2011 50th IEEE Conference on Decision and Control and
European Control Conference (CDC-ECC)
Orlando, FL, USA, December 12-15, 2011

Asymptotic Bias of Stochastic Gradient Search

Vladislav B. Tadić and A. Doucet

Abstract— **The asymptotic behavior of the stochastic gradient algorithm with a biased gradient estimator is analyzed. Relying on arguments based on differential geometry (Yomdin theorem and Lojasiewicz inequality), relatively tight bounds on the asymptotic bias of the iterates generated by such an algorithm are derived. The obtained results hold under mild and verifiable conditions and cover a broad class of complex stochastic gradient algorithms. Using these results, the asymptotic properties of the actor-critic reinforcement learning are studied.**

Index Terms— **Stochastic gradient search, biased gradient estimation, reinforcement learning.**

I. Introduction

Many problems in automatic control, system identification, signal processing, machine learning, operations research and statistics can be posed as a stochastic optimization problem, i.e., as the minimization (or maximization) of an unknown objective function whose values are available only through noisy observations. Such a problem can efficiently be solved by stochastic gradient search (also known as the stochastic gradient algorithm). Stochastic gradient search is a procedure of the stochastic approximation type which iteratively approximates the minima of the objective function using a statistical or Monte Carlo estimator of the gradient (of the objective function). Not rarely, the estimator is biased, since the consistent gradient estimation is often computationally expensive or not available at all. As a result of the biased gradient estimation, the stochastic gradient search is biased, too, i.e., the corresponding algorithm does not converge to the set of minima, but to its vicinity. In order to interpret the results produced by such an algorithm and to tune the algorithm's parameters (e.g., to achieve a better bias/variance balance and a better convergence rate), the knowledge about the asymptotic bias of the algorithm iterates is crucially needed.

Despite its practical and theoretical importance, the asymptotic behavior of the stochastic gradient search with biased gradient estimation (also referred to as the biased stochastic gradient search) has not attracted much attention in the literature on stochastic optimization and stochastic approximation. To the best of the present author's knowledge, the asymptotic properties of the biased stochastic gradient search (and the biased stochastic approximation) have only been analyzed in [5], [8], [9] and [10]. Although these results provide a good insight into the asymptotic behavior of the biased gradient search, they hold under restrictive conditions which are very hard to verify for complex nonlinear algorithms. Moreover, unless the objective function is of a simple form (e.g., convex), none of the results of [5], [8], [9], [10] offers an explicit bound on the asymptotic bias of the algorithm iterates.

In this paper, we study the asymptotic behavior of the biased gradient search. Using arguments based on differential geometry (Yomdin theorem and Lojasiewicz inequality), we derive relatively tight bounds on the asymptotic bias of the algorithm iterates. The obtained results hold under mild and easily verifiable conditions and cover a broad class of complex stochastic gradient algorithms. In this paper, we show how the results can be applied to the asymptotic analysis of actor-critic reinforcement learning.

The paper is organized as follows. The main results are presented in Section II, where the stochastic gradient search with additive noise is analyzed. In Section III, the asymptotic bias of the stochastic gradient search with Markovian dynamics is studied. In Section IV, the asymptotic bias of actor-critic reinforcement learning is assessed using the general results obtained in Sections II and III.

II. Main Results

In this section, the asymptotic behavior of the following algorithm is analyzed:

$$\theta_{n+1} = \theta_n - \alpha_n(\nabla f(\theta_n) + \xi_n), \quad n \geq 0. \tag{1}$$

Here, $f : \mathbb{R}^{d_\theta} \to \mathbb{R}$ is a differentiable function, while $\{\alpha_n\}_{n\geq 0}$ is a sequence of positive real numbers. θ_0 is an $\mathbb{R}^{d_\theta}$-valued random variable defined on a probability space $(\Omega, \mathcal{F}, P)$, while $\{\xi_n\}_{n\geq 0}$ is an $\mathbb{R}^{d_\theta}$-valued stochastic process defined on the same probability space. To allow more generality, we assume that for each $n \geq 0$, ξ_n is a random function of $\theta_0, \ldots, \theta_n$. In the area of stochastic optimization, recursion (1) is known as the stochastic gradient search (or the stochastic gradient algorithm). The recursion minimizes function $f(\cdot)$, which is usually referred to as the objective function. Term $\nabla f(\theta_n) + \xi_n$ is interpreted as a gradient estimator (i.e., an estimator of $\nabla f(\theta_n)$), while ξ_n represents the estimator's noise (or error). For further details, see [22], [27] and references given therein.

Throughout the paper, the following notation is used. The Lebesgue measure is denoted by $m(\cdot)$, while $\|\cdot\|$ and $d(\cdot,\cdot)$ stand for the Euclidean norm and the Euclidean distance (respectively). S and A are the sets of stationary points and the critical values of $f(\cdot)$, i.e.,

$$S = \{\theta \in \mathbb{R}^{d_\theta} : \nabla f(\theta) = 0\}, \quad A = \{f(\theta) : \theta \in S\}.$$

For a compact set $Q \subset \mathbb{R}^{d_\theta}$ and $\varepsilon \in (0, \infty)$, $A_{Q,\varepsilon}$ denotes the set of ε-critical points of $f(\cdot)$ contained in the f-image

V. B. Tadić is with the Department of Mathematics, University of Bristol, Bristol BS8 1TW, United Kingdom. (v.b.tadic@bristol.ac.uk).

