Explicit Mean-Square Error Bounds for Monte-Carlo and Linear Stochastic Approximation

Shuhang Chen
University of Florida

Adithya M. Devraj
University of Florida

Ana Bušić
Inria and DI ENS

Sean Meyn
University of Florida

Abstract

This paper concerns error bounds for recursive equations subject to Markovian disturbances. Motivating examples abound within the fields of Markov chain Monte Carlo (MCMC) and Reinforcement Learning (RL), and many of these algorithms can be interpreted as special cases of stochastic approximation (SA). It is argued that it is not possible in general to obtain a Hoeffding bound on the error sequence, even when the underlying Markov chain is reversible and geometrically ergodic, such as the M/M/1 queue. This is motivation for the focus on mean square error bounds for parameter estimates. It is shown that mean square error achieves the optimal rate of $O(1/n)$, subject to conditions on the step-size sequence. Moreover, the exact constants in the rate are obtained, which is of great value in algorithm design.

1 Introduction

Many questions in statistics and the area of reinforcement learning (RL) are concerned with computation of the root of a function in the form of an expectation: $\bar{f}(\theta) = \mathsf{E}[f(\theta, \Phi)]$, where Φ is a vector valued random variable, and $\theta \in \mathbb{R}^d$. The value θ^* satisfying $\bar{f}(\theta^*) = 0$ is most commonly approximated through some version of the stochastic approximation (SA) algorithm (Robbins and Monro, 1951; Borkar, 2008). In its basic form, this is the recursive algorithm

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

in which $\{\alpha_n\}$ is a non-negative gain sequence, and $\{\Phi_n\}$ is a sequence of random variables whose distribution converges to that of Φ as $n \to \infty$. The sequence is a Markov chain in the applications of interest in this paper.

There is a large body of work on conditions for convergence of this recursion, and also a Central Limit Theorem (CLT): with $\widetilde{\theta}_n = \theta_n - \theta^*$,

$$\lim_{n\to\infty} \widetilde{\theta}_n = 0 \qquad \textit{almost surely}$$

$$\lim_{n\to\infty} \sqrt{n}\widetilde{\theta}_n = N(0, \Sigma_\theta) \qquad \textit{in distribution}$$

The $d \times d$ matrix Σ_θ is known as the *asymptotic covariance* (Benveniste et al., 2012; Kushner and Yin, 1997).

Soon after the stochastic approximation algorithm was first introduced in Robbins and Monro (1951); Blum (1954). Chung et al. (1954) identified the optimal CLT covariance and techniques to obtain the optimum for scalar recursions. This can be cast as a form of *stochastic Newton-Raphson* (SNR) (Devraj and Meyn, 2017a,b; Devraj et al., 2019; Devraj, 2019). Gradient free methods [or *stochastic quasi Newton-Raphson* (SQNR)] appeared in later work: The first example was proposed in Venter et al. (1967), which was shown to obtain the optimal variance for a one-dimensional SA recursion. The algorithm obtains estimates of the SNR gain $-A^{-1}$ (see (2) below), through a procedure similar to Kiefer and Wolfowitz (1952). Ruppert (1985)

Proceedings of the 23rdInternational Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

proposed an extension of Venter's algorithm for vector-valued functions .

The averaging technique of Ruppert and Polyak is a two-time-scale algorithm that is also designed to achieve the optimal asymptotic covariance (Ruppert, 1988; Polyak, 1990; Polyak and Juditsky, 1992). A two-time-scale variant of the SNR algorithm known as "Zap-SNR" was proposed in Devraj and Meyn (2017a,b); Devraj et al. (2019); Devraj (2019), with applications to RL. Zap algorithms are stable and convergent under mild assumptions (Devraj and Meyn, 2017a; Chen et al., 2019a).

