2024 IEEE 63rd Conference on Decision and Control (CDC)
December 16-19, 2024. MiCo, Milan, Italy

2024 IEEE 63rd Conference on Decision and Control (CDC) | 979-8-3503-1633-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/CDC56724.2024.10886821

Markovian Foundations for Quasi-Stochastic Approximation in Two Timescales

Caio Kalil Lauand and Sean Meyn[1]

Abstract— **Many machine learning and optimization algorithms can be cast as instances of stochastic approximation (SA). The convergence rate of these algorithms is known to be slow, with the optimal mean squared error (MSE) of order $O(n^{-1})$. In prior work it was shown that MSE bounds approaching $O(n^{-4})$ can be achieved through the framework of quasi-stochastic approximation (QSA); essentially SA with careful choice of deterministic exploration. These results are extended to two time-scale algorithms, as found in policy gradient methods of reinforcement learning and extremum seeking control. The extensions are made possible in part by a new approach to analysis, grounded in the theory of Lyapunov exponents, allowing for the interpretation of two timescale algorithms as instances of single timescale QSA. The general theory is illustrated with applications to extremum seeking control.**

Index Terms— **Stochastic Approximation, Averaging Theory, Extremum Seeking Control**

I. Introduction

Stochastic approximation (SA) remains a significant topic for research since its birth in the 1950s, particularly in the machine learning and optimization communities. Early application to reinforcement learning may be found in [15], [36], [37]; see [30], [29], [31] for more recent advances for applications to machine learning.

Any SA algorithm is designed to solve a root-finding problem $\bar f(\theta^*) = 0$, in which $\bar f : \mathbb{R}^d \to \mathbb{R}^d$ may be expressed $\bar f(\theta) = \mathsf{E}[f(\theta, \Phi)]$ for $\theta \in \mathbb{R}^d$ with Φ a random variable. The basic recursion is

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

where $\{\Phi_n\}$ is a sequence of random vectors converging to Φ (in distribution) and $\{\alpha_n\}$ is a step-size sequence.

The recursion is designed to mimic the *mean flow*, defined as the ODE $\dot{\vartheta}_t = \bar f(\vartheta_t)$. Analysis of (1) proceeds by comparing parameter estimates with solutions to the mean flow [7]. Common choices for $\{\alpha_n\}$ include (i) vanishing step-size sequences of the form $\alpha_n = n^{-\rho}, \rho \in (1/2, 1]$; and (ii) constant step-sizes, in which $\alpha_n \equiv \alpha > 0$. Under general conditions on $\bar f$ and $\boldsymbol{\Phi}$, the choice (i) leads to almost sure convergence of $\{\theta_n\}$ to θ^* for each initial condition $\theta_0 \in \mathbb{R}^d$. For (ii), there is little hope for convergence, though convergence typically holds for the averaged parameters [30], [8], [21], [7].

Bounds on mean squared error (MSE) are necessarily of order $\mathsf{E}[\|\theta_n - \theta^*\|^2] = O(\alpha_n)$ in most cases. This can be improved to $O(n^{-1})$ for vanishing step-sizes with $\rho < 1$, and even for constant step-size algorithms in special cases [30], [29], [21] through the application of *averaging*.

Recent results have established much better MSE bounds when $\boldsymbol{\Phi}$ is deterministic: of order $O(\alpha_n^4)$ with averaging, and when $\boldsymbol{\Phi}$ is a mixture of sinusoids with carefully selected frequencies [19]. Numerical experiments illustrating this substantial acceleration can be found in [19, §3.2].

This deterministic analogue of SA, known as quasi-stochastic approximation (QSA), is largely motivated by applications such as reinforcement learning and gradient-free optimization, in which the algorithm designer also designs $\boldsymbol{\Phi}$ for the purpose of "exploration".

The goal of this paper is to extend the theory of QSA to algorithms in two timescales, for which the algorithm objective is to solve a pair of root finding problems $\bar g(\theta^*, \lambda^*) = \bar h(\theta^*, \lambda^*) = 0$. For notational convenience it is assumed that the dimensions of θ and λ are the same, so that $\bar g, \bar h : \mathbb{R}^{2d} \to \mathbb{R}^d$.

It is assumed that $\bar g$ and $\bar h$ are defined as expectations, expressed here in sample path form

$$\bar g(\theta, \lambda) := \lim_{T\to\infty} \frac{1}{T} \int_0^T g(\theta, \lambda, \xi_t)\, dt \tag{2a}$$

$$\bar h(\theta, \lambda) := \lim_{T\to\infty} \frac{1}{T} \int_0^T h(\theta, \lambda, \xi_t)\, dt \tag{2b}$$

The *probing signal* $\{\xi_t\} \subseteq \mathbb{R}^m$ is of the form $\xi_t = G(\Phi_t)$ in which $\boldsymbol{\Phi}$ is the state process for a dynamical system, interpreted as a deterministic Markov process. Justification for the above limits may be found in [27], [23] under the assumptions of the paper.

The QSA algorithms considered in this paper are expressed in continuous time,

$$\tfrac{d}{dt}\Theta_t = a_t g(\Theta_t, \Lambda_t, \xi_t) \tag{3a}$$

$$\tfrac{d}{dt}\Lambda_t = b_t h(\Theta_t, \Lambda_t, \xi_t) \tag{3b}$$

in which the gain processes $\{a_t, b_t\}$ are non-negative.

Three gain choice settings might be considered:

1. Constant gain: $a_t \equiv \alpha$ and $b_t \equiv \beta$, with $0 < \alpha \ll \beta$. The pair of ODEs (3) is known as a singular perturbation model [13]. The analytical approach of [23] based on the *perturbative mean flow* extends easily to this setting.

2. Vanishing gain: Both tend to zero, and $b_t/a_t \to \infty$ as $t \to \infty$. This is favored in actor-critic methods and in some applications to optimization [15], [16], [4], [3], [2].

*This work was supported by ARO award W911NF2010055 and NSF grant CCF 2306023.

[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`

3. Mixed case: a_t is vanishing, but $b_t \equiv \beta > 0$ is held fixed. One example is extremum seeking control, in which the fast ODE emerges as the state of a high pass filter [24]. Another example is policy gradient methods for reinforcement learning, in which Λ_t is the state process of a dynamical system to be controlled [27].

For any choice of gain, the pair (3) has the *two-time scale* property: a_t is small compared to b_t.

It is assumed that there is a continuous function $\lambda^* \colon \mathbb{R}^d \to \mathbb{R}^d$ satisfying $\overline{h}(\theta, \lambda^*(\theta)) = 0$ for each θ. Analysis is then based on a family of mean flow equations:

$$\tfrac{d}{dt}\vartheta_t = \overline{g}(\vartheta_t, \lambda^*(\vartheta_t)) \tag{4a}$$

$$\tfrac{d}{dt}\lambda_t^\theta = \overline{h}(\theta, \lambda_t^\theta) \tag{4b}$$

The two ODEs do not interact: $\theta \in \mathbb{R}^d$ is held fixed in the *fast* ODE (4b), and the *slow* ODE (4a) is autonomous, since the second argument of $\overline{g}$ is $\lambda^*(\vartheta_t)$.

The present paper focuses on the mixed-gain setting because the applications are most compelling in current research, and because the analysis is most interesting.

As in most papers on two time-scale algorithms, we introduce for the purposes of analysis the family of QSA ODEs parameterized by $\theta \in \mathbb{R}^d$:

$$\tfrac{d}{dt}\Lambda_t^\theta = \beta h(\theta, \Lambda_t^\theta, \xi_t) \tag{5}$$

Under the assumptions imposed, we establish the existence and uniqueness of a steady-state distribution μ_θ for $(\Lambda_t^\theta, \Phi_t)$.

Analysis of the full QSA ODE combines elements of two classical approaches:

1. Singular perturbations. A family of models is considered, parameterized by small $\beta > 0$. In this case the goal is to establish $\Lambda_t \approx \lambda^*(\Theta_t)$ for large t, along with error bounds.

2. Parameter dependent noise. (3a) is regarded as a single timescale algorithm, in which the driving noise $(\boldsymbol{\Lambda}, \boldsymbol{\Phi})$ is parameter dependent. Its mean flow is defined by

$$\tfrac{d}{dt}\vartheta_t = \overline{g}_0(\vartheta_t)\,, \quad \overline{g}_0(\theta) = \int g(\theta, \lambda, G(z))\mu_\theta(d\lambda, dz) \tag{6}$$

The general single timescale algorithm then takes the form

$$\tfrac{d}{dt}\Theta_t = a_t[\overline{g}_0(\Theta_t) + \widetilde{\Xi}_t^0] \tag{7}$$

in which $\widetilde{\Xi}_t^0 := g(\Theta_t, \Lambda_t, \xi_t) - \overline{g}_0(\Theta_t)$.