A. Doucet is with the Department of Statistics, University of Oxford, Oxford OX1 3TG, United Kingdom. (adoucet@stats.ox.ac.uk).

of Q, i.e.,

$$A_{Q,\varepsilon} = \{f(\theta) : \theta \in Q, \|\nabla f(\theta)\| \leq \varepsilon\}.$$

For $t \in (0,\infty)$ and $n \geq 0$, $a(n,t)$ is the integer defined as

$$a(n,t) = \min\left\{k \geq n : \sum_{i=n}^{k} \alpha_i > t\right\}.$$

The algorithm (1) is analyzed under the following assumptions:

Assumption 2.1: $\lim_{n\to\infty} \alpha_n = 0$ and $\sum_{n=0}^{\infty} \alpha_n = \infty$.

Assumption 2.2: There exist $\mathbb{R}^{d_\theta}$-valued stochastic processes $\{\zeta_n\}_{n\geq 0}$ and $\{\rho_n\}_{n\geq 0}$ (defined on $(\Omega, \mathcal{F}, P)$) such that $\xi_n = \zeta_n + \rho_n$ for each $n \geq 0$ and such that

$$\lim_{n\to\infty} \max_{n\leq k<a(n,t)} \left\|\sum_{i=n}^{k} \alpha_i \zeta_i\right\| = 0, \tag{2}$$

$$\limsup_{n\to\infty} \|\rho_n\| < \infty \tag{3}$$

w.p.1 on $\{\sup_{n\geq 0} \|\theta_n\| < \infty\}$ for any $t \in (0,\infty)$.

Assumption 2.3: There exists a real number $p \in (0,1]$ and for any compact set $Q \subset \mathbb{R}^{d_\theta}$, there exists another real number $M_Q \in [1,\infty)$ such that

$$m(A_{Q,\varepsilon}) \leq M_Q \varepsilon^p \tag{4}$$

for all $\varepsilon \in [0,\infty)$.

Remark 2.1: Due to the Yomdin theorem [33, Theorem 1.2] (also known as the quantitative version of the Morse-Sard theorem), Assumption 2.3 holds if $f(\cdot)$ is q times differentiable and $d_\theta < q < \infty$. In this case, $p = (q - d_\theta)/(q-1)$. A further insight can be provided for the case when $f(\cdot)$ is real-analytic: If $f(\cdot)$ is real-analytic, Yomdin theorem and Lojasiewicz inequality [19, Theorem 17], [20] imply that for any compact set $Q \subset R^{d_\theta}$, there exit real numbers $r_Q \in (0,1]$, $M_Q, N_Q \in [1,\infty)$ such that

$$m(A_{Q,\varepsilon}) \leq M_Q \varepsilon, \quad d(\theta, S) \leq N_Q \|\nabla f(\theta)\|^{r_Q} \tag{5}$$

for all $\varepsilon \in [0,\infty)$, $\theta \in Q$. Quantities p and r_Q are usually referred to as the Yomdin and Lojasiewicz exponents.

Assumption 2.1 corresponds to the step-size sequence $\{\alpha_n\}_{n\geq 0}$ and is commonly used in the asymptotic analysis of stochastic gradient and stochastic approximation algorithms. Assumption 2.2 is a noise condition. It can be interpreted as a decomposition of the gradient estimator's noise $\{\xi_n\}_{n\geq 0}$ into a zero-mean sequence $\{\zeta_n\}_{n\geq 0}$ (which is averaged out by step-sizes $\{\alpha_n\}_{n\geq 0}$, see (2)) and the estimator's bias $\{\rho_n\}_{n\geq 0}$ (which is almost surely bounded, see (3)). Assumption 2.2 is satisfied if $\{\zeta_n\}_{n\geq 0}$ are martingale-differences and $\{\rho_n\}_{n\geq 0}$ are a continuous function of $\{\theta_n\}_{n\geq 0}$. It also holds for gradient search with Markovian dynamics (see Section III). If the gradient estimator is unbiased (i.e., $\lim_{n\to\infty} \rho_n = 0$ w.p.1), Assumption 2.2 reduces to the well-known Kushner-Clark condition, the weakest noise assumption under which the almost sure convergence of (1) can be demonstrated. Assumption 2.3 is related to the stability of the gradient flow $d\theta/dt = -\nabla f(\theta)$, or more specifically, to the geometry of the stationary points and the critical values of $f(\cdot)$. As explained in Remark 2.1, this assumption holds if $f(\cdot)$ is at least $d_\theta + 1$ times differentiable. Although such a degree of differentiability can be considered as a restrictive condition, it holds for the objective functions of many stochastic gradient algorithms used in system identification, signal processing, machine learning and statistics. E.g., in Section IV, we show that the objective function associated with actor-critic reinforcement learning is smooth (i.e., infinitely many times differentiable). In [32], we prove the same property for the objective functions associated with sequential Monte Carlo methods for the identification of non-linear non-Gaussian state-space models. In [30], we show the analyticity for the objective functions associated with the recursive maximum likelihood estimation in hidden Markov models. It is also worth mentioning that the objective functions associated with principal component analysis (as well as with many other adaptive signal processing algorithms) are often polynomial or rational, and hence, smooth and analytic, too (see e.g., [12] and references cited therein).

In order to state the main results of this section, we need some further notation. For a compact set $Q \subset \mathbb{R}^{d_\theta}$, Λ_Q denotes the event

$$\Lambda_Q = \liminf_{n\to\infty} \{\theta_n \in Q\} = \bigcup_{n=0}^{\infty} \bigcap_{k=n}^{\infty} \{\theta_k \in Q\}.$$

$L_Q \in [1,\infty)$ stands for an upper bound of $\|\nabla f(\cdot)\|$ on Q and for a Lipschitz constant of $f(\cdot)$, $\nabla f(\cdot)$ on the same set. Moreover, ρ is the random variable defined by

$$\rho = \limsup_{n\to\infty} \|\rho_n\|.$$

With this notation, our main result on the asymptotic bias of the recursion (1) can be stated as follows.