The asymptotic covariance in the CLT for the recursion (1) has an explicit form under general conditions (Kushner and Yin, 2003, Chapter 10, Theorem 3.3). Assuming that the root θ^* is unique, denote

$$A = \partial \bar{f}\,(\theta^*)\,, \qquad \Delta_n = f(\theta^*, \Phi_n) \tag{2}$$

and consider the linear approximation:

$$\widetilde{\theta}_{n+1} = \widetilde{\theta}_n + \alpha_{n+1}[A\widetilde{\theta}_n + \Delta_{n+1}]\,. \tag{3}$$

Subject to the assumption that $\frac{1}{2}I + A$ is Hurwitz (i.e., $\text{Real}(\lambda) < -\frac{1}{2}$ for each eigenvalue of A), the $d \times d$ matrix Σ_θ is the unique positive semi-definite solution to the Lyapunov equation

$$[\tfrac{1}{2}I + A]\Sigma + \Sigma[\tfrac{1}{2}I + A]^\intercal + \Sigma_\Delta = 0 \tag{4}$$

in which Σ_Δ is also an asymptotic covariance: the covariance matrix appearing in the CLT for the sequence $\{\Delta_n\}$ (which may be expressed in terms of a solution to a Poisson equation - see Kushner and Yin (2003, Chapter 10, Theorem 2.2)).

The goal of this paper is to demonstrate that the CLT is far less *asymptotic* than it may appear. For this we focus analysis on the linearization (3), along with first steps towards analysis of the nonlinear recursions. Subject to assumptions on A and the Markov chain, we establish the bound

$$\text{Cov}\,(\theta_n) = n^{-1}\Sigma_\theta + O(n^{-1-\delta}) \tag{5}$$

with $\text{Cov}\,(\theta_n) = \mathsf{E}[\widetilde{\theta}_n\widetilde{\theta}_n^\intercal]$. The bound is refined under further assumptions:

$$\text{Cov}\,(\theta_n) = n^{-1}\Sigma_\theta + n^{-2}\Sigma_{\theta,2} + O(n^{-2-\delta}) \tag{6}$$

where $\delta > 0$, and $\Sigma_{\theta,2}$ solves a second Lyapunov equation based on a second Poisson equation.

It is hoped that these results will be helpful in construction and performance analysis of algorithms in machine learning, statistics and RL.

The reader may ask, why not search directly for a version of Hoeffding's inequality:

$$\mathsf{P}\{\|\widetilde{\theta}_n\| \ge \varepsilon\} \le b_0 \exp(-nI_0(\varepsilon)) \tag{7}$$

where $b_0 > 0$ is fixed, and I_0 is a convex function that is strictly positive and finite in a region $0 < \varepsilon^2 \le \bar{\varepsilon}^2$. The answer is that such bounds are not always possible even for the simplest SA recursions, even when the Markov chain is geometrically ergodic. This is clarified in the first general example:

Markov Chain Monte Carlo As a prototypical example of stochastic approximation, Markov chain Monte Carlo (MCMC) proceeds by constructing an ergodic Markov chain $\mathbf{\Phi}$ with invariant measure π so as to estimate $\pi(F) = \int F(z)\,\pi(dz)$ for some function $F : \mathsf{Z} \to \mathbb{R}^d$ via,

$$\theta_{n+1} = \frac{1}{n+1}\sum_{k=1}^{n+1} F(\Phi_k) \tag{8}$$

This is an instance of the SA recursion (1):

$$\theta_{n+1} = \theta_n + \frac{1}{n+1}(-\theta_n + F(\Phi_{n+1})) \tag{9}$$

Subtracting $\theta^* = \pi(F)$ from both sides of (9) gives, with $\widetilde{\theta}_n = \theta_n - \pi(F)$,

$$\widetilde{\theta}_{n+1} = \widetilde{\theta}_n + \frac{1}{n+1}(-\widetilde{\theta}_n + F(\Phi_{n+1}) - \pi(F))$$

which is (3) in a special case: $A = -I$, $\Delta_{n+1} = F(\Phi_{n+1}) - \pi(F)$ and $\alpha_n = 1/n$.

A significant part of the literature on MCMC focuses on finding Markov chains whose marginals approach the invariant measure π quickly. Error estimates for MCMC have only been studied under rather restrictive settings. For instance, under the assumption of uniform ergodicity of $\mathbf{\Phi}$ and uniform boundedness of F (which rarely hold in practice outside of a finite state space), a generalized Hoeffding's inequality was obtained in

Glynn and Ormoneit (2002) to obtain the PAC-style error bound (7). We can not expect Hoeffding's bound if either of these assumptions is relaxed. Consider the simplest countable state space Markov chain: the M/M/1 queue with uniformization, defined with $\mathsf{Z} = \{0, 1, 2, \dots\}$ and

$$\Phi_{n+1} = \begin{cases} \Phi_n + 1 & \text{prob. } \alpha \\ \max(\Phi_n - 1, 0) & \text{prob. } \mu = 1 - \alpha \end{cases}$$