Contributions The main contributions of the paper are summarized herein.

(i) Thm. 2.1 establishes the *perturbative mean flow* (p-mean flow) representation for the "fast" QSA ODE (3b):

$$\begin{aligned} \tfrac{d}{dt}\Lambda_t &= \beta[\overline{h}(\Theta_t, \Lambda_t) - \beta\overline{\Upsilon}^{\text{ff}}(\Theta_t, \Lambda_t) + \mathcal{W}_t] \\ \mathcal{W}_t &= \sum_{i=0}^{2} \beta^{2-i} \frac{d^i}{dt^i}\mathcal{W}_t^i \end{aligned} \tag{8}$$

in which $\{\mathcal{W}_t^i, \overline{\Upsilon}_t^{\text{ff}} : i = 0, 1, 2\}$ are smooth functions of time identified in the theorem. Moreover, conditions are identified under which $\overline{\Upsilon}^{\text{ff}}$ is identically zero, which has valuable implications to algorithm design.

The representation in (8) invites filtering techniques for error attenuation: the terms $\{\mathcal{W}_t^i : i = 1, 2\}$ are zero-mean and can be attenuated through a second order low-pass filter.

(ii) Estimation error bounds are obtained in Thm. 2.2: for a vector $\theta^\beta \in \mathbb{R}^d$ satisfying $\|\theta^\beta - \theta^*\| = O(\beta)$,

$$\|\Theta_t - \theta^\beta\| = O(a_t)$$

$$\limsup_{t\to\infty} \|\Lambda_t - \lambda^*(\theta^*)\| = O(\beta)$$

(iii) The introduction of filtering in Thm. 2.3 yields attenuation of estimation error. In particular, the limiting parameter $\theta^\beta \in \mathbb{R}^d$ satisfies $\|\theta^\beta - \theta^*\| = O(\beta^2)$.

(iv) Portions of the results given in Thm. 2.2 and Thm. 2.3 are based on the justification of the ODE (7) as an approximation to (3a). Theory is based on Lyapunov exponents to establish the existence of unique invariant measures $\{\mu_\theta : \theta \in \mathbb{R}^d\}$, and solutions to Poisson's equation for $\boldsymbol{\Psi}^\theta = (\boldsymbol{\Lambda}^\theta, \boldsymbol{\Phi})$. Criteria and consequences of a negative Lyapunov exponent are contained in Thm. 2.4.

(v) Examples in Section III illustrate application of the general theory in (i)–(iv).

Extremum seeking control (ESC) is given as an example of mixed-gain two timescale QSA, for which the mean vector fields $\overline{g}$ and $\overline{g}_0$ are identified. The standard ESC algorithm and the 1SPSA algorithm of Spall [34] are not globally stable in general, even when gradient descent is stable, because $\overline{g}_0$ is not Lipschitz continuous. It is shown in [23] that ESC will have a finite escape time from "large" initial conditions even when the objective Γ is a strictly convex quadratic; this paper also provides a remedy through the introduction of a state dependent "exploration gain", which is shown to result in global stability. Extension to the two timescale setting is presented in Section III-C.

Literature Survey Research in singular perturbation theory was extremely active within the control systems community in the 1970s, later serving as a foundation for adaptive control. See [33] for its century long history and [14], [11], [13], [32] for more comprehensive literature surveys on the topic.

Almost sure convergence of $\{\theta_n\}$ to θ^* for two timescale SA was established in [6] under the assumption that the sequence of estimates is uniformly bounded almost surely. Bounds on the MSE appeared soon after for the special case of linear SA [17]. Extensions to non-linear recursions were presented in [28], while criteria for boundedness of estimates appeared in [18]. To the best of our knowledge, the first appearance of two timescale SA with $\boldsymbol{\Phi}$ deterministic is the gradient free optimization algorithm in [3].

Gradient free optimization methods concern the estimation of $\theta^{\text{opt}} \in \arg\min \Gamma(\theta)$ based solely on evaluations of the function $\Gamma \colon \mathbb{R}^d \to \mathbb{R}$, without access to its gradient. A solution based on stochastic approximation (SA) was proposed in the early 1950s by Kiefer and Wolfowitz and refinements followed over the years [12], [34]. ESC theory followed a parallel development, and in fact was born far before the introduction of SA [35], [24].

The introduction of tools from the Markov processes literature to QSA began in [27, Ch. 4], and matured significantly in [19], [23] following the discovery of conditions to ensure existence of well behaved solutions to Poisson's equation. This led to the p-mean flow in [23], which is a refinement of the noise decomposition of [25], [26] based upon Poisson's equation for Markov chains.

A function analogous to $\bar{\Upsilon}^{\mathsf{ff}}$ also appears in the p-mean flow for single timescale QSA. It is shown in [23], [21] that this term is not only a major source of estimation error in constant gain algorithms, but also may slow down convergence rates when the gain is vanishing.

Organization This paper is organized into three additional sections. Section II provides formal statements of contributions (i)–(iv), along with details on assumptions. Section III contains examples based upon the mixed gain algorithms that are the focus of this paper. Conclusions and directions for future research are contained in Section IV. Most of the technical analysis is postponed to the Appendix.

II. Two Timescale Quasi-Stochastic Approximation

A. Assumptions

The m-dimensional probing signal is assumed to be of the form $\xi_t = G_0(\xi_t^0)$ in which $G_0 \colon \mathbb{R}^K \to \mathbb{R}^m$ is continuous, and for fixed frequencies and phases $\{\omega_k\,,\, \phi_k : 1 \le k \le K\}$,

$$\xi_t^0 = [\cos(2\pi[\,\omega_1 t + \phi_1]);\, \dots;\, \cos(2\pi[\,\omega_K t + \phi_K])] \qquad (9)$$

Theory requires an alternative representation, and stronger assumptions: throughout the paper we take $\xi_t = G(\Phi_t)$, in which $\mathbf{\Phi}$ is the K-dimensional clock process that evolves in a compact set of the Euclidean space denoted $\Omega \subset \mathbb{C}^K$. It has entries $\Phi_t^i = \exp(2\pi j[\omega_i t + \phi_i])$ for each i and t and is the state process for a dynamical system $\frac{d}{dt}\Phi_t = W\Phi_t$, where $W := 2\pi j \mathrm{diag}(\omega_i)$.

The following compact notation is adopted for (3):

$$\tfrac{d}{dt} X_t = \mathrm{g}_t f(X_t, \xi_t)\,, \quad \mathrm{g}_t = \begin{bmatrix} a_t I & 0 \\ 0 & \beta I \end{bmatrix} \qquad (10)$$

$$\bar{f}(x) := \lim_{T\to\infty} \frac{1}{T} \int_0^T f(x, \xi_t)\, dt \qquad (11)$$

We assume that $\bar{f}(x^*) = 0$ with $x^* = (\theta^*; \lambda^*(\theta^*))$.

Assumptions are summarized in the following:

(A0i) $\xi_t = G_0(\xi_t^0)$ for all t, with ξ_t^0 defined in (9), and the function $G_0 \colon \mathbb{R}^K \to \mathbb{R}^m$ is analytic.

(A0ii) The frequencies $\{\omega_1\,, \dots\,, \omega_K\}$ are distinct, of the form $\omega_i = \log(a_i/b_i) > 0$ and with $\{a_i, b_i\}$ distinct positive integers.

(A1) $a_t = (1+t)^{-\rho}$ with $\frac{1}{2} < \rho < 1$ and $b_t \equiv \beta > 0$.

(A2) The functions f and $\bar{f}$ are Lipschitz continuous: for a constant $L_f < \infty$ and all x', $x \in \mathbb{R}^{2d}\,, \xi, \xi' \in \mathbb{R}^m$,

$$\begin{aligned} \|f(x', \xi) - f(x, \xi)\| &\le L_f \|x' - x\| \\ \|f(x, \xi') - f(x, \xi)\| &\le L_f \|\xi' - \xi\| \\ \|\bar{f}(x') - \bar{f}(x)\| &\le L_f \|x' - x\| \end{aligned}$$

(A3) For each $\theta \in \mathbb{R}^d$, the ODE $\frac{d}{dt}\lambda_t^\theta = \bar{h}(\theta, \lambda_t^\theta)$ has a unique globally asymptotically stable equilibrium $\lambda^*(\theta)$, where $\lambda^* : \mathbb{R}^d \to \mathbb{R}^d$ satisfies, for a constant $L_\lambda < \infty$,

$$\|\lambda^*(\theta) - \lambda^*(\theta')\| \le L_\lambda \|\theta - \theta'\|\,, \quad \theta, \theta' \in \mathbb{R}^d$$

Moreover, the ODE $\frac{d}{dt}\vartheta_t = \bar{g}(\vartheta_t, \lambda^*(\vartheta_t))$ has a unique globally asymptotically stable equilibrium θ^*.

