Quasi-Stochastic Approximation and Off-Policy Reinforcement Learning

Preprint

Andrey Bernstein,[1] Yue Chen,[1] Marcello Colombino,[1]
Emiliano Dall'Anese,[2] Prashant Mehta,[3] and Sean Meyn[4]

[1] National Renewable Energy Laboratory
[2] University of Colorado Boulder
[3] University of Illinois at Urbana-Champaign
[4] University of Florida, Gainesville

Presented at the 2019 IEEE Conference on Decision and Control (IEEE CDC)
Nice, France
December 11–13, 2019

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Contract No. DE-AC36-08GO28308

Conference Paper
NREL/CP-5D00-73518
December 2019

Quasi-Stochastic Approximation and Off-Policy Reinforcement Learning

Preprint

Andrey Bernstein,[1] Yue Chen,[1] Marcello Colombino,[1] Emiliano Dall'Anese,[2] Prashant Mehta,[3] and Sean Meyn[4]

[1] National Renewable Energy Laboratory
[2] University of Colorado Boulder
[3] University of Illinois at Urbana-Champaign
[4] University of Florida, Gainesville

Suggested Citation
Bernstein, Andrey, Yue Chen, Marcello Colombino, Emiliano Dall'Anese, Prashant Mehta, and Sean Meyn. 2019. *Quasi-Stochastic Approximation and Off-Policy Reinforcement Learning: Preprint*. Golden, CO: National Renewable Energy Laboratory. NREL/CP-5D00-73518. https://www.nrel.gov/docs/fy20osti/73518.pdf.

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Contract No. DE-AC36-08GO28308

Conference Paper
NREL/CP-5D00-73518
December 2019

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

NOTICE

This work was authored in part by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. This work was supported in part by the Laboratory Directed Research and Development (LDRD) Program at NREL. The views expressed herein do not necessarily represent the views of the DOE or the U.S. Government.

Cover Photos by Dennis Schroeder: (clockwise, left to right) NREL 51934, NREL 45897, NREL 42160, NREL 45891, NREL 48097, NREL 46526.

NREL prints on paper that contains recycled content.

Quasi-Stochastic Approximation and Off-Policy Reinforcement Learning

Andrey Bernstein* Yue Chen* Marcello Colombino* Emiliano Dall'Anese‡ Prashant Mehta§ Sean Meyn¶

Abstract— **The Robbins-Monro stochastic approximation algorithm is a foundation of many algorithmic frameworks for reinforcement learning (RL), and often an efficient approach to solving (or approximating the solution to) complex optimal control problems. However, in many cases practitioners are unable to apply these techniques because of an inherent high variance. This paper aims to provide a general foundation for "quasi-stochastic approximation," in which all of the processes under consideration are deterministic, much like quasi-Monte-Carlo for variance reduction in simulation. The variance reduction can be substantial, subject to tuning of pertinent parameters in the algorithm. This paper introduces a new coupling argument to establish optimal rate of convergence provided the gain is sufficiently large. These results are established for linear models, and tested also in non-ideal settings.**

A major application of these general results is a new class of RL algorithms for deterministic state space models. In this setting, the main contribution is a class of algorithms for approximating the value function for a given policy, using a different policy designed to introduce exploration.

I. Introduction and Proposed Framework

Stochastic approximation concerns the root-finding problem $\overline{f}(\theta^*) = 0$, where $\theta^* \in \mathbb{R}^d$ is a parameter to be computed or approximated, and $\overline{f} : \mathbb{R}^d \to \mathbb{R}^d$ is defined using the following expectation

$$\overline{f}(\theta) := \mathsf{E}[f(\theta, \xi)] , \qquad \theta \in \mathbb{R}^d , \tag{1}$$

in which $f : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ and ξ is an m-dimensional random vector. With this problem in mind, the stochastic approximation (SA) method of Robbins and Monro [1], [2] involves recursive algorithms to estimate the parameter θ^*. The simplest algorithm is defined by the following recursion (n is the iteration index):

$$\theta_{n+1} = \theta_n + a_n f(\theta_n, \xi_n) , \qquad n \geq 0, \tag{2}$$

where $\boldsymbol{\xi} := \{\xi_n\}$ is an exogenous m-dimensional stochastic process, $a_n > 0$ is the step size, and $\theta_0 \in \mathbb{R}^d$ is given. For consistency with (1), it is assumed that the distribution of ξ_n converges to that of ξ as $n \to \infty$; e.g., $\boldsymbol{\xi}$ is an ergodic Markov process.

The motivation for the SA recursion (and also an important tool for convergence analysis) is the associated ordinary differential equation (ODE):

$$\tfrac{d}{du}\chi(u) = \overline{f}\,(\chi(u)) . \tag{3}$$

Under general assumptions, including boundedness of the stochastic recursion (2), the limit points of (2) are a subset of the stationary points of the ODE; that is, solutions to $\overline{f}(\theta^*) = 0$. See [2], [3] and the earlier monographs [4], [5].

The upshot of stochastic approximation is that it can be implemented without knowledge of the function f or of the distribution of ξ; rather, it can rely on observations of the sequence $\{f(\theta_n, \xi_n)\}$. This is one reason why these algorithms are valuable in the context of reinforcement learning (RL) [2], [6], [7], [8], [9]. In such cases, the driving noise is typically modeled as a Markov chain.

The present paper considers a *quasi-stochastic approximation* (QSA) algorithm, in which the "noise" is generated from a *deterministic* signal rather than a stochastic process. We opt for the continuous-time model:

$$\tfrac{d}{dt}\theta(t) = a(t) f(\theta(t), \xi(t)) . \tag{4}$$

The entries of the vector-valued process $\boldsymbol{\xi}$ may be chosen to be sums of sinusoids with irrationally related frequencies. The continuous time setting is adopted mainly for simplicity of exposition, especially for the convergence analysis; results can be extended to the discrete-time setting, but are omitted due to space constraints.

One motivation for the proposed framework was to provide foundations for the Q-learning algorithm introduced in [10], which treats nonlinear optimal control in continuous time. In [10] it was found in numerical experiments that the rate of convergence is superior to the ones of traditional applications of Q-learning. The present paper provides explanations for this fast convergence, and presents a methodology to design algorithms with optimal rate of convergence.

A. Contributions

Contributions of the present paper are explained in terms of theoretical advancements for the QSA and applications.

Authors are in alphabetical order.