This is a reversible, geometrically ergodic Markov chain when $\rho = \alpha/\mu < 1$, with geometric invariant measure. It is shown in Meyn (2007) that the error bound (7) fails for most unbounded functions F. The question is looked at in greater depth in Duffy and Meyn (2010, 2014), where asymptotic bounds are obtained for the special case $F(z) \equiv z$. An asymptotic version of (7) is obtained for the lower tail:

$$\lim_{n\to\infty} \frac{1}{n} \log\Big(\mathsf{P}\{\widetilde{\theta}_n \le -\varepsilon\}\Big) = -I_0(\varepsilon) \tag{10}$$

A different scaling is required for the upper tail:

$$\lim_{n\to\infty} \frac{1}{n} \log\Big(\mathsf{P}\Big\{\frac{\widetilde{\theta}_n}{n} \ge \varepsilon\Big\}\Big) = -J_0(\varepsilon) \tag{11}$$

The functions I_0 and J_0 are strictly positive and finite valued in some domain $0 < \varepsilon < \varepsilon_0$. Consequently, the finite-n bound (7) is not attainable for this simple Markov model.

Reinforcement Learning The theory of this paper also applies to TD-learning. Consider a Markov chain $\boldsymbol{X}$ evolving on a (Polish) state space X. Given a cost function $c : \mathsf{X} \to \mathbb{R}$, and a discount factor $\beta \in (0, 1)$, the goal in TD-learning is to approximate the solution $h : \mathsf{X} \to \mathbb{R}$ to the Bellman equation:

$$h(x) = c(x) + \beta \mathsf{E}[h(X_{n+1}) \mid X_n = x] \tag{12}$$

With $\Phi_{n+1} := (X_{n+1}, X_n)$, this becomes

$$\mathsf{E}[\mathcal{D}(h, \Phi_{n+1}) \mid \Phi_0 \dots \Phi_n] = 0 \tag{13}$$

where $\mathcal{D}(h, \Phi_{n+1}) := c(X_n) + \beta h(X_{n+1}) - h(X_n)$. Equation (13) may be regarded as motivation for the TD-learning algorithms of Sutton (1988); Tsitsiklis and Van Roy (1997).

Consider a linearly parameterized family of candidate approximations $\{h^\theta(x) = \theta^\intercal \psi(x) : \theta \in \mathbb{R}^d\}$, where $\psi : \mathsf{X} \to \mathbb{R}^d$ denotes the d basis functions. The goal in TD-learning is to solve the *Galerkin relaxation* of (13):

$$\mathsf{E}[\mathcal{D}(h^{\theta^*}, \Phi_{n+1})\zeta_n] = 0 \tag{14}$$

where $\{\zeta_n\}$ is a d-dimensional stochastic process, adapted to $\boldsymbol{\Phi}$, and the expectation is with respect to the steady state distribution. The TD(0) algorithm is the SA recursion (1) applied to solve (14) with $\zeta_n \equiv \psi(X_n)$:

$$\begin{aligned} \theta_{n+1} &= \theta_n + \alpha_{n+1} d_{n+1} \psi(X_n) \\ d_{n+1} &= c(X_n) + \beta h^{\theta_n}(X_{n+1}) - h^{\theta_n}(X_n) \end{aligned} \tag{15}$$

Denoting

$$\begin{aligned} A_{n+1} &:= \psi(X_n)\big(\beta\psi(X_{n+1}) - \psi(X_n)\big)^\intercal \\ b_{n+1} &:= -c(X_n)\psi(X_n) \end{aligned}$$

the algorithm (15) can be rewritten as:

$$\theta_{n+1} = \theta_n + \alpha_{n+1}\big(A_{n+1}\theta_n - b_{n+1}\big) \tag{16}$$

Note that θ^* from (14) solves the linear equation $\mathsf{E}[A_{n+1}]\theta^* = \mathsf{E}[b_{n+1}]$. Subtracting θ^* from both sides of (16) gives, with $\widetilde{\theta}_n = \theta_n - \theta^*$,

$$\widetilde{\theta}_{n+1} = \widetilde{\theta}_n + \alpha_{n+1}[A\widetilde{\theta}_n + \tilde{A}_{n+1}\widetilde{\theta}_n + \Delta_{n+1}] \tag{17}$$

where $\Delta_{n+1} = A_{n+1}\theta^* - b_{n+1}$, $\tilde{A}_{n+1} = A_{n+1} - A$. We show through coupling that (17) and (3) have the same asymptotic covariance if the matrix $\frac{1}{2}I + A$ is Hurwitz (see Thm. 2.5).