(A4) $\limsup_{t\to\infty} \|X_t\| \le b^\bullet < \infty$.

(A5) The vector fields f and $\bar{f}$ are each twice continuously differentiable. Moreover, the matrices $\bar{A}^{\mathsf{s}*} := \partial_\theta \bar{g}(\theta^*, \lambda^*(\theta^*))$ and $\{\bar{A}^{\mathsf{f}}(\theta) := \partial_\lambda \bar{h}(\theta, \lambda^*(\theta)) : \theta \in \mathbb{R}^d\}$ are assumed Hurwitz.

(A6) There exist functions $V^\circ \colon \mathbb{R}^d \to \mathbb{R}_+$ and $V^\bullet \colon \mathbb{R}^d \to \mathbb{R}_+$ with bounded gradients satisfying the following bounds for each $\theta, \lambda \in \mathbb{R}^d$:

$$\begin{aligned} \partial_\lambda V^\bullet(\theta, \lambda) \cdot \bar{h}(\theta, \lambda) &\le -\|\lambda - \lambda^*(\theta)\|^2 \\ \partial_\theta V^\circ(\theta) \cdot \bar{g}(\theta, \lambda^*(\theta)) &\le -\|\theta - \theta^*\|^2 \end{aligned}$$

Assumptions (A1)–(A6) are variations on standard assumptions in the stochastic approximation literature [7], [21], [5]. It is conjectured that the ultimate boundedness assumption (A4) will follow from the other assumptions.

Assumption (A0) is far from standard. It is required to obtain error bounds in the main results of the paper.

Markovian foundations The theory in this paper rests on recognition that $\mathbf{\Phi}$ is a Markov process. Some key tools from the theory of Markov processes are firstly ergodicity: $\mathbf{\Phi}$ admits a unique invariant measure π, the uniform distribution on Ω. These observations justify the law of large numbers (LLN) (2)—see [27], [23].

Many of the results of this paper rest on solutions to Poisson's equation, whose definition depends upon context. Here we recall the formulation from [23]: Let $\mathsf{Y} := \mathbb{R}^{2d} \times \Omega$ denote the state space for (Θ, Λ, Φ). We denote $\bar{u}(x) := \int_\Omega u(x,z)\, \pi(dz)$ and $\tilde{u}(x,z) := u(x,z) - \bar{u}(x)$ for any $(x,z) \in \mathsf{Y}$ and continuous vector-valued function u on Y. We say that $\hat{u}$ is the *solution* to Poisson's equation with *forcing function* u if

$$\int_0^T \tilde{u}(x, \Phi_t)\, dt = \hat{u}(x, \Phi_0) - \hat{u}(x, \Phi_T) \qquad (12)$$

for each $T \ge 0$ and $x \in \mathbb{R}^{2d}$.

When there is no risk of confusion, we write u_t instead of $u(X_t, \Phi_t)$ for functions of the larger state process (X_t, Φ_t).

Solutions are not unique, since we may always add a constant to obtain another solution. Throughout the paper it is assumed that the solution is normalized so that $\int \hat{u}(x,z)\, \pi(dz) = 0$ for each x.

If u is smooth in its first variable, then the directional derivatives in the directions g and h are denoted

$$\begin{aligned} &[\mathcal{D}^g u](x,z) = \partial_\theta u\,(x,z) \cdot g(x, G(z)) \\ &[\mathcal{D}^h u](x,z) = \partial_\lambda u\,(x,z) \cdot h(x, G(z))\,, \quad (x,z) \in \mathsf{Y} \end{aligned} \qquad (13)$$

Assumption A0 and Poisson's equation Assumption (A0) may seem overly restrictive, but to-date this is the only known condition under which we can establish solutions to Poisson's equation, and bounds on these solutions [19], [23].

A second implication of (A0) concerns the nuisance term $\bar{\Upsilon}^{\mathsf{ff}}$ appearing in (8). Results in [19], [23] may be extended to the setting of this paper to conclude that $\bar{\Upsilon}^{\mathsf{ff}}(x) \equiv 0$ for each x under (A0), with (A0ii) of particular importance (as counter-examples in this prior work show).

The function $\bar{\Upsilon}^{\mathsf{ff}}$ is constructed based on the solution to Poisson's equation with forcing function $u \equiv f$. The solution $\hat{f}$ shares the same smoothness properties as f; its Jacobian with respect to x, denoted $\widehat{A}(x,z){:=}\partial_x \hat{f}\,(x,z)$, is a uniformly bounded function on Y.

Denote $\Upsilon(x,z) := -\widehat{A}(x,z) f(x, G(z))$. This is a function $\Upsilon\colon \mathsf{Y} \to: \mathbb{R}^{2d}$ which admits the representation,

$$\begin{aligned} \Upsilon(x,z) &= \begin{bmatrix} \Upsilon^{\mathsf{ss}}(x,z) + \Upsilon^{\mathsf{fs}}(x,z) \\ \Upsilon^{\mathsf{sf}}(x,z) + \Upsilon^{\mathsf{ff}}(x,z) \end{bmatrix} \\ \text{where} \quad \Upsilon^{\mathsf{ss}} &= -[\mathcal{D}^g \hat{g}]\,, \quad \Upsilon^{\mathsf{sf}} = -[\mathcal{D}^g \hat{h}] \\ \Upsilon^{\mathsf{fs}} &= -[\mathcal{D}^h \hat{g}]\,, \quad \Upsilon^{\mathsf{ff}} = -[\mathcal{D}^h \hat{h}] \end{aligned} \tag{14}$$

Denoting $\bar{\Upsilon}(x) = \int \Upsilon(x,z)\, \uppi(dz)$ for $x \in \mathbb{R}^{2d}$, which has a similar decomposition, the term $\bar{\Upsilon}^{\mathsf{ff}}$ appearing in (8) is precisely $\bar{\Upsilon}^{\mathsf{ff}}(x) = \int \Upsilon^{\mathsf{ff}}(x,z)\, \uppi(dz)$.

B. Main Results

The p-mean flow representation for the fast QSA ODE is described in Thm. 2.1; it is one of several steps required in establishing convergence of $\boldsymbol{\Theta}$ and the estimation error bounds in Thm. 2.2.

The terms in the representation are defined based on $\hat{f}$ and Υ. The functions Υ and $\hat{f}$ are themselves treated as forcing functions in Poisson's equation, with solutions denoted $\hat{\hat{f}}$ and $\widehat{\Upsilon}$, respectively.

Theorem 2.1 (Perturbative Mean Flow): Suppose that (A0)-(A2) hold. Then, the p-mean flow representation (8) holds with $\bar{\Upsilon}^{\mathsf{ff}} \equiv 0$, $\mathcal{W}^2_t := \hat{\hat{h}}_t$,

$$\mathcal{W}^0_t := -[\mathcal{D}^h \widehat{\Upsilon}^{\mathsf{ff}}]_t + \frac{a_t}{\beta^2}\Big[r_t [\mathcal{D}^g \hat{\hat{h}}]_t - \Upsilon^{\mathsf{sf}}_t - \beta [\mathcal{D}^g \widehat{\Upsilon}^{\mathsf{ff}}]_t \Big] \tag{15a}$$

$$\mathcal{W}^1_t := -[\mathcal{D}^h \hat{\hat{h}}]_t + \widehat{\Upsilon}^{\mathsf{ff}}_t - \frac{a_t}{\beta}[\mathcal{D}^g \hat{\hat{h}}]_t \tag{15b}$$

where $r_t := \rho/(1+t)$.

The rate at which $\boldsymbol{\Theta}$ converges is the same as what would be expected from single timescale QSA ODEs with vanishing gain. Similarly, the estimation error bounds for $\boldsymbol{\Lambda}$ that follow are identical to what is obtained for single timescale QSA with constant gain, as well as in the averaging literature [23], [11].

Theorem 2.2 (Convergence): Suppose (A0)–(A6) hold. Then, there exist finite constants $\beta^0, b^{2.2}$ such that for any $0 < \beta \le \beta^0$ there is $\theta^\beta \in \mathbb{R}^d$ satisfying $\|\theta^* - \theta^\beta\| \le b^{2.2}\beta$, and

(i) $\|\Theta_t - \theta^\beta\| \le b^{2.2} a_t$ for $t \ge 0$.