Theorem 2.1: Let Assumptions 2.1 – 2.3 hold. Then, for any compact set $Q \subset \mathbb{R}^{d_\theta}$, there exists a real number $K_Q \in [1,\infty)$ (depending only on L_Q and M_Q) such that

$$\limsup_{n\to\infty} \|\nabla f(\theta_n)\| \leq K_Q \rho^{p/2}, \tag{6}$$

$$\limsup_{n\to\infty} f(\theta_n) - \liminf_{n\to\infty} f(\theta_n) \leq K_Q \rho^p \tag{7}$$

w.p.1 on Λ_Q.

The proof of Theorem 2.1 is provided in [31]. As a direct consequence of the Yomdin theorem, Lojasiewicz inequality and Theorem 2.1, the following result is obtained.

Corollary 2.1: Let Assumptions 2.1 and 2.2 hold.

(i) Suppose that $f(\cdot)$ is q-times differentiable and that $d_\theta < q < \infty$. Then, all conclusions of Theorem 2.1 hold with $p = (q - d_\theta)/(q-1)$.

(ii) Suppose that $f(\cdot)$ is real-analytic. Then, the conclusions of Theorem 2.1 hold with $p = 1$. Moreover, for any compact set $Q \subset R^{d_\theta}$, there exists a real number $K_Q \in [1,\infty)$ (depending only on L_Q, M_Q and N_Q) such that

$$\limsup_{n\to\infty} \|\nabla f(\theta_n)\| \leq K_Q \rho^{r_Q/2}$$

w.p.1 on Λ_Q.

In the literature on stochastic optimization, it is well-known that stochastic gradient search with an unbiased gradient estimator (the case where $\rho = 0$) exhibits the following asymptotic behavior: $\lim_{n\to\infty} \nabla f(\theta_n) = 0$, $\lim_{n\to\infty} d(\theta_n, S) = 0$ and $\{f(\theta_n)\}_{n\geq 0}$ converges (i.e., $\limsup_{n\to\infty} f(\theta_n) = \liminf_{n\to\infty} f(\theta_n)$). When the gradient estimator is biased (i.e., $\rho > 0$), this is not true any more: Now, the quantities

$$\limsup_{n\to\infty} \|\nabla f(\theta_n)\|, \tag{8}$$

$$\limsup_{n\to\infty} d(\theta_n, S), \tag{9}$$

$$\limsup_{n\to\infty} f(\theta_n) - \liminf_{n\to\infty} f(\theta_n) \tag{10}$$

are strictly positive. However, it is reasonable to expect these quantities to decrease in ρ and to tend to zero when $\rho \to 0$. Hence, the quantities (8) – (10) and the way they depend on ρ can be considered as a sensible characterization of the asymptotic behavior and the asymptotic bias of the gradient search with biased gradient estimation. In the case of algorithm (1), such a characterization is contained in Theorem 2.1 and Corollary 2.1: Both the theorem and its corollary provide relatively tight, explicit bounds on the quantities (8) – (10) in terms of the gradient estimator's bias ρ and the Yomdin and Lojasiewicz exponents. Apparently, the results of Theorem 2.1 and Corollary 2.1 are of a local nature: They hold only on the event where algorithm (1) is stable (i.e., where sequence $\{\theta_n\}_{n\geq 0}$ belongs to a compact set Q). Stating results on the asymptotic bias of stochastic gradient search in such a local form is quite sensible due to the following reasons. The stability of stochastic gradient search is based on well-understood arguments which are rather different from the arguments the proofs of Theorem 2.1 and Corollary 2.1 rely on. Moreover, as illustrated in Appendix and [31], it is relatively easy to get a global version of Theorem 2.1 and Corollary 2.1 by combining them with the methods used to verify or to ensure the stability (e.g., with the results of [1, Section II.1.9], [4] and [10]). It is also worth mentioning that local asymptotic results are typical in the areas of stochastic optimization and stochastic approximation (e.g., similarly as Theorem 2.1 and Corollary 2.1, most of the results of [1, Part II] hold only on set Λ_Q).

Gradient algorithms with biased gradient estimation are often used in the areas of system identification [14], [23], machine learning [2], [6], [16], operations research [11], [15], and statistics [7], [26]. To interpret the result produced by such an algorithm and to tune the algorithm's parameters (e.g., to achieve better bias/variance balance and convergence rate), the knowledge of the asymptotic bias is crucially needed. Despite this fact, the asymptotic behavior of the gradient search with biased gradient estimation has not received much attention in the literature on stochastic optimization and stochastic approximation. To the best of the present author's knowledge, the asymptotic properties of the biased stochastic gradient search and biased stochastic approximation has been studied only in [5, Section 5.3], [8], [9], [10, Section 2.7]. Although these results provide a good insight into the asymptotic behavior of the biased gradient search, they hold only if $f(\cdot)$ is unimodal or if $\{\theta_n\}_{n\geq 0}$ is contained in the domain of an asymptotically stable attractor of $d\theta/dt = -\nabla f(\theta)$. Moreover, unless $f(\cdot)$ is of a simple form (e.g., convex), the results of [5, Section 5.3], [8], [9], [10, Section 2.7] do not provide any explicit bound on the asymptotic bias of the gradient search with biased gradient estimation. Unfortunately, in the case of complex stochastic gradient algorithms (such as those studied in Section IV), $f(\cdot)$ is usually multimodal and very complicated (with lot of unisolated local extrema and saddle points). Consequently, for such algorithms, not only it is hard to verify the assumptions adopted in [5, Section 5.3], [8], [9], [10, Section 2.7], but these assumptions are likely not to hold at all.