The matrix $A = \mathsf{E}[A_{n+1}]$ is Hurwitz under general conditions, and the sequence of estimates $\{\theta_n\}$ converges to θ^* (Tsitsiklis and Van Roy, 1997). However, when the discount factor β is close to 1, it can be shown that $\lambda_{\max} > -\frac{1}{2}$ (where $\lambda_{\max}$ denotes the largest eigenvalue of A), and is in fact close to 0 under mild additional assumptions (Devraj and Meyn, 2017a; Devraj, 2019; Devraj et al., 2020). Full details and finer results are presented in Theorems 2.4, 2.6 and the discussion that follows.

The SNR algorithm is defined as follows:

$$\theta_{n+1} = \theta_n - \alpha_{n+1} d_{n+1} \widehat{A}_{n+1}^{-1} \psi(X_n) \tag{18}$$

$$\widehat{A}_{n+1} = \widehat{A}_n + \alpha_{n+1}\big[A_{n+1} - \widehat{A}_n\big] \tag{19}$$

Under the assumption that the matrix sequence $\{\widehat{A}_n : n \geq 0\}$ is invertible for each n, the sequence of estimates obtained using (18,19) are identical to the parameter estimates obtained using the LSTD(0) algorithm (Devraj and Meyn, 2017a; Devraj, 2019). Consequently, the LSTD(0) algorithm achieves the optimal asymptotic covariance.

Q-learning and many other RL algorithms can also be cast as SA recursions. They are no longer linear, but it is anticipated that bounds can be obtained through linearization (Gerencser, 1999).

Literature Survey Finite time performance bounds for linear SA were obtained in many prior papers, subject to the assumption that the noise sequence $\{\Delta_n\}$ appearing in (3) is a martingale difference sequence (Dalal et al., 2017; Lakshminarayanan and Szepesvari, 2018). Much of the literature on finite time bounds for linear SA recursions with Markovian noise has been recent.

For constant step-size algorithms with step-size α, it follows from analysis in Borkar and Meyn (2000) that the pair process (θ_n, Φ_n) is a geometrically ergodic Markov chain, and the covariance of θ_n is $O(\alpha)$ in steady state. Finite time bounds of order $O(\alpha)$ were obtained in Tadić (2006); Bhandari et al. (2018); Srikant and Ying (2019); Hu and Syed (2019). Unfortunately, these bounds are not tight, and hence their value for algorithm design is limited.

Mean-square error bounds have also been obtained for diminishing step-size algorithms, to establish the optimal rate of convergence $\mathsf{E}[\|\widetilde{\theta}_n\|^2] \leq b_\theta/n$ (Srikant and Ying, 2019; Bhandari et al., 2018; Chen et al., 2019b). The constant b_θ is a function of the mixing time of the underlying Markov chain. These results require strong assumptions (uniform ergodicity of the Markov chain), and do not obtain the optimal constant $b_\theta^* = \operatorname{trace}(\Sigma_\theta)$. Finite time bounds are obtained in Karimi et al. (2019) for $\mathsf{E}[\|\bar{f}(\theta_n)\|^2]$. This may be a more relevant performance criterion, but the resulting bounds obtained to-date are not tight.

Contributions The main contribution of this paper is a general framework for analyzing the finite time performance of linear SA algorithms with Markovian noise, and vanishing step-size (required to achieve the optimal convergence rate of Chung-Ruppert-Polyak). The M/M/1 queue example illustrates plainly that Markovian noise introduces challenges not seen in the "white noise" setting, and that the finite-n error bound (7) cannot be obtained without substantial restrictions. Even under the assumptions of Glynn and Ormoneit (2002) (uniform ergodicity, and bounded noise), the resulting bounds are *extremely loose* and hence may give little insight for algorithm design. Our approach allows us to obtain explicit bounds under weak assumptions. In particular, the V-uniform assumption imposed in this work is far weaker than geometric mixing.