(ii) $\limsup_{t\to\infty} \|\Lambda_t - \uplambda^*(\theta^*)\| \le b^{2.2}\beta$.

We next introduce a second order filter to reduce estimation error,

$$\tfrac{d^2}{dt^2}\Lambda^{\mathsf{F}}_t + 2\gamma\zeta \tfrac{d}{dt}\Lambda^{\mathsf{F}}_t + \gamma^2 \Lambda^{\mathsf{F}}_t = \gamma^2 \Lambda_t\,, \tag{16}$$

which is used in the slow dynamics,

$$\tfrac{d}{dt}\Theta_t = a_t g(\Theta_t, \Lambda^{\mathsf{F}}_t, \xi_t) \tag{17}$$

We impose the constraint on the natural frequency, $\gamma = O(\beta)$.

Theorem 2.3 (Error Attenuation): Suppose that the second-order filter is chosen subject to the following constraints: the damping ratio $\zeta \in (0,1)$ is independent of β, and a constant $\eta > 0$ is also fixed to define the natural frequency, $\gamma = \eta\beta$ for each β. If in addition (A0)–(A6) hold, then, there exists β^0 such that for any $0 < \beta \le \beta^0$ there is $\theta^\beta \in \mathbb{R}^d$ satisfying $\|\theta^* - \theta^\beta\| \le b^{2.3}\beta^2$, and

(i) $\|\Theta_t - \theta^\beta\| \le b^{2.3} a_t$ for $t \ge 0$.

(ii) $\limsup_{t\to\infty} \|\Lambda^{\mathsf{F}}_t - \uplambda^*(\theta^*)\| \le b^{2.3}\beta^2$.

It is conjectured that averaging techniques similar to the ones employed in [27], [19] can be applied to $\boldsymbol{\Theta}$, yielding convergence rates of order $O(a_t^2)$, compared to $O(a_t)$ in Thms. 2.2 and 2.3.

C. Lyapunov exponents and Poisson's equation

A key step in establishing convergence of $\boldsymbol{\Theta}$ consists of justifying the interpretation of (3a) as an instance of single timescale QSA with mean vector field (6). This requires a different formulation of Poisson's equation.

Consider for each $\theta \in \mathbb{R}^d$ the ODE (5). The joint process $\{\Psi^\theta_t = (\Lambda^\theta_t, \Phi_t) : t \ge 0\}$ is the state process of a dynamical system (hence a Markov process). Its Lyapunov exponent is defined as follows,

$$\mathcal{L}_\Lambda := \lim_{t\to\infty} \frac{1}{t}\log(\|\mathcal{S}_t\|) \tag{18a}$$

in which the *sensitivity process* is defined by

$$\mathcal{S}_t := \frac{\partial}{\partial \Lambda^\theta_0}\Lambda^\theta_t \tag{18b}$$

Thm. 2.4 establishes that the Lyapunov exponent is negative, which bring two crucial consequences for the system (5): existence and uniqueness of an invariant measure $\upmu_\theta$ for $\boldsymbol{\Psi}^\theta$, and solutions to Poisson's equation for functions of $\boldsymbol{\Psi}^\theta$. Similar conclusions are obtained in [20] for single timescale QSA with constant gain.

Let $u\colon \mathsf{Y} \to \mathbb{R}$ be a Lipschitz continuous function. Poisson's equation addressed in following takes the following form: on denoting $\bar{u}_0(\theta) := \int u(\theta, \lambda, z)\, \upmu_\theta(d\lambda, dz)$ and $\tilde{u}_0(x,z) := u(x,z) - \bar{u}_0(\theta)$ for any $(x,z) \in \mathsf{Y}$, the solution to Poisson's equation $\hat{u}_0$ satisfies the defining identity,

$$\int_0^T \tilde{u}_0(\theta, \Psi^\theta_t)\, dt = \hat{u}_0(\theta, \Psi^\theta_0) - \hat{u}_0(\theta, \Psi^\theta_T)\,, \quad T \ge 0\,. \tag{19}$$

Once again the solution is assumed normalized, with $\int \hat{u}_0(\theta, \lambda, z)\, \upmu_\theta(d\lambda, dz) = 0$ for each θ.

Theorem 2.4: (i) If (A0)–(A6) hold then there is $\beta_0 > 0$ and a continuous function $B\colon \mathsf{Y} \to \mathbb{R}_+$ such that the Lyapunov exponent is negative for $0 < \beta \le \beta_0$. Moreover, there is a constant $\delta > 0$ such that

$$\|\Lambda^\theta_t - \Lambda^{\theta,0}_t\| \le B(\theta, \lambda, z)\exp(-\delta\beta t) \tag{20}$$

where $(\theta, \lambda, z) \in \mathsf{Y}$ is the initial condition, and $\Lambda^{\theta,0}_t$ is the solution to (5) with initial condition $(\theta, 0, z)$.

(ii) Suppose that (A0)–(A3) hold, along with (20), and the the dynamical system (5) is ultimately bounded for any θ. Then there is a unique invariant measure for $\boldsymbol{\Psi}^\theta$ with compact support. Moreover, there is a solution $\hat{u}_0$ to (19) that is zero mean and locally Lipschitz continuous, whenever u is locally Lipschitz continuous.

Proof: Part (i) follows from arguments in [20, §2].

Parts of (ii) are also based on analysis from [20]: subject to (20), for each $\theta \in \mathbb{R}^d$ we can apply a technique known as *coupling from the past* to construct a stationary realization $\{(\Lambda_t^{\theta,\infty}, \Phi_t^\infty) : -\infty < t < \infty\}$. Its common marginal is $\upmu_\theta$.

Solutions to Poisson's equation are obtained by examining the construction of this stationary realization: $\Lambda_t^{\theta,\infty} = \phi_\infty^\theta(\Phi_t^\infty)$ for each t, for a smooth function ϕ_∞^θ. For any initial condition $\Phi_0 = z$ the process $\Lambda_t^{\theta,z} = \phi_\infty^\theta(\Phi_t)$ is a solution to (5), with initial condition $\Lambda_0^{\theta,z} = \phi_\infty^\theta(z) \in \mathbb{R}^d$.

The construction of $\hat{u}_0$ now proceeds in two steps: first consider, for each $\Psi_0^\theta = (\lambda, z)$,

$$\hat{u}_1(\theta, \lambda, z) = \int_0^\infty \tilde{u}_1(\theta, \Psi_t^\theta)\, dt\,, \qquad \tilde{u}_1 = \tilde{u}_0 - \tilde{u}_2$$
$$\textit{with} \quad \tilde{u}_2(\Phi_t) = \tilde{u}_0(\theta, \phi_\infty^\theta(\Phi_t), \Phi_t)$$

This satisfies Poisson's equation with forcing function $\tilde{u}_1$: $\frac{d}{dt}\hat{u}_1(\theta, \Psi_t^\theta) = -\tilde{u}_1(\theta, \Psi_t^\theta)$ for all $t \geq 0$.

Next, let $\hat{u}_2(\theta, z)$ denote the solution to Poisson's equation in the form (12) with forcing function $\tilde{u}_2$. The desired solution to (19) is then $\hat{u}_0 = \hat{u}_1 - \hat{u}_2$. ■

Upon obtaining existence of solutions $\hat{u}_0$ to (19), analysis proceeds as follows:

(i) By construction we have for each $\theta \in \mathbb{R}^d$,

$$\tfrac{d}{dt}\hat{g}_0(\theta, \Lambda_t^\theta, \Phi_t) = -\tilde{g}_0(\theta, \Lambda_t^\theta, \xi_t) = \varepsilon_t^a + \varepsilon_t^b$$

in which $\varepsilon_t^a := \beta\partial_\lambda \hat{g}_0(\theta, \Lambda_t^\theta, \Phi_t) \cdot h(\theta, \Lambda_t^\theta, \xi_t)$ and $\varepsilon_t^b := \partial_\Phi \hat{g}_0(\theta, \Lambda_t^\theta, \Phi_t) \cdot W\Phi_t$.

(ii) On replacing θ with Θ_t we arrive at a similar expression:

$$\begin{aligned} \tfrac{d}{dt}\hat{g}_0(\Theta_t, \Lambda_t, \Phi_t) &= a_t \partial_\theta \hat{g}_0(\Theta_t, \Lambda_t, \Phi_t) \cdot g_t + \varepsilon_t^a + \varepsilon_t^b \\ &= -\tilde{g}_0(\Theta_t, \Lambda_t, \xi_t) - a_t \Upsilon_t^0 \end{aligned}$$

in which $\Upsilon_t^0 := \partial_\theta \hat{g}_0(\Theta_t, \Lambda_t, \Phi_t) \cdot g(\Theta_t, \Lambda_t, \xi_t)$.