Relying on the Yomdin theorem (a quantitative version of the Morse-Sard theorem) and the Lojasiewicz inequality, Theorem 2.1 and Corollary 2.1 overcome the difficulties described in the previous paragraph. Both the theorem and its corollary allow the objective function to be multimodal (with manifolds of unisolated extrema and saddle points), do not require $d\theta/dt = -\nabla f(\theta)$ to have an asymptotically stable attractor and do not assume (apriori) any particular behavior of $\{\theta_n\}_{n\geq 0}$. In addition to this, Theorem 2.1 and Corollary 2.1 provide relatively tight explicit bounds on the asymptotic bias of algorithm (1). Moreover, as illustrated in Section IV, the theorem and its corollary cover a relatively broad class of stochastic gradient algorithms used in reinforcement learning. In another paper [32], using Theorem 2.1, we analyze sequential Monte Carlo methods for the identification of nonlinear non-Gaussian state-space models and we demonstrate that the asymptotic bias of these methods converges to zero polynomially in the number of particles.

III. Stochastic Gradient Search with Markovian Dynamics

In order to illustrate the results of Section II and to set up a framework for the analysis carried out in Section IV, we apply Theorem 2.1 to stochastic gradient algorithms with Markovian dynamics. These algorithms are defined by the following difference equation:

$$\theta_{n+1} = \theta_n - \alpha_n(F(\theta_n, Z_{n+1}) + \rho_n), \quad n \geq 0. \tag{11}$$

In this recursion, $F : \mathbb{R}^{d_\theta} \times \mathbb{R}^{d_z} \to \mathbb{R}^{d_\theta}$ is a Borel-measurable function, while $\{\alpha_n\}_{n\geq 0}$ is a sequence of positive real numbers. θ_0 is an $\mathbb{R}^{d_\theta}$-valued random variable defined on a probability space $(\Omega, \mathcal{F}, P)$, while $\{Z_n\}_{n\geq 0}$ and $\{\rho_n\}_{n\geq 0}$ are $\mathbb{R}^{d_z}$ and $\mathbb{R}^{d_\theta}$-valued stochastic processes defined on the same probability space. More specifically, ρ_n is a random function of $\theta_0, \ldots, \theta_n$ for each $n \geq 0$. $\{Z_n\}_{n\geq 0}$ is a Markov process controlled by $\{\theta_n\}_{n\geq 0}$, i.e., there exists a family of transition probability kernels $\{\Pi_\theta(\cdot,\cdot)\}_{\theta\in\mathbb{R}^{d_\theta}}$ on $\mathbb{R}^{d_z}$ such that

$$P(Z_{n+1} \in B | \theta_0, Z_0, \ldots, \theta_n, Z_n) = \Pi_{\theta_n}(Z_n, B) \tag{12}$$

w.p.1 for any Borel-measurable set $B \subseteq \mathbb{R}^{d_z}$ and $n \geq 0$. In the context of stochastic gradient search, $F(\theta_n, Z_{n+1}) + \rho_n$ represents a gradient estimator (i.e., an estimator of $\nabla f(\theta_n)$).

The algorithm (11) is analyzed under the following assumptions.

Assumption 3.1: $\sum_{n=0}^{\infty} \alpha_n = \infty$, $\sum_{n=0}^{\infty} \alpha_n^2 < \infty$ and $\sum_{n=0}^{\infty} |\alpha_n - \alpha_{n+1}| < \infty$.

Assumption 3.2: There exist a differentiable function $f : \mathbb{R}^{d_\theta} \to \mathbb{R}$ and a Borel-measurable function $\tilde{F} : \mathbb{R}^{d_\theta} \times \mathbb{R}^{d_z} \to \mathbb{R}^{d_\theta}$ such that $\nabla f(\cdot)$ is locally Lipschitz continuous and

$$F(\theta, z) - \nabla f(\theta) = \tilde{F}(\theta, z) - (\Pi \tilde{F})(\theta, z) \tag{13}$$

for each $\theta \in \mathbb{R}^{d_\theta}$, $z \in \mathbb{R}^{d_z}$, where $(\Pi \tilde{F})(\theta, z) = \int \tilde{F}(\theta, z') \Pi_\theta(z, dz')$.

Assumption 3.3: For any compact set $Q \subset \mathbb{R}^{d_\theta}$, there exists a Borel-measurable function $\varphi_Q : \mathbb{R}^{d_z} \to [1, \infty)$ such that

$$\begin{aligned}
&\max\{\|F(\theta, z)\|, \|\tilde{F}(\theta, z)\|, \|(\Pi \tilde{F})(\theta, z)\|\} \leq \varphi_Q(z), \\
&\|(\Pi \tilde{F})(\theta', z) - (\Pi \tilde{F})(\theta'', z)\| \leq \varphi_Q(z) \|\theta' - \theta''\|, \\
&\sup_{n \geq 0} E\left(\varphi_Q^2(Z_{n+1}) I_{\{\tau_Q > n\}} | \theta_0 = \theta, Z_0 = z\right) < \infty
\end{aligned}$$

for all $\theta, \theta', \theta'' \in Q$, $z \in \mathbb{R}^{d_z}$, where $\tau_Q = \inf\{n \geq 0 : \theta_n \notin Q\}$.

Assumption 3.4: $\limsup_{n \to \infty} \|\rho_n\| < \infty$ w.p.1 on $\{\sup_{n \geq 0} \|\theta_n\| < \infty\}$.

The main results on the asymptotic bias of the recursion (11) reads as follows.

Theorem 3.2: Let Assumptions 3.1 – 3.4 hold, and suppose that $f(\cdot)$ (introduced in Assumption 3.2) satisfies Assumption 2.3. Then, all conclusions of Theorem 2.1 are true.

The proof of Theorem 3.2 is provided in [31].