Our starting point is the classical martingale approximation of the noise used in CLT analysis of Markov chains (Meyn and Tweedie, 2009, Chapter 17), and used in the analysis of SA recursions since Metivier and Priouret (1984). For each n, the random vector Δ_n is expressed as the sum of a martinagle difference and a telescoping term. The solution of the linear recursion (3) is decomposed as a sum of the respective responses:

$$\widetilde{\theta}_n = \widetilde{\theta}_n^{\mathcal{M}} + \widetilde{\theta}_n^{\mathcal{T}} \tag{20}$$

The challenge is to obtain explicit bounds on the mean square error for each term.

A vector-valued sequence $\{e_n\}$ converges to zero at rate $1/n^{\varrho_0}$ if

$$\lim_{n\to\infty} n^{\varrho}\|e_n\| = \begin{cases} 0, & \text{if } \varrho < \varrho_0 \\ \infty, & \text{if } \varrho > \varrho_0 \end{cases}$$

Bounds for the mean square error are obtained in Thm. 2.4, subject to conditions on both the matrix A and the noise sequence. In summary, under general assumptions on $\{\Delta_n\}$,

(i) The bound (5) holds if $\frac{1}{2}I + A$ is Hurwitz.

(ii) (6) holds if $I + A$ is Hurwitz.

(iii) If there is an eigenvalue of A with $\text{Real}(\lambda) > -\frac{1}{2}$, and corresponding left-eigenvector v that lies outside of the null-space of Σ_Δ, then

$$\lim_{n\to\infty} n^{2\rho}\mathsf{E}[|v^{\intercal}\widetilde{\theta}_n|^2] = \begin{cases} 0, & \varrho < \varrho_0 \\ \infty, & \varrho > \varrho_0 \end{cases} \tag{21}$$

with $\rho_0 = |\text{Real}(\lambda)|$. The convergence of $\mathsf{E}[\|\widetilde{\theta}_n\|^2]$ to zero is thus no faster than $n^{-2\rho_0}$.

2 Mean Square Convergence

Notation and Background Consider the linear SA recursion (3), with the noise sequence $\{\Delta_n\}$ defined in (2). We use the following notation to represent the noise as a function of Φ_n:

$$f^*(\Phi_n) := \Delta_n = f(\theta^*, \Phi_n) \tag{22}$$

A form of geometric ergodicity is assumed throughout. To apply standard theory, we assume that the state space Z is *Polish* (the standing assumption in Meyn and Tweedie (2009)). We fix a measurable function $V\colon \mathsf{Z} \to [1,\infty)$, and let L_∞^V denote the set of measurable functions $g\colon \mathsf{Z} \to \mathbb{R}$ satisfying

$$\|g\|_V := \sup_{z\in\mathsf{Z}} \frac{|g(z)|}{V(z)} < \infty$$

The Markov chain $\mathbf{\Phi}$ is assumed to be V-*uniformly ergodic*: there exists $\rho \in (0,1)$, and $B_V < \infty$ such that for each $g \in L_\infty^V$, $z \in \mathsf{Z}$,

$$\begin{aligned}\Big|\mathsf{E}[g(\Phi_n) \mid \Phi_0 = z] - \pi(g)\Big| \\ \le B_V \|g\|_V \rho^n V(z)\,, \quad n \ge 0\end{aligned} \tag{23}$$

where π is the unique invariant measure, and $\pi(g) = \int g(z)\,\pi(dz)$ is the steady state mean.

The uniform bound (23) is not a strong assumption. For example, it is satisfied for the M/M/1 queue described below (8) with $V(z) = \exp(\varepsilon_0 z)$, for $\varepsilon_0 > 0$ sufficiently small, with $z \in \mathsf{Z} = \{0, 1, \dots\}$ (Meyn and Tweedie, 2009, Theorem. 16.4.1).

The following are imposed throughout:

Assumptions:

(A1) The Markov process $\mathbf{\Phi}$ is V-uniformly ergodic, with unique invariant measure π.

(A2) The $d\times d$ matrix A is Hurwitz, and the step-size sequence $\alpha_n \equiv 1/n$, $n \ge 1$.

(A3) $f^* : \mathsf{Z} \to \mathbb{R}^d$ satisfies $\|f_i^{*2}\|_V < \infty$ and $\pi(f_i^*) = 0$ for each i.