*A.B., M.C., and Y.C. are with NREL in Golden Colorado (name.lastname@nrel.gov)

‡E.D. is the Department of ECEE at the University of Colorado Boulder (Emiliano.Dallanese@colorado.edu)

§P.G.M. is with the Department of MAE at the University of Illinois Urbana-Champaign (mehtapg@illinois.edu)

¶S.M. is with the Department of ECE at the University of Florida in Gainesville (meyn@ece.ufl.edu)

Acknowledgements: Financial support from NSF CMMI grant 146277, NSF CPS grant 1646229, and ARO grant W911NF1810334 is gratefully acknowledged. This work was authored in part by the National Renewable Energy Laboratory, managed and operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. This work was supported in part by the Laboratory Directed Research and Development (LDRD) Program at NREL. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

Analysis: As in the classical SA algorithm, analysis is based on consideration of the associated ODE (3) in which the "averaged" vector field is given by the ergodic average:

$$\overline{f}(\theta) = \lim_{T\to\infty} \frac{1}{T}\int_0^T f(\theta,\xi(t))\,dt, \quad \text{for all } \theta \in \mathbb{R}^d. \tag{5}$$

The paper will introduce pertinent assumptions in Section IV to ensure that the limit (5) exists, and that the averaged ODE (3) has a unique globally asymptotically stable stationary point θ^*. It will be shown that the QSA (4) converges to the same limit. Relative to convergence theory in the stochastic setting, new results concerning rates of convergence will be offered in Section IV.

The variance analysis outlined in Section IV begins by considering a linear setting $\bar{f}(\theta) = A(\theta - \theta^*)$, with A Hurwitz. The linearity assumption is typical in much of the literature on variance for stochastic approximation and is justified by constructing a linearized approximation for the original nonlinear algorithm [11], [5]. Rates of convergence of nonlinear QSA is beyond the scope of this paper and will be pursued in future work.

Under the assumption that $I + A$ is Hurwitz (that is, each eigenvalue λ of A satisfies $\text{Re}(\lambda) < -1$), it will be shown that the optimal rate of convergence of $1/t$ can be obtained. In particular, there is a constant $\overline{\sigma} < \infty$ such that the following holds for each initial condition $\theta(0)$:

$$\limsup_{t\to\infty} t\|\theta(t) - \theta^*\| \le \overline{\sigma} \tag{6}$$

This assumption is stronger than what is imposed to obtain the Central Limit Theorem for stochastic approximation, which requires $\text{Re}(\lambda) < -\frac{1}{2}$. On the other hand, the conclusions for stochastic approximation algorithms are weaker, where the above bound is replaced by

$$\limsup_{t\to\infty} t\mathsf{E}[\|\theta(t) - \theta^*\|^2] \le \overline{\sigma}^2 \tag{7}$$

That is, the rate is $1/\sqrt{t}$ rather than $1/t$ [4], [5].

The most compelling applications are: (i) gradient-free optimization methods, based on ideas from extremum-seeking control [12], [13]; and (ii) RL for deterministic control systems. Q-learning with function approximation is reviewed, following [10]. It is shown that the most straightforward application of RL does not satisfy the conditions of the paper, and in fact may not be stable. In view of these challenges, a new class of "off policy" RL algorithms are introduced. These algorithms have attractive numerical properties, and are suitable for application to approximate policy iteration.

B. Literature review

The first appearance of QSA methods appears to have originated in the domain of quasi-Monte Carlo methods applied to finance [14], [15]. Rates of convergence were obtained in [16], but with only partial proofs, and without the coupling bounds reported here.

Gradient-free optimization has been studied in several, seemingly disconnected lines of work. The Kiefer-Wolfowitz algorithm is the classical gradient-free optimization method that uses finite-difference approximation the gradient [17]. For a d-dimensional problem, it perturbs each dimension separately and requires $2d$ function evaluations. The simultaneous perturbation stochastic approximation (SPSA) algorithm uses random perturbations that are zero-mean independent variables [18], requiring two function evaluations at each update. Deterministic perturbations in SPSA are proposed in [19]. Another line of work, typically known as "bandit optimization" (see e.g., [20], [21], [22]) leverages a stochastic estimate of the gradient, based on a single or multiple evaluations of the objective function. Such algorithms have been analyzed extensively using tools similar to the classical SA approach, with similar conclusion on the high variance of the estimates [23]. In addition, the gradient-free technique termed "extremum-seeking control" (ESC) [12], [13] adopts sinusoidal signals as perturbations to estimate the gradient; it is a special application of the QSA theory developed in this paper. Stability of the classic ESC feedback scheme was analyzed in e.g., [24], [25].

The rate of convergence result (7) is an interpretation of classical results in the SA literature. Under mild conditions, the "limsup" can be replaced by a limit, and moreover the Central Limit Theorem holds for the scaled error process $\{\sqrt{t}[\theta(t) - \theta^*]\}$ [4], [5], [2]. In these works, the asymptotic covariance is the solution to a Lyapunov equation, derived from the linearized ODE and the noise covariance. The results in the QSA setting are different. It is shown in Theorem 4.3 that under the Hurwitz assumption on $I+A$, the scaled parameter estimates $\{t[\theta(t)-\theta^*]\}$ *couple* with another process, obtained by integrating the noise process. There is a large literature on techniques to minimize the asymptotic variance in stochastic approximation, including Ruppert-Polyak-Juditsky (RPJ) averaging [26], [27], or adaptive gain selection, resulting in the stochastic Newton-Raphson (SNR) algorithm [28], [5].

There is a large literature on techniques to minimize the asymptotic variance in stochastic approximation, including Ruppert-Polyak-Juditsky (RPJ) averaging [26], [27], or adaptive gain selection, resulting in the stochastic Newton-Raphson (SNR) algorithm [28], [5]. The problem of optimizing the rate for QSA (e.g., minimizing the bound $\overline{\sigma}$ in (6)) through choice of algorithm parameters is not trivial. This is because coupling occurs only when the eigenvalues of A satisfy $\text{Re}(\lambda) < -1$.

The fixed-policy Q-learning algorithm introduced here may be regarded as an *off policy* TD-learning algorithm (or SARSA) [29], [30]. The standard TD and SARSA algorithms are not well-suited to deterministic systems since the introduction of exploration creates bias. By definition, an off policy method allows an arbitrary stable input, which can be chosen to speed value function estimation. Q-learning also allows for exploration, but this is a nonlinear algorithm that often presents numerical challenges, and there is little theory to support this class of algorithms beyond special cases such as optimal stopping, or the complex "tabular" case for finite state-space models [29], [30]. In the special case of linear systems with quadratic cost, the off-policy TD learning algorithm introduced here reduces to [31].

Organization: The remainder of this paper is organized as follows. Sections II and III contain several general application areas for QSA, along with numerical examples. Stability and convergence theory is summarized in Section IV, with most technical proofs deferred to [32]. Conclusions and future directions for research are summarized in Section V.

II. Motivational Application Examples

To motivate the QSA theory, this section briefly discusses quasi Monte-Carlo and gradient-free optimization. A deeper look at applications to optimal control, which is the main focus of this paper, will be given in Section III.