Assumption 3.1 is related to the sequence $\{\alpha_n\}_{n \geq 0}$. It holds if $\alpha_n = 1/n^a$ for $n \geq 1$, where $a \in (1/2, 1]$ is a constant. Assumptions 3.2 – 3.4 correspond to the stochastic process $\{Z_n\}_{n \geq 0}$ and are common for the asymptotic analysis of stochastic approximation algorithms with Markovian dynamics. Assumptions 3.2 – 3.4 have been introduced by Metivier and Priouret and later generalized by Kushner, Yin and their co-workers (see [1, Part II], [17] and references cited therein). However, neither the results of Metivier and Priouret, nor the results of Kushner, Yin and their co-workers provide any information on the asymptotic bias of the gradient search with biased gradient estimation.

Regarding Theorem 3.2, the following note is in order. As already mentioned in the beginning of the section, the purpose of the theorem is illustrating the results of Theorem 2.1 and providing a framework for studying the examples presented in the next sections. Since these examples perfectly fit into the framework developed by Metivier and Priouret, more general assumptions and settings of [17] are not considered here in order to keep the exposition as concise as possible.

IV. Example 1: Actor-Critic Reinforcement Learning

In this section, we apply Theorems 2.1, 3.2 and Corollary 2.1 to the asymptotic analysis of actor-critic algorithms for Markov decision processes.

To define an average-cost Markov decision process with a parameterized randomized control, we need the following notation. $d_\theta \geq 1$, $N_x > 1$ and $N_y > 1$ are integers, while $\mathcal{X} = \{1, \ldots, N_x\}$ and $\mathcal{Y} = \{1, \ldots, N_y\}$. $\phi(x, y)$ is a non-negative (real-valued) function of $(x, y) \in \mathcal{X} \times \mathcal{Y}$. $p(x'|x, y)$ is a non-negative function of $(x, x', y) \in \mathcal{X} \times \mathcal{X} \times \mathcal{Y}$ satisfying $\sum_{x' \in \mathcal{X}} p(x'|x, y) = 1$ for each $x \in \mathcal{X}$, $y \in \mathcal{Y}$. $q_\theta(y|x)$ is a non-negative function of $(\theta, x, y) \in \mathbb{R}^{d_\theta} \times \mathcal{X} \times \mathcal{Y}$ with the following properties: It is differentiable in θ for any $x \in \mathcal{X}$, $y \in \mathcal{Y}$ and fulfills $\sum_{y \in \mathcal{Y}} q_\theta(y|x) = 1$ for all $\theta \in \mathbb{R}^{d_\theta}$, $x \in \mathcal{X}$. For $\theta \in \mathbb{R}^{d_\theta}$, $\{(X_n^\theta, Y_n^\theta)\}_{n \geq 0}$ is an $\mathcal{X} \times \mathcal{Y}$-valued Markov chain which is defined on a (canonical) probability space $(\Omega, \mathcal{F}, P_\theta)$ and which admits

$$\begin{aligned}
&P_\theta(X_{n+1}^\theta = x', Y_{n+1}^\theta = y' | X_n^\theta = x, Y_n^\theta = y) \\
&= q_\theta(y'|x') p(x'|x, y)
\end{aligned}$$

for each $x, x' \in \mathcal{X}$, $y, y' \in \mathcal{Y}$. $f(\cdot)$ is a function defined by

$$f(\theta) = \lim_{n \to \infty} E_\theta(\phi(X_n^\theta, Y_n^\theta))$$

for $\theta \in \mathbb{R}^{d_\theta}$. With this notation, an average-cost Markov decision problem with parameterized randomized control can be defined as the minimization of $f(\cdot)$. In the literature on reinforcement learning and operations research, $\{X_n^\theta\}_{n \geq 0}$ is called a controlled Markov chain, while $\{Y_n^\theta\}_{n \geq 0}$ are control actions. $\{p(x'|x, y)\}_{x, x' \in \mathcal{X}, y \in \mathcal{Y}}$ are the chain transition probabilities, while $\{q_\theta(y|x)\}_{x \in \mathcal{X}, y \in \mathcal{Y}}$ are the action likelihoods. θ is a parameter indexing the action likelihoods. For further details on Markov decision processes, see [3] and references cited therein.

Since $f(\cdot)$ and its gradient rarely admit a close-form expression, $f(\cdot)$ is minimized using methods based on stochastic gradient search and Monte Carlo gradient estimation. One of the most sophisticated methods of this kind is the actor-critic algorithm proposed by Konda and Tsitsiklis in [16]. This algorithm is defined by the following equations:

$$V_{n+1} = \lambda V_n + s_{\theta_n}(Y_{n+1}|X_{n+1}), \tag{14}$$

$$\begin{aligned}
W_{n+1} =& (s_{\theta_n}(Y_{n+1}|X_{n+1}) - s_{\theta_n}^T(Y_n|X_n))^T \eta_n' \\
&+ c(X_n, Y_n) - \eta_n'',
\end{aligned} \tag{15}$$

$$\theta_{n+1} = \theta_n - \alpha_n s_{\theta_n}(Y_{n+1}|X_{n+1}) s_{\theta_n}^T(Y_{n+1}|X_{n+1}) \eta_n', \tag{16}$$

$$\eta_{n+1}' = \eta_n' + \beta_n V_{n+1} W_{n+1}, \tag{17}$$

$$\eta_{n+1}'' = \eta_n'' + \beta_n(\phi(X_{n+1}, Y_{n+1}) - \eta_n''), \quad n \geq 0. \tag{18}$$