For any $g \in L_\infty^V$, denote $\tilde{g}(z) = g(z) - \pi(g)$, and

$$\hat{g}(z) = \sum_{n=0}^{\infty} \mathsf{E}[\tilde{g}(\Phi_n) \mid \Phi_0 = z] \tag{24}$$

It is evident that $\hat{g} \in L_\infty^V$ under (A1). Further conclusions are summarized below. Thm. 2.1 (i) follows immediately from (A1). Part (ii) follows from (i) and Meyn and Tweedie (2009, Lemma 15.2.9) (the chain is also $\sqrt{V}$-uniformly ergodic).

Theorem 2.1. *The following conclusions hold for a V-uniformly ergodic Markov chain:*

(i) *The function $\hat{g} \in L_\infty^V$ defined in (24) has zero mean, and solves Poisson's equation:*

$$\mathsf{E}[\hat{g}(\Phi_{k+1}) \,|\, \Phi_k = z] = \hat{g}(z) - \tilde{g}(z) \tag{25}$$

(ii) *If $g^2 \in L_\infty^V$, then $\hat{g}^2 \in L_\infty^V$.* □

Assumption (A3) implies that the sequence $\{\Delta_n\}$ appearing in (3) is zero mean for the stationary version of the Markov chain $\mathbf{\Phi}$. Its asymptotic covariance (appearing in the Central Limit Theorem) is denoted

$$\Sigma_\Delta = \sum_{k=-\infty}^{\infty} \mathsf{E}_\pi[\Delta_k \Delta_0^{\intercal}] \tag{26}$$

where the expectations are in steady state.

A more useful representation of Σ_Δ is obtained through a decomposition of the noise sequence based on Poisson's equation. This now standard technique was introduced in the SA literature in the 1980s (Metivier and Priouret, 1984).

With f^* defined in (22), denote by $\hat{f}$ a solution to Poisson's equation:

$$\mathsf{E}[\hat{f}\,(\Phi_{k+1}) \mid \Phi_k = z] = \hat{f}(z) - f^*(z) \tag{27}$$

This is in fact d separate Poisson equations since $f^*\colon \mathsf{Z} \to \mathbb{R}^d$. It is assumed for convenience that the solutions are normalized so $\hat{f}$ has zero steady-state mean. This is justified by the fact that $\hat{f} - \pi(\hat{f})$ also solves (27) under assumption (A3). The fact that $\hat{f}_i^2 \in L_\infty^V$ for $1 \le i \le d$ follows from Thm. 2.1 (ii).

We then write, for $n \ge 1$,

$$\Delta_n = f^*(\Phi_n) = \Delta_{n+1}^m + Z_n - Z_{n+1}$$

where $Z_n = \hat{f}(\Phi_n)$ and $\Delta_{n+1}^m = Z_{n+1} - \mathsf{E}[Z_{n+1} \mid \mathcal{F}_n]$ is a martingale difference sequence. Each of

the sequences is bounded in L_2, and the asymptotic covariance (26) is expressed

$$\Sigma_\Delta = \mathsf{E}_\pi[\Delta_n^m \Delta_n^{m\intercal}] \tag{28}$$

where the expectation is taken in steady-state. The equivalence of (28) and (26) appears in Meyn and Tweedie (2009, Theorem 17.5.3) for the case in which Δ_n is scalar valued; the generalization to vector valued processes involves only notational changes.

Decomposition and Scaling We now explain the decomposition (20). Each of the two sequences $\{\widetilde{\theta}_n^{\mathcal{M}}, \widetilde{\theta}_n^{\mathcal{T}}\}$ evolves as a SA sequence, with different inputs and initial conditions:

$$\mathcal{E}_{n+1}^{\mathcal{M}} = \mathcal{E}_n^{\mathcal{M}} + \alpha_{n+1}\big[A\mathcal{E}_n^{\mathcal{M}} + \Delta_{n+2}^m\big] \tag{29a}$$
$$\mathcal{E}_{n+1}^{\mathcal{T}} = \mathcal{E}_n^{\mathcal{T}} + \alpha_{n+1}\big[A\mathcal{E}_n^{\mathcal{T}} + Z_{n+1} - Z_{n+2}\big] \tag{29b}$$

with $\mathcal{E}_0^{\mathcal{M}} = \mathcal{E}_0$ and $\mathcal{E}_0^{\mathcal{T}} = 0$.