A. *Quasi Monte-Carlo*

Consider the problem of obtaining the integral over the interval $[0,1]$ of a function $y\colon \mathbb{R} \to \mathbb{R}$. To fit the QSA model (4), let $\xi(t) := t$ (modulo 1), and set

$$f(\theta, \xi) := y(\xi) - \theta. \tag{8}$$

The averaged function is then given by

$$\overline{f}(\theta) = \lim_{T\to\infty} \frac{1}{T}\int_0^T f(\theta, \xi(t))\, dt = \int_0^1 y(t)\, dt - \theta$$

so that $\theta^* = \int_0^1 y(t)\, dt$. Algorithm (4) is given by:

$$\tfrac{d}{dt}\theta(t) = a(t)[y(\xi(t)) - \theta(t)]. \tag{9}$$

The numerical results that follow are based on the function $y(t) = e^{4t}\sin(100t)$. This exotic function was among many tested – it is used here only because the conclusions are particularly striking.

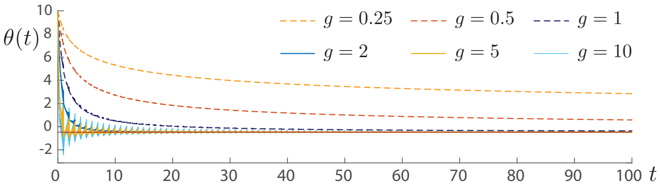

Fig. 1: Sample paths of Quasi Monte-Carlo estimates.

The differential equation was approximated using a standard Euler scheme with step-size 10^{-3}. Two algorithms are compared in the numerical results that follow: standard Monte-Carlo, and versions of the deterministic algorithm (9), differentiated by the gain $a(t) = g/(t+1)$. Fig. 1 shows typical sample paths of the resulting estimates for a range of gains; in each case the algorithm was initialized with $\theta(0) = 10$. The true mean is $\theta^* \approx -0.4841$.

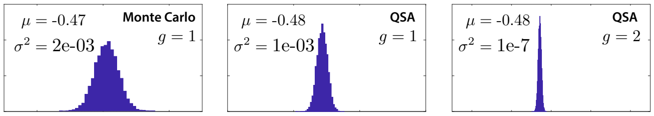

Fig. 2: Histograms of Monte-Carlo and Quasi Monte-Carlo estimates after 10^4 independent runs.

Independent trials were conducted to obtain variance estimates. In each of 10^4 independent runs, the common initial condition was drawn from $N(0, 10)$, and the estimate was collected at time $T = 100$. Fig. 2 shows three histograms of estimates for standard Monte-Carlo, and QSA using gains $g = 1$ and 2. An alert reader must wonder: *why is the variance reduced by 4 orders of magnitude when the gain is increased from* 1 *to* 2*?* The relative success of the high-gain algorithm is explained in Section IV.

B. *Gradient-Free Optimization*

Consider the unconstrained optimization problem:

$$\min_{\theta\in\mathbb{R}^d} J(\theta). \tag{10}$$

The goal is to minimize this function based on observations of $J(x(t))$, where the signal $\boldsymbol{x}$ is chosen by design. It is assumed that $J\colon \mathbb{R}^d \to \mathbb{R}$ is convex, twice continuously differentiable, and that it has a unique minimizer, denoted θ^*. Computation of the optimizer is thus equivalent to obtaining a zero of the gradient of J. The goal is to design QSA algorithms that seek solutions to the equation $\overline{f}(\theta^*) = 0$, where

$$\overline{f}(\theta) := H\nabla J(\theta)\,, \qquad \theta \in \mathbb{R}^d\,. \tag{11}$$

The choice of the invertible matrix H is part of the algorithm design.

We design the signal $\boldsymbol{x}$ as the sum of two terms $x(t) = \theta(t) + \varepsilon\xi(t)$, $t \ge 0$, where $\varepsilon > 0$ and $\xi_i(t) = \sqrt{2}\sin(\omega_i t)$, for $\omega_i \neq \omega_j$ for all $i \neq j$. It can be shown that this process satisfies:

$$\lim_{T\to\infty} \frac{1}{T}\int_{t=0}^T \xi(t)\, dt \;=\; 0 \tag{12}$$

$$\lim_{T\to\infty} \frac{1}{T}\int_{t=0}^T \xi(t)\xi(t)^{\intercal}\, dt \;=\; I \tag{13}$$

For a given $\theta \in \mathbb{R}^d$, consider then the second-order Taylor expansion of the objective function around θ:

$$\begin{aligned} J(\theta + \varepsilon\xi(t)) &= J(\theta) \\ &\quad + \varepsilon\xi(t)^{\intercal}\nabla J(\theta) + \frac{1}{2}\varepsilon^2\xi(t)^{\intercal}\nabla^2 J(\theta)\xi(t) + o(\varepsilon^2). \end{aligned}$$

Define $f(\theta, \xi) := -\xi J(\theta + \varepsilon\xi)$. It is easy to verify that under (12) and (13), one has that:

$$\overline{f}(\theta) := \lim_{T\to\infty} \frac{1}{T}\int_{t=0}^T f(\theta, \xi(t))\, dt = -\varepsilon\nabla J(\theta) + \mathrm{Err}(\varepsilon) \tag{14}$$

where $\|\mathrm{Err}(\varepsilon)\| \le O(\varepsilon^2)$. Thus, based on (4), the following algorithm seeks for (approximate) zeros of ∇J:

$$\begin{aligned} \tfrac{d}{dt}\theta(t) &= -a(t)\xi(t)J(x(t)) \\ x(t) &= \theta(t) + \varepsilon\xi(t)\,. \end{aligned} \tag{15}$$

In fact, (15) is a stylized version of *the extremum-seeking algorithm* of [12]. The gain $\boldsymbol{a}$ is typically assumed constant in this literature, and there is a large literature on how to improve the algorithm, such as through the introduction of a linear filter on the measurements $\{J(x(t))\}$. It is hoped that the results of this paper can be used to guide algorithm design in this application.

III. QSA for Reinforcement Learning

In this section we show how QSA can be used to speed up the exploration phase that is needed for policy evaluation in reinforcement learning.

A. *Off-policy TD Learning*

Consider the nonlinear state space model

$$\tfrac{d}{dt}x(t) = g(x(t), u(t))\,, \qquad t \geq 0$$

with $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$. Given a cost function $c\colon \mathbb{R}^{n+m} \to \mathbb{R}$, and a feedback law $u(t) = \phi(x(t))$, let J denote the associated value function:

$$J^\phi(x) = \int_0^\infty c(x(t), \phi(x(t)))\, dt\,, \qquad x = x(0).$$

The goal of policy evaluation (or TD-learning [33]) is to approximate this value function based on input-output measurements. It is assumed in [33] that the joint process $(\boldsymbol{x}, \boldsymbol{u})$ is an ergodic Markov chain, which presents an obvious challenge in this deterministic setting: this ergodic steady state will typically be degenerate. It is common to introduce noise, as in Q-learning [10], and also a discount factor in the definition of J to ensure that $J(x) < \infty$ for all x. Following these modifications, the approximation objective has been changed significantly: rather than approximating the original value function J, the algorithm will provide an approximation for the value function with discounting, and with a randomized policy. Exploration and/or discounting may create significant distortion in the value function.