(iii) Step (ii) motivates the representation of (3a) as the single timescale QSA ODE (7) with $\widetilde{\Xi}_t^0 = -\tilde{g}_0(\Theta_t, \Lambda_t, \xi_t) = -\frac{d}{dt}\hat{g}_0(\Theta_t, \Lambda_t, \Phi_t) - a_t \Upsilon_t^0$.

III. Examples

A. Q learning

Consider the infinite horizon optimal control problem

$$J^*(x) = \min \int_0^\infty c(\mathcal{X}_t, \mathcal{U}_t)\, dt\,, \quad \tfrac{d}{dt}\mathcal{X}_t = F(\mathcal{X}_t, \mathcal{U}_t)$$

in which the the state $\mathcal{X}_t$ and state $\mathcal{U}_t$ evolve on $\mathbb{R}^n$, $\mathbb{R}^m$ respectively, $c\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}_+$ is continuous, and the minimum is over all continuous inputs. Provided the minimum exists to define a smooth function J^*, we define $Q^*(x, u) = c(x, u) + \partial_x J^*(x) \cdot F(x, u)$; one version of the *Q-function*. The HJB equation tells us that the optimal policy is state feedback, $\phi^*(x) = \arg\min_u Q^*(x, u)$.

In Q-learning we have a parameterized family of approximations $\{Q^\theta : \theta \in \mathbb{R}^d\}$, and an algorithm is devised to obtain θ^* so that Q^{θ^*} is close to Q^*. Each parameter defines a policy $\phi^\theta \in \arg\min_u Q^\theta(x, u)$, and one can expect that ϕ^{θ^*} will approximate ϕ^* if the Q-function approximation is sufficiently tight.

There are many Q-learning algorithms available in the deterministic, continuous time setting of this paper. In most cases these may be cast as (3) in which $\Lambda_t = \mathcal{X}_t$ plays the role of the fast variable, with β emerging via a time-transformation. Provided that the assumptions of Thm. 2.4 are satisfied, we obtain convergence of the algorithm, along with ergodicity of the state process $\mathcal{X}$.

The theory in this paper does not require restrictive assumptions on the input used for training. Indeed, the bulk of the theory of Q-learning is based on policies for training that are *oblivious*, which in this deterministic setting means that $\mathcal{U}_t = \xi_t$ for all t, in which the exploration signal does not depend upon Θ_t (the estimate of θ^* at time t). The theory in this paper allows for far more efficient inputs for training, such as the *epsilon-greedy* input $\mathcal{U}_t = \phi^{\Theta_t}(\mathcal{X}_t) + \varepsilon\xi_t$.

B. Linear model

A general model takes the form

$$f(X_t) = [\bar{A} + A_t^\circ]X_t + B_t \tag{21}$$

in which the $(2d) \times (2d)$ matrix valued process $\{A_t^\circ\}$ has zero mean. Letting $b \in \mathbb{R}^{2d}$ denote the mean of $\{B_t\}$, and assuming $\bar{A}$ is invertible, we have $(\theta^*; \lambda^*) = -\bar{A}^{-1}b$. We take $b = 0$ without loss of generality throughout the remainder of this section.

A single time-scale QSA algorithm is appropriate if $\bar{A}$ is Hurwitz. If not, consider the decomposition into four $d \times d$ blocks: in Matlab notation, $\bar{A} = [\bar{A}^{\mathsf{s}}, \bar{A}^{\mathsf{sf}}; \bar{A}^{\mathsf{fs}}, \bar{A}^{\mathsf{f}}]$. We then have $\lambda^*(\theta) = -[\bar{A}^{\mathsf{f}}]^{-1}\bar{A}^{\mathsf{fs}}\theta$, and (4a) becomes the d-dimensional ODE, $\frac{d}{dt}\vartheta = A^\bullet\vartheta$, with $A^\bullet = \bar{A}^{\mathsf{s}} - \bar{A}^{\mathsf{sf}}[\bar{A}^{\mathsf{f}}]^{-1}\bar{A}^{\mathsf{fs}}$. Success of the two timescale algorithm requires that each of the two matrices $A^\bullet$, $\bar{A}^{\mathsf{f}}$ be Hurwitz.

The vector field (21) differs from the linear model of [17] because there is multiplicative noise, which significantly complicates analysis: see discussion in [23].

Consider for example the mixed-gain QSA ODE (10) with $\bar{A} = [\alpha, \alpha; -2, -1]$, $B_t = [\sin(\omega_1 t), \sin(\omega_2 t)]^\intercal$ and $A_t = IB_t$.

Hence, $\bar{f}(x) = \bar{A}x$ and $x^* = 0$. The matrix $\bar{A}$ is Hurwitz only for $0 < \alpha < 1$. When $\alpha > 1$, the benefits of a two timescale algorithm for stabilization become clear: $\lambda^*(\theta) = -2\theta$, giving stable dynamics for the mean flow:

$$\tfrac{d}{dt}\vartheta_t = \bar{g}(\vartheta_t, \lambda^*(\vartheta_t)) = -\alpha\vartheta_t$$

C. Extremum-seeking control

This approach to gradient-free optimization begins with the construction of approximate gradients $\{\widetilde{\nabla}_t\Gamma : t \geq 0\}$ based on perturbed observations of the form $\mathcal{Y}_t := \mathcal{Y}(\Theta_t, \xi_t) = \Gamma(\Theta_t + \epsilon_t\xi_t)$, in which $\epsilon_t \equiv \epsilon(\Theta_t)$ is known as the *probing* *gain*. Two possibilities result in a Lischitz QSA algorithm provided $\nabla\Gamma$ is Lipschitz continuous:

$$\begin{aligned}\epsilon(\theta) &= \varepsilon\sqrt{1+\Gamma(\theta)} \\ \epsilon(\theta) &= \varepsilon\sqrt{1+\|\theta-\theta^{\mathsf{ctr}}\|^2/\sigma_p^2}\end{aligned} \tag{22}$$

where in the first choice it is assumed without loss of generality that Γ takes on non-negative values, θ^{ctr} is an a-priori estimate of θ^{opt} and σ_p plays the role of the standard deviation around this prior.

The observations are normalized

$$\mathcal{Y}^{\mathsf{n}}_t := \mathcal{Y}^{\mathsf{n}}(\Theta_t, \xi_t) = \frac{1}{\epsilon_t}\Gamma(\Theta_t + \epsilon_t\xi_t) \tag{23}$$

To obtain $\{\Theta_t\}$, the probing signal and $\{\mathcal{Y}^{\mathsf{n}}_t\}$ are fed as input to a sequence of filters [1].

The first is a high-pass filter with state process Λ:

$$\begin{aligned}\tfrac{d}{dt}\Lambda_t &= \mathrm{F}\Lambda_t + \mathrm{G}u_t \\ y_t &= \mathrm{H}^{\intercal}\Lambda_t + \mathrm{J}u_t\end{aligned} \tag{24}$$

with $(\mathrm{F},\mathrm{G},\mathrm{H},\mathrm{J})$ of compatible dimension. In this equation, u_t is the scalar input and y_t the scalar output.

The output of (24) is expressed $y_t = [\mathrm{M}u]_t$ with transfer function $\mathrm{M}(s) := \mathrm{H}^{\intercal}(Is-\mathrm{F})^{-1}\mathrm{G}+\mathrm{J}$. We define $\check{\mathcal{Y}}^{\mathsf{n}}_t = [\mathrm{M}\mathcal{Y}^{\mathsf{n}}]_t$ and $\check{\xi}^i_t = [\mathrm{M}\xi^i]_t$ for $1 \le i \le m$, where ξ^i_t denotes the i^{th} component of the probing signal.

The $m+1$ outputs of (24) are then fed into a low pass filter with state process Θ:

$$\tfrac{d}{dt}\Theta_t = -a_t[\sigma(\Theta_t - \theta^{\mathsf{ctr}}) + a_t\check{\xi}_t\check{\mathcal{Y}}^{\mathsf{n}}_t] \tag{25}$$

with θ^{ctr} as defined in (22).