In this recursion, $\lambda \in [0, 1)$ is a constant (usually referred to as the discounting factor), while $\{\alpha_n\}_{n \geq 0}$ and $\{\beta_n\}_{n \geq 0}$ are sequences of positive real numbers. Function $s_\theta(y|x)$ is defined by

$$s_\theta(y|x) = \nabla_\theta \log(q_\theta(y|x))$$

for $\theta \in \mathbb{R}^{d_\theta}$, $x \in \mathcal{X}$, $y \in \mathcal{Y}$. θ_0, η_0', V_0 are $\mathbb{R}^{d_\theta}$-valued random variables, while η_0'', W_0 are $\mathbb{R}$-valued random variables. $\{X_n\}_{n\geq 1}$ and $\{Y_n\}_{n\geq 1}$ are (respectively) $\mathcal{X}$ and $\mathcal{Y}$ valued stochastic processes which are generated through Monte Carlo simulations: For each $n \geq 0$, X_{n+1} is simulated from $p(\cdot|X_n, Y_n)$ independently of θ_0, η_0', η_0'', V_0, W_0, $X_0, Y_0, \ldots, X_n, Y_n$, while Y_{n+1} is simulated from $q_{\theta_n}(\cdot|X_{n+1})$ independently of θ_0, η_0', η_0'', V_0, W_0, $X_0, Y_0, \ldots, X_n, Y_n, X_{n+1}$. Hence, $\{(X_n, Y_n)\}_{n\geq 1}$ satisfies

$$P(X_{n+1} = x, Y_{n+1} = y|\theta_0, X_0, Y_0, \ldots, \theta_n, X_n, Y_n) = q_{\theta_n}(y|x)p(x|X_n, Y_n)$$

w.p.1 for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$, $n \geq 1$.

Remark 4.2: In algorithm (14) – (18), recursion (16) is a gradient search minimizing $f(\cdot)$, while term

$$s_{\theta_n}(Y_{n+1}|X_{n+1})s_{\theta_n}^T(Y_{n+1}|X_{n+1})\eta_n'$$

is a Monte Carlo estimator of $\nabla f(\theta_n)$. This estimator is biased and its bias is proportional to $1 - \lambda$ (see [16]).

Algorithm (14) – (18) is analyzed under the following assumptions.

Assumption 4.1: $\lim_{n\to\infty} \alpha_n/\beta_n = 0$, $\sum_{n=0}^{\infty} \beta_n = \infty$, $\sum_{n=0}^{\infty} \beta_n^2 < \infty$ and $\sum_{n=0}^{\infty} |\beta_n - \beta_{n+1}| < \infty$.

Assumption 4.2: $p(x'|x, y) > 0$ for each $x, x' \in \mathcal{X}$, $y \in \mathcal{Y}$.

Assumption 4.3: For each $x \in \mathcal{X}$, $y \in \mathcal{Y}$, $s_\theta(y|x)$ is locally Lipschitz continuous in θ on Θ.

To state the main results of this section, we need some further notation. δ is a real numbers defined as

$$\delta = \min\{p(x'|x, y) : x, x' \in \mathcal{X}, y \in \mathcal{Y}\}$$

(obviously $0 < \delta < 1$). For a compact set $Q \subset \mathbb{R}^{d_\theta}$ and fixed $x \in \mathcal{X}, y \in \mathcal{Y}$, $L_{1,Q}(x, y)$ is an upper bound in $\theta \in Q$ for $\|s_\theta(y|x)\|$, while $L_{2,Q}$ is a Lipschitz constant in $\theta \in Q$ for $q_\theta(y|x)$, $s_\theta(y|x)$. For the same Q, let

$$L_Q = \max\{|\phi(x, y)|, L_{1,Q}(x, y), L_{2,Q}(x, y) : x \in \mathcal{X}, y \in \mathcal{Y}\}.$$

With this notation, our results on the analytical properties of $f(\cdot)$ and asymptotic behavior of algorithm (14) read as follows.

Proposition 4.1: Let Assumption 4.2 hold.

(i) Suppose that $q_\theta(y|x)$ is q times differentiable in θ for each $\theta \in \mathbb{R}^{d_\theta}$, $x \in \mathcal{X}$, $y \in \mathcal{Y}$. Then, $f(\cdot)$ is q times differentiable on $\mathbb{R}^{d_\theta}$. If additionally $d_\theta < q < \infty$, then, for any compact set $Q \subset \mathbb{R}^{d_\theta}$, there exists a real number $M_Q \in [1, \infty)$ such that (4) holds for $p = (q - d_\theta)/(q - 1)$ and all $\varepsilon \in [0, \infty)$.

(ii) Suppose that $q_\theta(y|x)$ is real-analytic in θ for each $\theta \in \mathbb{R}^{d_\theta}$, $x \in \mathcal{X}$, $y \in \mathcal{Y}$. Then, $f(\cdot)$ is real-analytic on $\mathbb{R}^{d_\theta}$. Moreover, for any compact set $Q \subset \mathbb{R}^{d_\theta}$, there exist real numbers $r_Q \in (0, 1)$ and $M_Q, N_Q \in [1, \infty)$ such that (5) holds for all $\varepsilon \in [0, \infty)$, $\theta \in Q$.

Theorem 4.3: Let Assumptions 3.1, 4.2 and 4.3 hold.

(i) Suppose that $q_\theta(y|x)$ is q times differentiable in θ for each $\theta \in \mathbb{R}^{d_\theta}$, $x \in \mathcal{X}$, $y \in \mathcal{Y}$, where $d_\theta < q < \infty$. Let $p = (q - d_\theta)/(q - 1)$. Then, for any compact set $Q \subset \mathbb{R}^{d_\theta}$, there exists a real number $K_Q \in [1, \infty)$ (depending only on δ, L_Q, M_Q) such that

$$\limsup_{n\to\infty} \|\nabla f(\theta_n)\| \leq K_Q(1 - \lambda)^{p/2},$$
$$\limsup_{n\to\infty} f(\theta_n) - \liminf_{n\to\infty} f(\theta_n) \leq K_Q(1 - \lambda)^p$$

w.p.1 on Λ_Q.