The algorithm proposed here avoids these difficulties. The construction begins with a Q-function [10] defined with respect to the given policy:

$$Q^\phi(x, u) = J^\phi(x) + c(x, u) + g(x, u) \cdot \nabla J^\phi\,(x).$$

This function solves the fixed point equation

$$Q^\phi(x, u) = \underline{Q}^\phi(x) + c(x, u) + g(x, u) \cdot \nabla \underline{Q}^\phi\,(x) \tag{16}$$

in which we use the notational convention $\underline{F}(x) = F(x, \phi(x))$ for any function F. We consider a family of functions $Q^{\phi\theta}(x, u)$ parameterized by θ, and define the Bellman error for a given parameter as

$$\begin{aligned} \mathcal{E}^\theta(x, u) = &- Q^{\phi\theta}(x, u) + \underline{Q}^{\phi\theta}(x) + c(x, u) \\ &+ g(x, u) \cdot \nabla \underline{Q}^{\phi\theta}\,(x) \end{aligned} \tag{17}$$

The goal of policy evaluation is to create a data-driven algorithm that, without using information on the system's model, computes a parameter θ^* for which the Bellman error is small: for example, minimizes $\|\mathcal{E}^\theta\|$ in a given norm. In [10], ideas from [34] are used to construct a convex program for a related learning objective. In this paper, we propose an *off-policy* RL algorithm: the value function for ϕ is approximated while the actual input u of the system may be entirely unrelated.

We choose a feedback law with "excitation", of the form

$$u(t) = \kappa(x(t), \xi(t)) \tag{18}$$

where κ and $\boldsymbol{\xi}$ are such that the resulting state trajectories are bounded for each initial condition, and the joint process $(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{\xi})$ admits an ergodic steady state. The goal is to find θ^* that minimizes the mean square error:

$$\|\mathcal{E}^\theta\|^2 := \lim_{T\to\infty} \frac{1}{T} \int_0^T \bigl[\mathcal{E}^\theta(x(t), u(t))\bigr]^2\, dt. \tag{19}$$

Similarly to Section II-B, the first-order condition for optimality is expressed as a root-finding problem. Collecting together the definitions, we arrive at the following QSA steepest descent algorithm:

$$\begin{aligned} \tfrac{d}{dt}\theta(t) &= -a(t)\mathcal{E}^{\theta(t)}(x(t), u(t))\zeta^{\theta(t)}(t) \\ \zeta^\theta(t) &:= \nabla_\theta \mathcal{E}^\theta(x(t), u(t)) \end{aligned} \tag{20}$$

The vector process $\{\zeta^{\theta(t)}(t)\}$ is analogous to the *eligibility vector* defined in TD-learning [29], [30], [6].

Model-free realization. It appears from the definition (17) that the nonlinear model must be known. A model-free implementation is obtained on recognizing that for any parameter θ, and any state-input pair $(x(t), u(t))$,

$$\begin{aligned} \mathcal{E}^\theta(x(t), u(t)) = -Q^{\phi\theta}(x(t), u(t)) &+ \underline{Q}^{\phi\theta}(x(t)) \\ + c(x(t), u(t)) &+ \tfrac{d}{dt}\underline{Q}^{\phi\theta}(x(t)) \end{aligned} \tag{21}$$

(Approximate) Policy improvement algorithm (PIA): Given a policy ϕ and approximation $Q^{\phi\theta^*}$ for this policy, a new policy is obtained via:

$$\phi^+(x) = \arg\min_u Q^{\phi\theta^*}(x, u) \tag{22}$$

This procedure is repeated to obtain a recursive algorithm.

B. *Practical Implementation*

Given a basis of functions $\{\psi_i : 1 \leq i \leq d\}$, consider the linearly parameterized family

$$Q^{\phi\theta}(x, u) = d(x, u) + \theta^{\intercal}\psi(x, u)\,, \quad \theta \in \mathbb{R}^d. \tag{23}$$

Note that the Bellman error is a linear function of θ whenever this is true of $Q^{\phi,\theta}$. Consequently, minimization of (19) is a model-free linear regression problem, and the limit exists for any stable input. Moreover, the steepest descent algorithm (20) becomes linear. In fact, given (23), we define

$$\begin{aligned} \zeta(t) &:= [\psi(x(t), \phi(x(t))) - \psi(x(t), u(t)) \\ &\qquad + \tfrac{d}{dt}\psi(x(t), \phi(x(t)))] \\ b(t) &:= [c(x(t), u(t)) - d(x(t), u(t)) + d(x(t), \phi(x(t))) \\ &\qquad + \tfrac{d}{dt}d(x(t), \phi(x(t)))] \end{aligned}$$

Then $\mathcal{E}^{\phi,\theta}(x(t), u(t)) = b(t) + \zeta(t)^\top\theta$, and (20) becomes

$$\tfrac{d}{dt}\theta(t) = -a(t)\,\bigl[\zeta(t)^\top\,\theta(t) + b(t)\bigr]\,\zeta(t) \tag{24}$$

The convergence of (24) may be very slow if the matrix

$$G := \lim_{t\to\infty} \frac{1}{t} \int_0^t \zeta(\tau)\zeta(\tau)^\top \mathrm{d}\tau \tag{25}$$

is poorly conditioned (i.e., has some eigenvalues close to zero). Note that using G^{-1} as a matrix gain could solve this problem. The integral (25) can be estimated from data. This suggests an intuitive two-step procedure for the steepest descent algorithm (24)

$$\widehat{G}_t = \frac{1}{t}\int_0^t \zeta(\tau)\zeta(\tau)^\top \mathrm{d}\tau, \quad 0 \le t \le T \tag{26a}$$

$$\tfrac{d}{dt}\theta(t) = -a(t)\widehat{G}_T^{-1}\left[\zeta(t)^\top\theta(t) + b(t)\right]\zeta(t),\ t \ge T \tag{26b}$$

The results in Section IV suggest that this is indeed a good idea in order to achieve the optimal convergence rate $\mathcal{O}(1/t)$. To obtain this rate, the additional requirement is that $a(t) = g/(1+t)$, with $g > 1$.

C. Numerical example

Consider the LQR problem in which $g(x,u) = Ax + Bu$, and $c(x,u) = x^\top Mx + u^\top Ru$, with (A,B) controllable, $M \ge 0$ and $R > 0$. Given the known structure of the problem, we know that the function Q^ϕ associated with any linear policy $\phi(x) = Kx$, takes the form

$$Q^\phi = \begin{bmatrix} x \\ u \end{bmatrix}^\top \left(\begin{bmatrix} M & 0 \\ 0 & R \end{bmatrix} + \begin{bmatrix} A^\top P + PA + P & PB \\ B^\top P & 0 \end{bmatrix} \right) \begin{bmatrix} x \\ u \end{bmatrix},$$

where P solves the Lyapunov equation $A^\top P + PA + K^\top RK + Q = 0$ and therefore lies within the parametric class (23) in which $d(x,u) = c(x,u)$ and each ψ_i is a quadratic function of (x,u). For example, for the special case $n = 2$ and $m = 1$, we can take the quadratic basis

$$\{\psi_1, \dots, \psi_6\} = \{x_1^2, x_2^2, x_1x_2, x_1u, x_2u, u^2\}.$$

The algorithm (26b) was used in conjunction with the approximate PIA update (22) to obtain a sequence of policies, defined by state feedback, with $\phi_N(x) = K_N x$ at iteration N of the algorithm. The same input was used at each iteration:

$$u(t) = K^{\mathrm{e}}x(t) + \xi(t) \tag{27}$$

with $\xi(t) = \sum_{j=1}^q a_j \sin(\omega_j t + \phi_j)$, and $A - BK^{\mathrm{e}}$ Hurwitz. The gain K^{e} need not be the same K_N whose value function we wish to approximate.