We arrive at a two timescale QSA ODE of the form (10):

$$\begin{aligned}f(X_t,\xi^\circ_t) &= \begin{bmatrix}-\sigma I & -\check{\xi}_t\mathrm{H}^{\intercal} \\ 0 & \mathrm{F}\end{bmatrix}X_t + \begin{bmatrix}-\mathrm{J}\check{\xi}_t \\ \mathrm{G}\end{bmatrix}\mathcal{Y}^{\mathsf{n}}_t \\ \bar{f}(x) &= \begin{bmatrix}-\sigma I & 0 \\ 0 & \mathrm{F}\end{bmatrix}x + \begin{bmatrix}-\mathrm{J}\,\mathsf{E}[\check{\xi}\mathcal{Y}^{\mathsf{n}}(\theta,\xi)] \\ \mathrm{G}\,\mathsf{E}[\mathcal{Y}^{\mathsf{n}}(\theta,\xi)]\end{bmatrix}\end{aligned} \tag{26}$$

in which $\xi^\circ_t := (\xi_t, \check{\xi}_t)$ is a $2m$-dimensional probing signal and Λ_t is state process in (24) with input $u_t = \mathcal{Y}^{\mathsf{n}}_t$. The expectations in (26) are taken over Ω, upon recalling that $\xi_t = G(\Phi_t)$.

From (24) and (26), we conclude that $\beta = 1$ always. Moreover, the vector field $\bar{g}_0$ associated with (6) can be identified:

$$\bar{g}_0(\theta) = -(\sigma I\theta + \mathsf{E}[\check{\xi}\, H^{\intercal}\Lambda^\theta + J\check{\xi}\mathcal{Y}^{\mathsf{n}}(\theta,\xi)]) \tag{27}$$

in which $\xi_t = G(\Phi_t)$ and the expectation is taken in steady state and with respect to the measure $\upmu_\theta$ for $\Psi^\theta = (\Lambda^\theta, \Phi)$. Once we establish the conditions of Thm. 2.2, we conclude convergence: $\|\Theta_t - \theta^1\| = O(a_t)$, in which $\bar{g}_0(\theta^1) = 0$ and $\|\theta^1 - \theta^*\| = O(1)$.

It is not difficult to establish that the Lyapunov exponent is negative: for each $\theta \in \mathbb{R}^d$, the ODE (5) is linear with additive disturbance:

$$h(\theta, \Lambda^\theta_t, \xi_t) = \mathrm{F}\Lambda^\theta_t + \mathrm{G}\mathcal{Y}^{\mathsf{n}}(\theta,\xi_t) \tag{28}$$

The sensitivity process (18b) solves $\frac{d}{dt}\mathcal{S}_t = \mathrm{F}\mathcal{S}_t$ with $\mathcal{S}_0 = I$. The Lyapunov exponent (18a) can be identified $\mathcal{L}_\Lambda = \mathrm{Re}\,(\lambda_1)$, where λ_1 is an eigenvalue of F with maximal real part. Provided F is Hurwitz, $\mathrm{Re}\,(\lambda_1) < 0$ and Thm. 2.4 can be used to conclude that $\boldsymbol{\Psi}^\theta$ admits an unique invariant measure for each θ as well as existence of solutions to (19).

It remains to show that the mean flow with vector field $\bar{g}_0$ is globally asymptotically stable. This requires additional assumptions on the objective. Here we assume that $\nabla\Gamma$ is Lipschitz continuous and that its norm is coercive, so that the ODE $\frac{d}{dt}x = \nabla\Gamma(x)$ is ultimately bounded. The same conclusion holds for $\frac{d}{dt}\vartheta = \bar{g}_0(\vartheta)$ for sufficiently small $\varepsilon > 0$, and conditions on the filter. However, this is only possible when using a state-dependent probing gain.

Analysis of the mean flow begins with a Taylor series approximation of $\mathcal{Y}^{\mathsf{n}}(\theta,\xi)$ around θ:

$$\mathcal{Y}^{\mathsf{n}}(\theta,\xi_t) = \frac{1}{\epsilon(\theta)}\Gamma(\theta) + \xi_t\nabla\Gamma(\theta) + O(\epsilon) \tag{29}$$

which applied to (27) gives, with $M_0 = J\mathsf{E}[\check{\xi}\xi]$,

$$\bar{g}_0(\theta) = -(\sigma I\theta + \mathsf{E}[\check{\xi}\, H^{\intercal}\Lambda^\theta] + M_0\nabla\Gamma(\theta) + O(\epsilon)) \tag{30}$$

It remains to obtain a representation for $\mathsf{E}[\check{\xi}\, H^{\intercal}\Lambda^\theta]$. In view of (24), we have that for each θ,

$$\mathrm{H}^{\intercal}\Lambda^\theta_t = \int_{-\infty}^t e^{\mathrm{F}(t-\tau)}\mathrm{G}\mathcal{Y}^{\mathsf{n}}(\theta,\xi_\tau)\,d\tau = \gamma_0\frac{\Gamma(\theta)}{\epsilon(\theta)} + \check{\xi}^{\intercal}_t\nabla\Gamma(\theta)$$

where the last inequality follows from substitution of (29) and $\gamma_0 = -\mathrm{H}^{\intercal}F^{-1}G$ denotes the DC gain of (24). This implies $\mathsf{E}[\check{\xi}\, H^{\intercal}\Lambda^\theta] = \Sigma_{\check{\xi}}\nabla\Gamma(\theta)$ with $\Sigma_{\check{\xi}} := \mathsf{E}[\check{\xi}\check{\xi}^{\intercal}]$.

Together with (30), we obtain

$$\bar{g}_0(\theta) = -(\sigma I\theta + M\nabla\Gamma(\theta) + O(\epsilon))\,, \quad M = \Sigma_{\check{\xi}} + M_0 \tag{31}$$

Under passivity of (24) we have $M + M^{\intercal} > 0$. Consequently, the mean flow with vector field (27) is ultimately bounded for sufficiently small $\varepsilon > 0$ in either of the choices of exploration gain (22). Heuristic arguments in [23] lead to the same approximation as in (31) and the same stability conclusion provided the high-pass filter (24) is passive.

IV. Conclusions

This paper extends QSA theory to algorithms with two timescales in the mixed gain setting. Estimation error bounds for the general algorithm were obtained along with justification for its interpretation as a single timescale algorithm.

There are several open paths for research:

△ This paper concerns static root finding problems of the form $\bar{f}(x^*) = 0$. It would be exciting to investigate extensions of the p-mean flow to tracking problems, that is, when $\bar{f}$ is a function of time so that the root is time-varying $\{x^*_t\}$.

△ Establishing rates of convergence for two timescale QSA with vanishing gain is still an open problem. We conjecture that it is simple to extend Thm. 2.2 to this case, but can we achieve the $O(b_t^4)$ MSE rates in [19] for two timescales?

△ The implications of this work to online optimization and control will be considered in future research, following [9], [10].

References

[1] K. B. Ariyur and M. Krstić. *Real Time Optimization by Extremum Seeking Control*. John Wiley & Sons, Inc., New York, NY, 2003.

[2] S. Bhatnagar, M. C. Fu, S. I. Marcus, and S. Bhatnagar. Two-timescale algorithms for simulation optimization of hidden markov models. *Iie Transactions*, 33(3):245–258, 2001.

[3] S. Bhatnagar, M. C. Fu, S. I. Marcus, and P. J. Fard. Optimal structured feedback policies for abr flow control using two-timescale spsa. *IEEE/ACM Transactions on Networking*, 9(4):479–491, 2001.

[4] S. Bhatnagar, M. C. Fu, S. I. Marcus, and I.-J. Wang. Two-timescale simultaneous perturbation stochastic approximation using deterministic perturbation sequences. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 13(2):180–209, 2003.

[5] 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*, pages 1–50, 2021.

[6] V. S. Borkar. Stochastic approximation with two time scales. *Systems Control Lett.*, 29(5):291–294, 1997.

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

[8] 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.

[9] M. Colombino, E. Dall'Anese, and A. Bernstein. Online optimization as a feedback controller: Stability and tracking. *Trans. on Control of Network Systems*, 7(1):422–432, 2020.

[10] A. Hauswirth, S. Bolognani, G. Hug, and F. Dörfler. Optimization algorithms as robust feedback controllers. *arXiv:2103.11329*, 2021.

[11] H. K. Khalil. *Nonlinear systems*. Prentice-Hall, Upper Saddle River, NJ, 3rd edition, 2002.

[12] J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *Ann. Math. Statist.*, September 1952.

[13] P. Kokotović, H. K. Khalil, and J. O'Reilly. *Singular Perturbation Methods in Control: Analysis and Design*. Society for Industrial and Applied Mathematics, 1999.

[14] P. Kokotovic, R. O'Malley, and P. Sannuti. Singular perturbations and order reduction in control theory — an overview. *Automatica*, 12(2):123–132, 1976.

[15] V. R. Konda and V. S. Borkar. Actor-critic–type learning algorithms for Markov decision processes. *SIAM Journal on control and Optimization*, 38(1):94–123, 1999.

[16] V. R. Konda and J. N. Tsitsiklis. On actor-critic algorithms. *SIAM J. Control Optim.*, 42(4):1143–1166 (electronic), 2003.