(ii) Suppose that $q_\theta(y|x)$ is real-analytic in θ for each $\theta \in \mathbb{R}^{d_\theta}$, $x \in \mathcal{X}$, $y \in \mathcal{Y}$. Then, all conclusions of Part (i) hold with $p = 1$. Moreover, for any compact set $Q \subset \mathbb{R}^{d_\theta}$, there exists a real number $K_Q \in [1, \infty)$ (depending only on δ, L_Q, M_Q, N_Q) such that

$$\limsup_{n\to\infty} d(\theta_n, S) \leq K_Q(1 - \lambda)^{r_Q/2}$$

w.p.1 on Λ_Q.

The proof of Proposition 4.1 and Theorem 4.3 is provided in [31].

Assumptions 4.1 corresponds to the asymptotic properties of sequence $\{\beta_n\}_{n\geq 0}$. It is satisfied if $\beta_n = 1/n^b$ and $\alpha_n = 1/n^a$ for $n \geq 1$ and if $1/2 < b < a \leq 1$. Assumption 4.2 is related to the stability of the controlled Markov chain $\{X_n^\theta\}_{n\geq 0}$ and is often used in the asymptotic analysis of reinforcement learning algorithms (see e.g., [3]). Assumption 4.3 corresponds to the parameterization of the action likelihoods $q_\theta(y|x)$ and is almost always met in practice. For some commonly used parameterizations (such as exponential and trigonometric), $q_\theta(y|x)$ is not only Lipschitz continuously differentiable in θ, but also real-analytic.

Although actor-critic algorithms are widely used in reinforcement learning, the available literature does not give a quite satisfactory answer to the problem of their asymptotic behavior. To the best of the present author's knowledge, the existing results do not even guarantee that the asymptotic bias of recursion (14), (18) goes to zero as λ tends to one, More specifically, none of the existing results guarantees that $\{\theta_n\}_{n\geq 0}$ converges to a vicinity of $S = \{\theta : \nabla f(\theta) = 0\}$ and that the radius of the vicinity tends to zero as λ approaches one (e.g., [16], probably the strongest results of the kind, claims this only for a subsequence of $\{\theta_n\}_{n\geq 0}$). The main difficulty stems from the fact that actor-critic algorithms are so complex that the existing asymptotic results for biased stochastic gradient search and biased stochastic approximation [5, Section 5.3], [8], [9], [10, Section 2.7] cannot be applied. Relying on the results presented in Sections II and III, Theorem 4.3 overcomes these difficulties: Under mild and easily verifiable conditions, Theorem 4.3 not only guarantees that the asymptotic bias of algorithm (14), (18) converges to zero as λ tends to one, but also provides a relatively tight polynomial bound on the rate in terms of λ and the Yomdin and Lojasiewicz exponents.

APPENDIX
STABILITY OF STOCHASTIC GRADIENT SEARCH WITH MARKOVIAN DYNAMICS

In order to obtain a 'global' version of the results of Sections II and III, the almost sure stability of the stochastic

gradient search with random truncations and Markovian dynamics is studied here. Such an algorithm is defined by the following equations:

$$\theta'_{n+1} = \theta_n - \alpha_n(F(\theta_n, Z_{n+1}) + \rho_n), \tag{19}$$

$$\theta_{n+1} = \theta'_{n+1}\, I_{\{\|\theta'_{n+1}\| \leq \beta_{\sigma_n}\}} + \vartheta_0\, I_{\{\|\theta'_{n+1}\| > \beta_{\sigma_n}\}}, \tag{20}$$

$$\sigma_{n+1} = \sigma_n + I_{\{\|\theta'_{n+1}\| > \beta_{\sigma_n}\}}. \tag{21}$$

Here, $F(\cdot,\cdot)$, $\{\alpha_n\}_{n\geq 0}$, $\{Z_n\}_{n\geq 0}$ and $\{\rho_n\}_{n\geq 0}$ have the same meaning as in Section III. $\vartheta_0 \in R^{d_\theta}$ is a (deterministic) vector, while $\{\beta_n\}_{n\geq 0}$ is an increasing sequence of positive real numbers satisfying $\|\vartheta_0\| < \beta_0$ and $\lim_{n\to\infty} \beta_n = \infty$. θ_0 is an R^{d_θ}-valued random variable fulfilling $\|\theta_0\| < \beta_0$, while $\sigma_0 = 0$. The random truncation scheme has been proposed and analyzed in [8], [9], [10].

The stability of the algorithm (19) – (21) is analyzed under the following assumptions.

Assumption A.1: $\rho_n = g(\theta_n)$ for each $n \geq 0$, where $g : R^{d_\theta} \to R^{d_\theta}$ is a Borel-measurable locally bounded function.

Assumption A.2: $\liminf_{\|\theta\|\to\infty} f(\theta) = \infty$ and $\inf_{\theta\in\mathbb{R}^{d_\theta}} f(\theta) > -\infty$. Moreover, there exist real numbers $c \in (0,1)$, $r \in [1,\infty)$ such that

$$\|\nabla f(\theta)\| \geq c, \qquad \frac{\|g(\theta)\|}{\|\nabla f(\theta)\|} \leq c$$

for all $\theta \in \mathbb{R}^{d_\theta}$ satisfying $\|\theta\| \geq r$.

Assumption A.3: Given any compact set $Q \subset \mathbb{R}^{d_\theta}$,

$$\sup_{n\geq 0} E\left(\varphi_Q^2(Z_n)\middle|\, \theta_0 = \theta, Z_0 = z\right) < \infty$$

for all $\theta \in \mathbb{R}^{d_\theta}$, $z \in \mathbb{R}^{d_z}$.

The next proposition contains the main results on the stability of the algorithm (19) – (21).

Proposition 1.2: Let $\{\theta_n\}_{n\geq 0}$ be generated by recursion (19) – (21). Moreover, let Assumptions 3.1 – 3.3, A.1, A.2 and A.3 hold, and suppose that $f(\cdot)$ (introduced in Assumption 3.2) satisfies Assumption 2.3. Then, the following is true:

(i) There exists a real number $a \in [1,\infty)$ (depending only on r, $f(\cdot)$) such that $\limsup_{n\to\infty} \|\theta_n\| < a$ w.p.1.