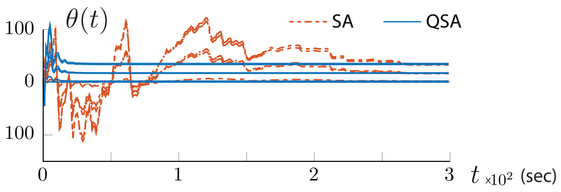

Fig. 3: Comparison of QSA and Stochastic Approximation (SA) for policy evaluation. It is observed that QSA converges significantly faster.

The algorithm was tested on the simple LQR example where the system is a double integrator with friction:

$$\dot{x} = \begin{bmatrix} 0 & -1 \\ 0 & -0.1 \end{bmatrix} x + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u, \quad M = I, \quad R = 10\,I. \tag{28}$$

In each experiment the input (27) was chosen for exploration, with $K^{\mathrm{e}} = [-1,-2]$ and ξ in (27) the sum of 24 sinusoids with frequency sampled uniformly between 0 and 50 rad/s, and phases sampled uniformly.

Figure 3 shows the evolution of the QSA algorithm for the evaluation of the policy $K = [-1, 0]$. The QSA algorithm is compared with the related SA algorithm in which ξ is "white noise" instead of a deterministic signal (formalized as an SDE). For implementation, both (26) and the linear system (28) were discretized with forward Euler discretization; time-step of 0.01s.

A plot of normalized policy error as a function of iteration N is shown in Figure 4.

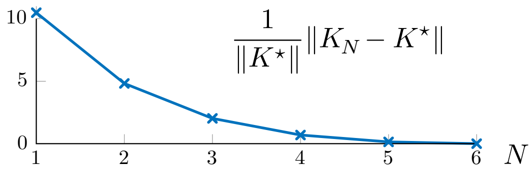

Fig. 4: Iterations of the policy improvement algorithm (PIA) (22) where each evaluation is performed by the model-free algorithm (26). The sequence of gain approximations obtained from the QSA PIA algorithm converge to the optimal gain $K^\star$.

IV. Convergence Analysis

The extension of stability and convergence results from the classical stochastic model (2) to the deterministic analog (4) requires some specialized analysis since the standard methods are not directly applicable. In particular, the first step in [2] and other references is to write (2) in the form,

$$\theta_{n+1} = \theta_n + a_n\big(\overline{f}(\theta_n) + M_n\big)\,,$$

where $\boldsymbol{M}$ is a martingale difference sequence (or a perturbation of such a sequence). This is possible when $\boldsymbol{\xi}$ is i.i.d., or for certain Markov $\boldsymbol{\xi}$ in (2). A similar transformation is not possible for any class of deterministic $\boldsymbol{\xi}$.

A. Assumptions for convergence

As in standard analysis of SA, the starting point is a temporal transformation: substitute in (4) the new time variable given by

$$u = g(t) := \int_0^t a(r)\,dr, \qquad t \ge 0.$$

The time-scaled process is then defined by

$$\widehat{\chi}(u) := \theta(g^{-1}(u)). \tag{29}$$

For example, if $a(r) = (1+r)^{-1}$, then

$$u = \log(1+t) \quad \text{and} \quad \xi(g^{-1}(u))) = \xi(e^u - 1). \tag{30}$$

The chain rule of differentiation gives

$$\tfrac{d}{du}\theta(g^{-1}(u)) = f(\theta(g^{-1}(u)), \xi(g^{-1}(u))).$$

That is, the time-scaled process solves the ODE,

$$\tfrac{d}{du}\widehat{\chi}(u) = f(\widehat{\chi}(u), \xi(g^{-1}(u))). \tag{31}$$

The two processes $\boldsymbol{\theta}$ and $\widehat{\boldsymbol{\chi}}$ differ only in time scale, and hence, proving convergence of one proves that of the other.

For the remainder of this section we will deal exclusively with $\widehat{\boldsymbol{\mathcal{X}}}$; it is on the 'right' time scale for comparison with $\boldsymbol{\mathcal{X}}$, the solution of (3).

Assumptions:

(A1) The system described by equation (3) has a globally asymptotically stable equilibrium at θ^*.

(A2) There exists a continuous function $V : \mathbb{R}^d \to \mathbb{R}_+$ and a constant $c_0 > 0$ such that, for any initial condition $\mathcal{X}(0)$ of (3), and any $0 \le T \le 1$, the following bounds hold whenever $\|\mathcal{X}(s)\| > c_0$,

$$V(\mathcal{X}(s+T)) - V(\mathcal{X}(s)) \le -T\|\mathcal{X}(s)\|.$$

(A3) There exists a constant $b_0 < \infty$, such that for all $\theta \in \mathbb{R}^d$, $T > 0$,

$$\left\| \frac{1}{T}\int_0^T f(\theta,\xi(t))\,dt - \bar{f}(\theta) \right\| \le \frac{b_0}{T}(1 + \|\theta\|)$$

(A4) There exists a constant $L < \infty$ such that the functions V, $\bar{f}$ and f satisfy the following Lipschitz conditions:

$$\begin{aligned} \|V(\theta') - V(\theta)\| &\le L\|\theta' - \theta\|, \\ \|\bar{f}(\theta') - \bar{f}(\theta)\| &\le L\|\theta' - \theta\|, \\ \|f(\theta',\xi) - f(\theta,\xi)\| &\le L\|\theta' - \theta\|, \quad \theta',\theta \in \mathbb{R}^d\,,\ \xi \in \mathbb{R}^m \end{aligned}$$

(A5) The process $\boldsymbol{a}$ is non-negative and monotonically decreasing, and as $t \to \infty$,

$$a(t) \downarrow 0, \qquad \int_0^t a(r)\,dr \to \infty.$$

Assumption (A1) determines uniquely the possible limit point of the algorithm. Assumption (A2) ensures that there is a Lyapunov function V with a strictly negative drift whenever $\boldsymbol{\mathcal{X}}$ escapes a ball of radius c_0. This assumption is used to establish boundedness of the trajectory $\widehat{\boldsymbol{\mathcal{X}}}$. Assumptions (A3) and (A4) are technical requirements essential to the proofs: (A3) is only slightly stronger than ergodicity of $\boldsymbol{\xi}$ as given by (5), while (A4) is necessary to control the growth of the respective functions. The process $\boldsymbol{a}$ in (A5) is a continuous time counterpart of the standard step size schedules in stochastic approximation, except that we impose monotonicity in place of square integrability.