[17] V. R. Konda and J. N. Tsitsiklis. Convergence rate of linear two-time-scale stochastic approximation. *Ann. Appl. Probab.*, 2004.

[18] C. Lakshminarayanan and S. Bhatnagar. A stability criterion for two timescale stochastic approximation schemes. *Automatica*, 2017.

[19] C. K. 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*, volume 35, pages 15743–15756. Curran Associates, Inc., 2022.

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

[21] C. K. Lauand and S. Meyn. The curse of memory in stochastic approximation. In *Proc. IEEE Conference on Decision and Control*, pages 7803–7809, 2023.

[22] C. K. Lauand and S. Meyn. Markovian foundations for quasi-stochastic approximation in two timescales: ext. version. *arXiv 2409.07842*, 2024.

[23] C. K. Lauand and S. Meyn. Quasi-stochastic approximation: Design principles with applications to extremum seeking control. *IEEE Control Systems Magazine*, 43(5):111–136, Oct 2023.

[24] S. Liu and M. Krstic. Introduction to extremum seeking. In *Stochastic Averaging and Stochastic Extremum Seeking*, Communications and Control Engineering. Springer, London, 2012.

[25] M. Métivier and P. Priouret. Applications of a Kushner and Clark lemma to general classes of stochastic algorithms. *Trans. on Information Theory*, 30(2):140–151, March 1984.

[26] 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.

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

[28] A. Mokkadem and M. Pelletier. Convergence rate and averaging of nonlinear two-time-scale stochastic approximation algorithms. *Ann. Appl. Probab.*, 16:1671–1702, 2006.

[29] 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*, pages 2947–2997, 2020.

[30] E. Moulines and F. R. Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems 24*, pages 451–459, 2011.

[31] A. Orvieto, H. Kersting, F. Proske, F. Bach, and A. Lucchi. Anticorrelated noise injection for improved generalization. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proc. Intl. Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*. PMLR, 17–23 Jul 2022.

[32] J. A. Sanders, F. Verhulst, and J. Murdock. *Averaging methods in nonlinear dynamical systems*, volume 59. Springer, 2007.

[33] D. R. Smith. *Singular-perturbation theory: an introduction with applications*. Cambridge University Press, 1985.

[34] J. C. Spall. *Introduction to stochastic search and optimization: estimation, simulation, and control*. John Wiley & Sons, 2003.

[35] Y. Tan, W. H. Moase, C. Manzie, D. Nešić, and I. Mareels. Extremum seeking from 1922 to 2010. In *Proc. of the 29th Chinese control conference*, pages 14–26. IEEE, 2010.

[36] J. Tsitsiklis. Asynchronous stochastic approximation and Q-learning. *Machine Learning*, 16:185–202, 1994.

[37] B. Van Roy. *Learning and Value Function Approximation in Complex Decision Processes*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, 1998. AAI0599623.

Appendix

P-mean flow The ODE (10) can be expressed in terms of its mean vector field:

$$\tfrac{d}{dt}\Theta_t = \mathrm{g}_t[\bar{f}(X_t) + \widetilde{\Xi}_t]\,, \quad \widetilde{\Xi}_t := f(X_t, \xi) - \bar{f}(X_t) \tag{32}$$

where $\widetilde{\Xi}_t = \begin{bmatrix} \widetilde{\Xi}^\theta_t \\ \widetilde{\Xi}^\lambda_t \end{bmatrix} = \begin{bmatrix} g(X_t, \xi) - \bar{g}(X_t) \\ h(X_t, \xi) - \bar{h}(X_t) \end{bmatrix}$.

The p-mean flow representation (2.1) follows from the next four lemmas.

Lemma 1.1: Suppose that for each $(x, \Phi_0) \in \mathsf{Y}$ and $t \ge 0$,

$$\tfrac{d}{dt}\widehat{F}(x, \Phi_t) := -\widetilde{F}(x, \xi_t) = -F(x, \xi_t) + \bar{F}(x)$$

where $\widehat{F} \colon \mathsf{Y} \to \mathbb{R}^d$ is C^1.

Then, on writing $\widehat{F}_t = \widehat{F}(X_t, \Phi_t)$, $\widetilde{F}_t = \widetilde{F}(X_t, \xi_t)$,

$$\tfrac{d}{dt}\widehat{F}_t = -\widetilde{F}_t + a_t[\mathcal{D}^g \widehat{F}](X_t, \Phi_t) + \beta[\mathcal{D}^h \widehat{F}](X_t, \Phi_t)$$

Proof: This follows from the chain rule and the definition of directional derivatives in (13): $\tfrac{d}{dt}\widehat{F}(X_t, \Phi_t) = -\widetilde{F}(X_t, \xi_t) + a_t \partial_\theta \widehat{F}(X_t, \Phi_t) \cdot g(X_t, \xi_t) + \beta \partial_\lambda \widehat{F}(X_t, \Phi_t) \cdot h(X_t, \xi_t)$. ■

Lemma 1.2: Under (A0)-(A2), $\widetilde{\Xi}_t$ admits the representation

$$\widetilde{\Xi}_t = -\tfrac{d}{dt}\hat{f}_t - \begin{bmatrix} a_t \Upsilon^{\mathsf{ss}}_t + \beta \Upsilon^{\mathsf{fs}}_t \\ a_t \Upsilon^{\mathsf{sf}}_t + \beta \Upsilon^{\mathsf{ff}}_t \end{bmatrix} \tag{33}$$

in which $\hat{f} \colon \mathsf{Y} \to \mathbb{R}^{2d}$ denotes the solution to Poisson's equation (12) with forcing function f.

Proof: Differentiating $\hat{f}_t$ with respect to time, we obtain from Lemma 1.1, $\widetilde{\Xi}_t = -\tfrac{d}{dt}\hat{f}(X_t, \Phi_t) + \widehat{A}(X_t, \Phi_t)\mathrm{g}_t f(X_t, \xi_t)$. The conclusion (33) then follows from the definitions in (10) and (14). ■

Lemma 1.3: If (A0)-(A2) hold, then with $r_t = \rho/(t+1)$,

$$\tfrac{d}{dt}\hat{h}_t = -r_t a_t[\mathcal{D}^g \hat{\hat{h}}]_t + a_t \tfrac{d}{dt}[\mathcal{D}^g \hat{\hat{h}}]_t + \beta \tfrac{d}{dt}[\mathcal{D}^h \hat{\hat{h}}]_t - \tfrac{d^2}{dt^2}\hat{\hat{h}}_t \tag{34}$$

Proof: Another application of Lemma 1.1 with $\widehat{F} = \hat{\hat{h}}$ gives $\frac{d}{dt}\hat{\hat{h}}_t = a_t[\mathcal{D}^g\hat{\hat{h}}]_t + \beta\frac{d}{dt}[\mathcal{D}^h\hat{\hat{h}}]_t - \hat{h}_t$. Differentiating both sides with respect to t once more yields (34):

$$\begin{aligned}\tfrac{d^2}{dt^2}\hat{\hat{h}}_t &= \tfrac{d}{dt}\{a_t[\mathcal{D}^g\hat{\hat{h}}]_t\} + \beta\tfrac{d}{dt}[\mathcal{D}^h\hat{\hat{h}}]_t - \tfrac{d}{dt}\hat{h}_t \\ &= -r_t a_t[\mathcal{D}^g\hat{\hat{h}}]_t + a_t\tfrac{d}{dt}[\mathcal{D}^g\hat{\hat{h}}]_t + \beta\tfrac{d}{dt}[\mathcal{D}^h\hat{\hat{h}}]_t - \tfrac{d}{dt}\hat{h}_t\end{aligned}$$

■

The final lemma is immediate from Lemma 1.1:

Lemma 1.4: Suppose that (A0)-(A2) hold. Then, $\Upsilon^{\mathsf{ff}}_t = \bar{\Upsilon}^{\mathsf{ff}}(X_t) + a_t[\mathcal{D}^g\widehat{\Upsilon}^{\mathsf{ff}}]_t + \beta[\mathcal{D}^h\widehat{\Upsilon}^{\mathsf{ff}}]_t - \frac{d}{dt}\widehat{\Upsilon}^{\mathsf{ff}}_t$. ■

Proof of Thm. 2.1. From (33) we obtain $\widetilde{\Xi}^\lambda_t = -\frac{d}{dt}\hat{h}_t - a_t\Upsilon^{\mathsf{sf}}_t - \beta\Upsilon^{\mathsf{ff}}_t$. The p-mean flow representation follows from substitution of the results in Lemmas 1.3 and 1.4. ■

Estimation error bounds Fix $\beta_0 > 0$ and let $0 < \beta \le \beta_0$ also be fixed. Upon defining a sequence of finite time intervals $\{T_n\}$ of length $T > 0$ in which $T_{n+1} - T_n := T/\beta$ for each n, we let $n_0 > 0$ denote the integer satisfying $a_{T_{n_0}} < \beta^2$.