(ii) There exists a real number $K \in [1,\infty)$ (depending only on L_{Q_a} and M_{Q_a}, where $Q_a = \{\theta \in \mathbb{R}^{d_\theta} : \|\theta\| \leq a\}$) such that

$$\limsup_{n\to\infty} \|\nabla f(\theta_n)\| \leq K\rho^{p/2},$$

$$\limsup_{n\to\infty} f(\theta_n) - \liminf_{n\to\infty} f(\theta_n) \leq K\rho^p$$

w.p.1.

Proposition 1.2 is just an extension of the results of [10] and [28]. Its proof is provided in [31].

REFERENCES

[1] A. Benveniste, M. Metivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximations,* Springer-Verlag, 1990.

[2] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming,* Athena Scientific, 1996.

[3] D. P. Bertsekas and J. N. Tsitsiklis, *Gradient convergence in gradient methods with errors,* SIAM Journal on Optimization, 10 (2000), pp. 627 – 642.

[4] V. S. Borkar and S. P. Meyn, *The ODE Method for Convergence of Stochastic Approximation and Reinforcement Learning,* SIAM Journal on Control and Optimization, 38 (2000), pp. 447 – 469.

[5] V. S. Borkar, *Stochastic Approximation: A Dynamical Systems Viewpoint,* Cambridge University Press, 2009.

[6] X.-R. Cao, *Stochastic Learning and Optimization,* Springer-Verlag, 2007.

[7] O. Cappe, E. Moulines, and T. Ryden, *Inference in Hidden Markov Models,* Springer-Verlag, 2005.

[8] H.-F. Chen, L. Guo, A.-J. Gao, *Convergence and robustness of the Robbins-Monro algorithm truncated at randomly varying bounds,* Stochastic Processes and Their Applications, 27 (1988), pp. 217 – 231.

[9] H.-F. Chen, A.-J. Gao, *Robustness analysis of stochastic approximation algorithms,* Stochastics and Stochastics Reports, 26 (1989), pp. 3 – 20.

[10] H.-F. Chen, *Stochastic Approximation and Its Application,* Kluwer, 2002.

[11] C. G. Cassandras and S. Lafortune, *Introduction to Discrete Event Systems,* 2nd Edition, Springer-Verlag, 2008.

[12] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications,* Wiley, 2002.

[13] A. Doucet, N. de Freitas and N. Gordon (Eds.), *Sequential Monte Carlo Methods in Practice,* Springer-Verlag, 2001.

[14] A. Doucet and V. B. Tadić, *Parameter estimation in general state-space models using particle methods,* Annals of the Institute of Statistical Mathematics, 55 (2003), pp. 409 – 422.

[15] S. G. Henderson, S. P. Meyn and V. B. Tadić, *Performance evaluation and policy selection in multiclass networks,* Discrete Event Dynamic Systems, 13 (2003), pp. 149 – 189.

[16] V. R. Konda and J. N. Tsitsiklis, *On actor-critic algorithms,* SIAM Journal on Control and Optimization, 42 (2003), pp. 1143 – 1166.

[17] H. J. Kushner and G. G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications,* 2nd edition, Springer-Verlag, 2003.

[18] L. Ljung, *System Identification: Theory for the User,* 2nd edition, Prentice Hall, 1999.

[19] S. Lojasiewicz, *Sur le problème de la division,* Studia Mathematica, 18 (1959), pp. 87 – 136.

[20] S. Lojasiewicz, *Sur la géométrie semi- et sous-analytique,* Annales de l'Institut Fourier (Grenoble), 43 (1993), pp. 1575 – 1595.

[21] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability,* 2nd Edition, Cambridge University Press, 2009.

[22] G. Ch. Pflug, *Optimization of Stochastic Models: The Interface Between Simulation and Optimization,* Kluwer 1996.

[23] G. Poyiadjis, A. Doucet and S. S. Singh, *Particle approximations of the score and observed information matrix in state-space models with applications to parameter estimation,* Biometrika 98 (2011), pp 65 – 80.

[24] B. T. Polyak and Y. Z. Tsypkin, *Criterion algorithms of stochastic optimization,* Automation and Remote Control, 45 (1984), pp. 766 – 774.

[25] B. T. Polyak, *Introduction to Optimization,* Optimization Software, 1987.

[26] T. Ryden, *On recursive estimation for hidden Markov models,* Stochastic Processes and Their Applications 66 (1997), pp. 79 – 96.

[27] J. C. Spall, *Introduction to Stochastic Search and Optimization,* Wiley, 2003.

[28] V. B. Tadić, *Stochastic approximation with random truncations, state-dependent noise and discontinuous dynamics,* Stochastics and Stochastics Reports, 64 (1998), pp. 283 – 326.

[29] V. B. Tadić and A. Doucet, *Exponential forgetting and geometric ergodicity for optimal filtering in general state-space models,* Stochastic Processes and Their Applications, 115 (2005), pp. 1408–1436.

[30] V. B. Tadić, *Analyticity, Convergence and Convergence Rate of Recursive Maximum Likelihood Estimation in Hidden Markov Models,* IEEE Transactions on Information Theory, 56 (2008), pp. 6406–6432.

[31] V. B. Tadić and A. Doucet, *Asymptotic Bias of Stochastic Gradient Search,* extended version of this paper.

[32] V. B. Tadić and A. Doucet, *Sequential Monte Carlo Methods for Identification of State-Space Models: Asymptotic Analysis,* in preparation.

[33] Y. Yomdin, *The Geometry of Critical and Near Critical Values of Differentiable Mappings,* Mathematische Annalen, 264 (1983), pp. 495 – 515.