Verifying (A2) for a linear system. Consider the ODE (3) in which $\bar{f}(x) = Ax$ with A a Hurwitz $d \times d$ matrix. There is a quadratic function $V_2(x) = x^{\intercal}Px$ satisfying the Lyapunov equation $PA + A^{\intercal}P = -I$, with $P > 0$. The function $V = k\sqrt{V_2}$, where the constant $k > 0$ is chosen suitably large, is a Lipschitz solution to (A2) for some finite c_0. ■

B. Convergence

The following is our main convergence result. The proof sketch is provided below; see the extended version [32] for the full proof.

Theorem 4.1: Under Assumptions (A1)–(A5), the solution to (4) converges to θ^* for each initial condition.

Define $\mathcal{X}^u(w)$, $w \ge u$, to be the unique solution to (3) 'starting' at $\widehat{\mathcal{X}}(u)$:

$$\tfrac{d}{dw}\mathcal{X}^u(w) = \bar{f}(\mathcal{X}^u(w)),\ w \ge u,\ \mathcal{X}^u(u) = \widehat{\mathcal{X}}(u). \tag{32}$$

The following result is required to prove Theorem 4.1.

Lemma 4.2: Under the assumptions of Theorem 4.1, for any $T > 0$, as $u \to \infty$,

$$\sup_{v\in[0,T]} \left\| \int_u^{u+v} \left[f(\widehat{\mathcal{X}}(w), \xi(g^{-1}(w))) - \bar{f}(\widehat{\mathcal{X}}(w))\right] dw \right\| \to 0$$

and $\sup_{v\in[0,T]} \|\widehat{\mathcal{X}}(u+v) - \mathcal{X}^u(u+v)\| \to 0$. ■

The proof of Lemma 4.2 is contained in [32]; the second limit is similar to Lemma 1 in Chapter 2 of [2].

Proof Sketch of Theorem 4.1: The first step in the proof is to establish ultimate boundedness of $\widehat{\mathcal{X}}(u)$: there exists $b < \infty$ such that for each $\theta \in \mathbb{R}^d$, there is a T_θ such that

$$\|\widehat{\mathcal{X}}(u)\| \le b \;\text{ for all } u \ge T_\theta\,,\ \widehat{\mathcal{X}}(0) = \theta$$

The (lengthy) proof is contained in [32].

Thus, for $u \ge T_\theta$, $\|\mathcal{X}^u(u)\| = \|\widehat{\mathcal{X}}(u)\| \le b$. By the definition of global asymptotic convergence, for every $\varepsilon > 0$, there exists a $\tau_\varepsilon > 0$, independent of the value $\mathcal{X}^u(u)$, such that $\|\mathcal{X}^u(u+v) - \theta^*\| < \varepsilon$ for all $v \ge \tau_\varepsilon$. Lemma 4.2 gives,

$$\begin{aligned} \limsup_{u\to\infty} &\|\widehat{\mathcal{X}}(u+\tau_\varepsilon) - \theta^*\| \\ &\le \limsup_{u\to\infty} \|\widehat{\mathcal{X}}(u+\tau_\varepsilon) - \mathcal{X}^u(u+\tau_\varepsilon)\| \\ &\quad + \limsup_{u\to\infty} \|\mathcal{X}^u(u+\tau_\varepsilon) - \theta^*\| \le \varepsilon. \end{aligned}$$

Since ε is arbitrary, we have the desired limit. ■

C. Variance

Let $\tilde{\theta}(t) := \theta(t) - \theta^*$ and $\nu(t) = (t+1)\tilde{\theta}(t)$. This section is devoted to providing conditions under which $\boldsymbol{\nu}$ is bounded, and there is a well defined covariance:

$$\overline{\Sigma}_\theta := \lim_{T\to\infty} \frac{1}{T}\int_0^T \nu(t)\nu(t)^{\intercal}\,dt. \tag{33}$$

Analysis requires additional assumptions on the "noise" process. It is also assumed that the model is linear and stable:

(A6) The function f is linear, $f(\theta,\xi) = A\theta + \xi$, the gain is $a(t) = 1/(t+1)$, and

(i) A is Hurwitz, and each eigenvalue $\lambda(A)$ satisfies $\text{Re}(\lambda) < -1$.

(ii) The function of time $\boldsymbol{\xi}$ is bounded, along with its partial integrals, denoted

$$\xi^I(t) = \int_0^t \xi(r)\,dr, \qquad \xi^{II}(t) = \int_0^t \xi^I(r)\,dr.$$

Assumption (A6) implies that $\bar{f}(\theta) = A\theta$, so that $\theta^* = 0$. The linearity assumption is typical in much of the literature on variance for stochastic approximation [11], [5], [2]. As in the SA literature, it is likely that the results of this section can be extended to nonlinear models via a Taylor-series approximation.

A typical example of Assumption (A6ii) is the case where the entries of $\boldsymbol{\xi}$ can be expressed as a sum of sinusoids:

$$\xi(t) = \sum_{i=1}^K v^i \sin(\phi_i + \omega_i t) \tag{34}$$

for fixed vectors $\{v^i\}$, phases $\{\phi_i\}$, and frequencies $\{\omega_i\}$.

Theorem 4.3 below implies that $\|\nu(t) - \xi^I(t)\| \to 0$, as $t \to \infty$. Consequently, the error covariance exists whenever there is a covariance for $\boldsymbol{\xi}^I$:

$$\overline{\Sigma}_\theta = \Sigma_{\xi^I} := \lim_{T\to\infty} \frac{1}{T} \int_0^T \xi^I(t) \xi^I(t)^{\intercal} \, dt.$$

This is easily computed for the special case (34).

Let $\bar{A} := I + A$ and fix a constant ε_S satisfying $0 < \varepsilon_S < -\text{Re}(\bar{\lambda})$ for each eigenvalue $\bar{\lambda}$ of $\bar{A}$; this is possible due to Assumption (A6i). Associated with the ODE $\frac{d}{dt}x(t) = (1+t)^{-1}\bar{A}x(t)$ is the *state transition matrix*:

$$S(t;r) = \exp\Bigl(\log\Bigl[\frac{1+t}{1+r}\Bigr]\bar{A}\Bigr)\,, \qquad r,t \ge 0. \tag{35}$$

It is easily shown that it satisfies the defining properties

$$S(t;t) = I\,, \quad \tfrac{d}{dt}S(t;r) = \frac{1}{t+1}\bar{A}S(t;r)\,, \quad r,t\ge 0. \tag{36}$$

