined with (19) implies $\|\check{G}\|_{L2} \le O(1/\alpha)$. ☐

We require bounds on Σ^G_{CLT} that are uniform in α, and for this we must consider stochastic processes of the form $\{\mathcal{G}_{k+\ell} = \mathcal{M}_{k+\ell}\theta_k + \gamma_{k+\ell}\}$. The result that follows enables a reduction to a more tractable process. See [12] for proofs of the results that follow.

Proposition A.2: Consider the stochastic process $\mathcal{G}_{k+\ell} = \mathcal{M}_{k+\ell}\theta_k + \gamma_{k+\ell}$, with $\mathcal{M}_{k+\ell} = \mathcal{M}(\Phi^{k+\ell}_{k+1})$ and $\gamma_{k+\ell} = \gamma(\Phi^{k+\ell}_{k+1})$. Suppose that $\mathcal{M} \colon \mathsf{X}^\ell \to \mathbb{R}^{d\times d}$ and $\gamma : \mathsf{X}^\ell \to \mathbb{R}^d$ satisfy for some $b_{\mathcal{G}} < \infty$ and all $x^\ell_1 \in \mathsf{X}^\ell$,

$$\mathsf{E}[\|\mathcal{M}(\Phi^{k+\ell}_{k+1})\|^4_F + \|\gamma(\Phi^{k+\ell}_{k+1})\|^4 \mid \Phi_k = x] \le b_{\mathcal{G}} v_+(x)$$

Then,

(i) $\mathcal{G}_{k+\ell} = \mathcal{D}^\circ_{k+\ell}\theta_k + M^\circ(\Phi_{k+1})\theta_k + u^\circ(\Phi_{k+1})$, in which $\{\mathcal{W}^{\mathcal{G}}_{k+\ell} = \mathcal{D}^\circ_{k+\ell}\theta_k\}$ is a martingale difference sequence with

$$\mathsf{E}[\|\mathcal{W}^{\mathcal{G}}_{k+\ell}\|^4_F \mid \Psi_k = z] \le \mathcal{V}(z)$$

$$\|M^\circ(x)\|_F + \|u^\circ(x)\| \le b^{\text{A.2}} v^{1/4}_+(x)$$

(ii) Its expectation in steady state is

$$\mathsf{E}_\varpi[\mathcal{G}_k] = \mathsf{E}_\varpi[\mathcal{G}^*_k] + O(\alpha)$$

where $\mathcal{G}^*_{k+\ell} = \mathcal{M}_{k+\ell}\theta^* + \gamma_{k+\ell}$.

Proposition A.3: Suppose that the stochastic process $\mathcal{G}_{k+\ell}$ satisfies the assumptions of Prop. A.2 with $\ell > 1$. Then, $\Sigma^{\mathcal{G}}_{\mathsf{CLT}} = O(1)$.