Proof of Thm. 2.2 (Outline) The complete proof is contained in the extended version of this article [22]. The outline that follows consists of ten steps. In the **i**th step, we denote by $b^{\mathbf{i}}$ a positive constant, independent of $\beta \in (0, \beta^0]$.

1. Under (A4), there exists $n_\bullet$ depending upon X_0 such that $\|X_t\| \le b^\bullet$ for all $t \ge T_{n_\bullet}$. Denote $\mathcal{T}_\circ := \max\{T_{n_\bullet}, T_{n_0}\}$.

2. Globally exponentially asymptotically stable (GEAS) ODEs are robust to bounded disturbances. Suppose that $\dot{x}^\circ = \bar{f}(x^\circ)$ is GEAS. Consider the ODE with disturbance,

$$\tfrac{d}{dt}x_t = \beta[\bar{f}^\circ(x_t) + w_t]\,, \quad \|w_t\| \le b^w + \beta^0\|x_t\| \qquad (35)$$

with b^w a positive constant. Then, the bound $\|x_t\| \le \frac{1}{2}\|x_0\| + b^{\mathbf{2}}$ holds for all $t > 0$ sufficiently large. This follows from Lyapunov stability arguments in [11, Ch. 5].

3. Integrating (3a) from T_n to t, we have

$$\|\Theta_t - \Theta_{T_n}\| \le b^{\mathbf{3}}T\frac{a_{T_n}}{\beta}\,, \quad \mathcal{T}_\circ < T_n < t \le T_{n+1} \qquad (36)$$

implying that $\boldsymbol{\Theta}$ is "quasi-static" over finite time intervals for large n.

4. The *scaled error* is defined by $Z^n_t := \frac{1}{\beta}(Y_t - \lambda^*(\Theta_{T_n}))$, $T_n < t \le T_{n+1}$, in which $Y_t := \Lambda_t - \beta\hat{h}_t$. On differentiating Z^n_t with respect to t, performing a first order Taylor series approximation to the function h in (3b), using (36) and (33) we obtain whenever $T_n > \mathcal{T}_\circ$,

$$\begin{gathered}\tfrac{d}{dt}Z^n_t = \beta[\bar{A}^{\mathsf{f}}(\Theta_{T_n})Z^n_t + w^Z_t] \\ \text{with } \|w^Z_t\| \le b^{\mathbf{4}} + O(\beta^0\min\{\|Z^n_t\|, \|Z^n_t\|^2\})\end{gathered} \qquad (37)$$

5. The following bounds hold for all $t > T_n > \mathcal{T}_\circ$, upon choosing T large enough: $\|Z^n_{T_{n+1}}\| \le \frac{1}{2}\|Z^n_{T_n}\| + b^{\mathbf{5}}$,

$$\begin{gathered}\|Z^{n+1}_{T_{n+1}} - Z^n_{T_{n+1}}\| \le b^{\mathbf{5}}T^2\frac{a_{T_n}}{\beta^2}\,, \\ \|Z^n_t - Z^n_{T_n}\| \le b^{\mathbf{5}}(\|Z^n_{T_n}\| + 1)\,,\end{gathered} \qquad (38)$$

The first bound is an application of **2** with $\bar{f}^\circ(x) = \bar{A}^{\mathsf{f}}(\Theta_{T_n})x$ and $w_t = w^Z_t$. The second bound follows similarly to **3** and the last bound is an application of the Bellman-Grönwall inequality [27, Prop. 4.2].

6. The collection of bounds in step **5** imply that there exists $\mathcal{T}_\lambda > \mathcal{T}_0$ depending upon X_0 such that $\|\Lambda_t - \lambda^*(\Theta_t)\| \le b^{\mathbf{6}}\beta$ for all $t \ge \mathcal{T}_\lambda$.

7. In view of (32), step **6** and the assumed Lipschitz continuity of g, the slow QSA ODE is re-written for large t: for a process $\{\Delta^\lambda_t\}$ satisfying $\|\Delta^\lambda_t\| \le b^{\mathbf{7}}$,

$$\tfrac{d}{dt}\Theta_t = a_t[g(\Theta_t, \lambda^*(\Theta_t), \xi_t) + \Delta^\lambda_t\beta]\,, \quad t > \mathcal{T}_\lambda. \qquad (39)$$

8. Arguments in the proof of [27, Thm. 4.15] can be used to conclude that (39) implies, $\limsup_{t\to\infty}\|\Theta_t - \theta^*\| \le b^{\mathbf{8}}\beta$.

9. Following steps (i)–(iii) outlined after Thm. 2.4, (3a) can be approximated by the single timescale algorithm (7). Then, the proofs of [27, Thms. 4.15 & 4.24] can be extended to (7) to obtain $\|\Theta_t - \theta^\beta\| \le b^{\mathbf{9}}a_t$, in which $\bar{g}_0(\theta^\beta) = 0$. Together with step **8**, this gives $\|\theta^\beta - \theta^*\| \le b^{\mathbf{9}}\beta$ and establishes part (i) of Thm. 2.2.

10. Part (i) of Thm. 2.2 implies part (ii) through Lipschitz continuity of λ^*, the bound in step **6** and the triangle inequality.

■

Filtering *Proof of Thm. 2.3* The proof extends arguments in [23], beginning with the the linearization $\overline{h}(x) = \bar{A}^{\mathsf{f}}(\theta)[\lambda - \lambda^*(\theta)] + \mathcal{E}(x)$ for any $x = (\theta, \lambda)$, with $\|\mathcal{E}(x)\| \le L_A\min(\|x\|, \|x\|^2)$ for some $L_A < \infty$. Applying (8),

$$\tfrac{d}{dt}\Lambda_t = \beta\big[\bar{A}^{\mathsf{f}}(\Theta_t)[\Lambda_t - \lambda^*(\Theta_t)] + \mathcal{E}_t + \mathcal{W}_t\big]$$

where $\mathcal{E}_t = \mathcal{E}(X_t)$ and we have used $\bar{\Upsilon}^{\mathsf{ff}} \equiv 0$ under (A0). Consequently,

$$\tfrac{d}{dt}\Lambda^{\mathsf{F}}_t = \beta\big[\bar{A}^{\mathsf{f}}(\Theta_t)\widetilde{\Lambda}^{\mathsf{F}}_t + \varepsilon_t + \mathcal{E}^{\mathsf{F}}_t + \mathcal{W}^{\mathsf{F}}_t\big] \qquad (40)$$

in which $\widetilde{\Lambda}^{\mathsf{F}}_t = \Lambda^{\mathsf{F}}_t - \lambda^*(\Theta_t)$, and letting $\{\mathrm{m}_t\}$ denote the impulse response for the low pass filter (16),

$$\mathcal{E}^{\mathsf{F}}_t = \int_0^t \mathrm{m}_{t-\tau}\mathcal{E}_\tau\,d\tau\,, \quad \mathcal{W}^{\mathsf{F}}_t = \int_0^t \mathrm{m}_{t-\tau}\mathcal{W}_\tau\,d\tau\,,$$

The error term ε_t depends on X_0: it involves the transient response of the filter, and the vanishing term

$$\int_0^t \mathrm{m}_{t-\tau}[\bar{A}^{\mathsf{f}}(\Theta_t) - \bar{A}^{\mathsf{f}}(\Theta_\tau)]\widetilde{\Lambda}^{\mathsf{F}}_\tau\,d\tau$$

Consequently ε_t is vanishing as $t \to \infty$.

We also have

$$\limsup_{t\to\infty}\|\mathcal{E}^{\mathsf{F}}_t\| \le \limsup_{t\to\infty}\|\mathcal{E}_t\| = O(\beta^2) \qquad (41a)$$

$$\limsup_{t\to\infty}\|\mathcal{W}^{\mathsf{F}}_t\| = O(\beta^2) \qquad (41b)$$

The upper bound for $\|\mathcal{E}_t\|$ in (41a) follows from Thm. 2.2 and the bound $\|\mathcal{E}_t\| \le L_A\|X_t\|^2$.

The bandwidth constraint on the low pass filter implies the bound for $\|\mathcal{W}^{\mathsf{F}}_t\|$ in (41b), exactly as in [23].

The bounds (41) combined with (40) and **2** establish Thm. 2.3 (ii), and (i) easily follows as in the proof of Thm. 2.2.

■