Theorem 4.3: Suppose Assumptions (A1)–(A6) hold. Then, for each initial condition $\theta(0)$,

$$\tilde{\theta}(t) = \frac{1}{t+1}\Bigl[\xi^I(t) + S(t;0)\tilde{\theta}(0)\Bigr] + O\Bigl(\frac{1}{(t+1)^{1+\delta_S}}\Bigr), \tag{37}$$

where $\delta_S = \min(\varepsilon_S, 1)$, and the final error term is independent of the initial condition $\tilde{\theta}(0)$. Consequently, the scaled error process satisfies the bound

$$\nu(t) = \xi^I(t) + O\Bigl(\frac{1+\|\tilde{\theta}(0)\|}{(t+1)^{\delta_S}}\Bigr). \tag{38}$$

■

The remarkable coupling bound (38) follows from (37) and Lemma 4.4 below. Coupling is illustrated here using the simple integration experiment of Section II-A. The representation (9) must be modified to fit the assumptions of the theorem. First, denote by $\boldsymbol{\xi}^0$ a periodic function of time whose sample paths define the uniform distribution on $[0,1]$: for any continuous function c,

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T c(\xi^0(t))\,dt = \int_0^1 c(x)\,dx.$$

Introduce a gain $g > 0$, and consider the error equation,

$$\tfrac{d}{dt}\tilde{\theta}(t) = \frac{g}{t+1}[y(\xi^0(t)) - \theta^* - \tilde{\theta}(t)] \tag{39}$$

The assumptions of the theorem are satisfied with $A = -g$ and $\xi(t) = g[y(\xi^0(t)) - \theta^*]$.

Figures 1 and 2 illustrate the qualitative conclusion of Theorem 4.3: that it is useful to choose $g > 1$ in (39), so that Assumption (A6i) is satisfied.

Coupling is illustrated in Fig. 5. The scaled errors $g^{-1}\boldsymbol{\nu}$ are compared since $\boldsymbol{\xi}$ grows linearly with g: we expect $g^{-1}\nu(t) \approx \int_0^t (y(\xi^0(r)) - \theta^*)$ for large t. The initial condition was set to $\theta(0) = 10$ in each experiment.

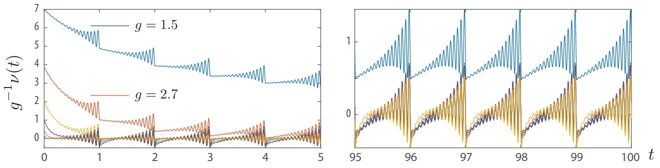

Fig. 5: Evolution of $\nu(t) = (1+t)\tilde{\theta}(t)$ using Quasi Monte-Carlo estimates for a range of gains.

The figure shows results using ten gains, approximately equally spaced on a logarithmic scale. The smallest gain is $g = 1.5$, and all other gains satisfy $g \ge 2$. Theorem 4.3 asserts that $|\nu(t) - \xi^I(t)| = O\bigl([1+t]^{-\delta_S}\bigr)$, where $\delta_S < 0.5$ for $g = 1.5$, and $\delta_S = 1$ for $g \ge 2$. The scaled errors $\{g^{-1}\nu(t) : 95 \le t \le 100\}$ are nearly indistinguishable when $g \ge 2$. The slower convergence for $g = 2.7$ is probably due to the term $S(t;0)\tilde{\theta}(0)$ appearing in (37).

Results using gains $g \le 1$ are omitted. As expected, $\boldsymbol{\nu}$ is unbounded for $g < 1$. For $g = 1$ the approximation (38) fails since $\nu(t)$ evolves near $\nu(0)$ for the entire run.

The proof of Theorem 4.3 leverages the following auxiliary results. Let $\tilde{\nu}(t) = \nu(t) - \xi^I(t)$, $t \ge 0$, denote the "second-order" error process.

Lemma 4.4: The scaled error processes solve the respective linear differential equations

$$\begin{aligned} \tfrac{d}{dt}\nu(t) &= \frac{1}{t+1}\bar{A}\nu(t) + \xi(t) \\ \tfrac{d}{dt}\tilde{\nu}(t) &= \frac{1}{t+1}\bar{A}\tilde{\nu}(t) + \frac{1}{t+1}\bar{A}\xi^I(t) \end{aligned} \tag{40}$$

The ODE for the second-order error admits the solution

$$\tilde{\nu}(t) = S(t,0)\tilde{\theta}(0) + \int_0^t \frac{1}{r+1} S(t;r)\bar{A}\xi^I(r)\,dr \tag{41}$$

where S is defined in (35). Under the eigenvalue assumptions in (A6), there exists $b_S < \infty$ such that

$$\|S(t;r)\|_2 \le b_S \Bigl[\frac{1+t}{1+r}\Bigr]^{-\varepsilon_S}$$

where $\|S(t;r)\|_2$ denotes the maximal singular value.

Proof: The representation follows from the state transition matrix interpretation (36). The bound on $\|S(t;r)\|_2$ easily follows. ■

The proof of the next result is contained in [32].

Lemma 4.5: For $t \ge 0$,

$$\begin{aligned} &\int_0^t \frac{1}{1+r} S(t;r)\bar{A}\xi^I(r)\,dr \\ &\quad = \frac{1}{1+t}\bar{A}\xi^{II}(t) - S(t;0)\bar{A}\xi^{II}(0) \\ &\qquad + \int_0^t \frac{1}{(1+r)^2} S(t;r)[I+\bar{A}]\bar{A}\xi^{II}(r)\,dr\,. \end{aligned} \tag{42}$$

There exists $b_\nu < \infty$ such that

$$\int_0^t \frac{1}{(1+r)^2}\|S(t;r)\|\,dr \le b_\nu \frac{1}{(1+t)^{\delta_S}}\,, \quad t \ge 0\,. \tag{43}$$

■

Proof of Theorem 4.3: Lemmas 4.4 and 4.5 give

$$\begin{aligned} \tilde{\nu}(t) &= S(t,0)\tilde{\theta}(0) + \mathcal{E}_{\tilde{\nu}}(t) \\ \mathcal{E}_{\tilde{\nu}}(t) &= \frac{1}{t+1}\bar{A}\xi^{II}(t) - S(t;0)\bar{A}\xi^{II}(0) \\ &\qquad + \int_0^t \frac{1}{(1+r)^2} S(t;r)[I+\bar{A}]\bar{A}\xi^{II}(r)\,dr\,. \end{aligned}$$

The two lemmas imply that $\|\mathcal{E}_{\tilde{\nu}}(t)\| \le O\bigl((1+t)^{-\delta_S}\bigr)$. ■

V. Conclusion

While QSA can result in significant improvement in convergence rate, the results of Section IV demonstrate that QSA algorithms must be implemented with care. If the gain does not satisfy the assumptions of Theorem 4.3 then the rate of convergence can be *slower* than obtained in an i.i.d. or Markovian setting.

There are many interesting topics for future research:

(i) Further work is required to extend Theorem 4.3 to the nonlinear algorithm.

(ii) Constant-gain algorithms are amenable to analysis using similar techniques.

(iii) Analysis of convergence under local stability assumptions (to local attractors) can be performed using tools similar to those used in the standard SA literature.

(iv) We are most interested in applications to control and optimization:

(a) On-line learning applications, in which the function f itself varies with time. That is, (4) is replaced by

$$\tfrac{d}{dt}\theta(t) = a f_t(\theta(t), \xi(t))\,,$$

Analysis will be far simpler than in a traditional SA setting.

(b) Applications to decentralized control using reinforcement learning techniques. In the LQR setting, the architecture for Q-learning or fixed-policy Q-learning might be informed by recent computational techniques for control synthesis [35].

References

[1] H. Robbins and S. Monro, "A stochastic approximation method," *Annals of Mathematical Statistics*, vol. 22, pp. 400–407, 1951.

[2] V. S. Borkar, *Stochastic Approximation: A Dynamical Systems Viewpoint*. Delhi, India and Cambridge, UK: Hindustan Book Agency and Cambridge University Press (jointly), 2008.

[3] V. S. Borkar and S. P. Meyn, "The ODE method for convergence of stochastic approximation and reinforcement learning," *SIAM J. Control Optim.*, vol. 38, no. 2, pp. 447–469, 2000, (also presented at the *IEEE CDC*, December, 1998).

[4] A. Benveniste, M. Métivier, and P. Priouret, *Adaptive algorithms and stochastic approximations*, ser. Applications of Mathematics (New York). Berlin: Springer-Verlag, 1990, vol. 22, translated from the French by Stephen S. Wilson.

[5] H. J. Kushner and G. G. Yin, *Stochastic approximation algorithms and applications*, ser. Applications of Mathematics (New York). New York: Springer-Verlag, 1997, vol. 35.

[6] D. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*. Cambridge, Mass: Atena Scientific, 1996.

[7] D. Huang, W. Chen, P. Mehta, S. Meyn, and A. Surana, "Feature selection for neuro-dynamic programming," in *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control*, F. Lewis, Ed. Wiley, 2011.

[8] A. M. Devraj and S. P. Meyn, "Fastest convergence for Q-learning," *ArXiv e-prints*, Jul. 2017.

[9] ——, "Zap Q-learning," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.

[10] P. G. Mehta and S. P. Meyn, "Q-learning and Pontryagin's minimum principle," in *Proc. of the IEEE Conf. on Dec. and Control*, Dec. 2009, pp. 3598–3605.

[11] V. R. Konda and J. N. Tsitsiklis, "Convergence rate of linear two-time-scale stochastic approximation," *Ann. Appl. Probab.*, vol. 14, no. 2, pp. 796–819, 2004. [Online]. Available: http://www.jstor.org/stable/4140429

[12] S. Liu and M. Krstic, "Introduction to extremum seeking," in *Stochastic Averaging and Stochastic Extremum Seeking*, ser. Communications and Control Engineering. London: Springer, 2012.

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

[14] B. Lapeybe, G. Pagès, and K. Sab, "Sequences with low discrepancy generalisation and application to Robbins-Monro algorithm," *Statistics*, vol. 21, no. 2, pp. 251–272, 1990.

[15] S. Laruelle and G. Pagès, "Stochastic approximation with averaging innovation applied to finance," *Monte Carlo Methods and Applications*, vol. 18, no. 1, pp. 1–51, 2012.

[16] S. Shirodkar and S. Meyn, "Quasi stochastic approximation," in *Proc. of the 2011 American Control Conference (ACC)*, July 2011, pp. 2429–2435.

[17] J. Kiefer and J. Wolfowitz, "Stochastic estimation of the maximum of a regression function," *Ann. Math. Statist.*, vol. 23, no. 3, pp. 462–466, 09 1952. [Online]. Available: https://doi.org/10.1214/aoms/1177729392

[18] J. C. Spall, "Multivariate stochastic approximation using a simultaneous perturbation gradient approximation," *IEEE Transactions on Automatic Control*, vol. 37, no. 3, pp. 332–341, March 1992.

[19] S. Bhatnagar, M. C. Fu, S. I. Marcus, and I.-J. Wang, "Two-timescale simultaneous perturbation stochastic approximation using deterministic perturbation sequences," *ACM Trans. Model. Comput. Simul.*, vol. 13, no. 2, pp. 180–209, 2003.

[20] A. D. Flaxman, A. T. Kalai, and H. B. McMahan, "Online convex optimization in the bandit setting: Gradient descent without a gradient," in *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, ser. SODA '05. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2005, pp. 385–394.

[21] B. Awerbuch and R. Kleinberg, "Online linear optimization and adaptive routing," *Journal of Computer and System Sciences*, vol. 74, no. 1, pp. 97 – 114, 2008, learning Theory 2004.

[22] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Foundations and Trends® in Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012.

[23] T. Chen and G. B. Giannakis, "Bandit convex optimization for scalable and dynamic iot management," *IEEE Internet of Things Journal*, vol. 6, no. 1, pp. 1276–1286, Feb 2019.

[24] M. Krstić and H.-H. Wang, "Stability of extremum seeking feedback for general nonlinear dynamic systems," *Automatica*, vol. 36, no. 4, pp. 595 – 601, 2000.

[25] H.-H. Wang and M. Krstić, "Extremum seeking for limit cycle minimization," *IEEE Transactions on Automatic Control*, vol. 45, no. 12, pp. 2432–2436, Dec 2000.

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

[27] B. T. Polyak and A. B. Juditsky, "Acceleration of stochastic approximation by averaging," *SIAM J. Control Optim.*, vol. 30, no. 4, pp. 838–855, 1992.

[28] D. Ruppert, "A Newton-Raphson version of the multivariate Robbins-Monro procedure," *The Annals of Statistics*, vol. 13, no. 1, pp. 236–245, 1985. [Online]. Available: http://www.jstor.org/stable/2241156

[29] C. Szepesvári, *Algorithms for Reinforcement Learning*, ser. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2010.

[30] R. Sutton and A. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press. On-line edition at http://www.cs.ualberta.ca/~sutton/book/the-book.html, 1998.

[31] S. Bradtke, B. Ydstie, and A. Barto, "Adaptive linear quadratic control using policy iteration," in *Proc. of the American Control Conf.*, vol. 3, 1994, pp. 3475–3479.

[32] A. Bernstein, Y. Chen, M. Colombino, E. Dall'Anese, S. Meyn, and P. Mehta, "Optimal rate of convergence for quasi-stochastic approximation," *(In preparation)*, 2019.

[33] J. N. Tsitsiklis and B. Van Roy, "An analysis of temporal-difference learning with function approximation," *IEEE Trans. Automat. Control*, vol. 42, no. 5, pp. 674–690, 1997.

[34] D. P. de Farias and B. Van Roy, "The linear programming approach to approximate dynamic programming," *Operations Res.*, vol. 51, no. 6, pp. 850–865, 2003.

[35] N. K. Dhingra, S. Z. Khong, and M. R. Jovanović, "The proximal augmented Lagrangian method for nonsmooth composite optimization," *IEEE Trans. Automat. Control*, 2018.