The second recursion admits a more tractable realization through the change of variables, $\Xi_n = \widetilde{\theta}_n^{\mathcal{T}} + \alpha_n Z_{n+1}$, $n \ge 1$.

Lemma 2.2. *The sequence $\{\Xi_n\}$ evolves as the SA recursion, with $\Xi_1 = Z_1$,*

$$\Xi_{n+1} = \Xi_n + \alpha_{n+1}\big[A\Xi_n - \alpha_n[I + A]Z_{n+1}\big] \tag{30}$$

□

Lemma 2.2 combined with (29) gives

$$\widetilde{\theta}_n = \widetilde{\theta}_n^{(1)} + \widetilde{\theta}_n^{(2)} + \widetilde{\theta}_n^{(3)} \tag{31}$$

where $\widetilde{\theta}_n^{(1)} = \widetilde{\theta}_n^{\mathcal{M}}$, $\widetilde{\theta}_n^{(2)} = \Xi_n$, and $\widetilde{\theta}_n^{(3)} = -\alpha_n Z_{n+1}$ for $n \ge 1$. Note that $\widetilde{\theta}_n^{\mathcal{T}} = \widetilde{\theta}_n^{(2)} + \widetilde{\theta}_n^{(3)}$.

It is more convenient to work directly with the recursion for the scaled sequence:

Lemma 2.3. *For any $\varrho \in (0, 1/2]$, the scaled sequence $\widetilde{\theta}_n^{\varrho} = n^{\varrho}\widetilde{\theta}_n$ admits the recursion,*

$$\mathcal{E}_{n+1}^{\varrho} = \mathcal{E}_n^{\varrho} + \alpha_{n+1}\big[\varrho_n \mathcal{E}_n^{\varrho} + A(n,\varrho)\mathcal{E}_n^{\varrho} + (n+1)^{\varrho}\Delta_{n+1}\big] \tag{32}$$

where $\varrho_n = \varrho + \varepsilon(n,\varrho)$ with $\varepsilon(n,\varrho) = O(n^{-1})$, and $A(n,\varrho) = (1+n^{-1})^{\varrho}A$. □

Denote $\widetilde{\theta}_n^{\varrho,(i)} = n^{\varrho}\widetilde{\theta}_n^{(i)}$ for each i. Lemma 2.3 combined with (31) gives

$$\widetilde{\theta}_n^{\varrho} = \widetilde{\theta}_n^{\varrho,(1)} + \widetilde{\theta}_n^{\varrho,(2)} + \widetilde{\theta}_n^{\varrho,(3)} \tag{33}$$

The first two sequences evolve as SA recursions:

$$\mathcal{E}_{n+1}^{\varrho,(1)} = \mathcal{E}_n^{\varrho,(1)} + \alpha_{n+1}\big[[\varrho_n I + A(n,\varrho)]\mathcal{E}_n^{\varrho,(1)} + (n+1)^{\varrho}\Delta_{n+2}^m\big] \tag{34a}$$
$$\mathcal{E}_{n+1}^{\varrho,(2)} = \mathcal{E}_n^{\varrho,(2)} + \alpha_{n+1}\big[[\varrho_n I + A(n,\varrho)]\mathcal{E}_n^{\varrho,(2)} - (n+1)^{\varrho}\alpha_n[I + A]Z_{n+1}\big] \tag{34b}$$

with initial conditions $\mathcal{E}_0^{\varrho,(1)} = \mathcal{E}_0^{\varrho}$, $\mathcal{E}_1^{\varrho,(2)} = \Xi_1$.

Mean Square Error Bounds Fix the initial condition $(\Phi_0, \widetilde{\theta}_0)$, and denote $\mathrm{Cov}\,(\theta_n) = \mathsf{E}[\widetilde{\theta}_n\widetilde{\theta}_n^{\intercal}]$ and $\Sigma_Z = \mathsf{E}_\pi[Z_n Z_n^{\intercal}]$. The following summarizes bounds on the convergence rate of $\mathsf{E}[\|\widetilde{\theta}_n\|^2] = \mathrm{trace}\,(\mathrm{Cov}\,(\theta_n))$.

Theorem 2.4. *Suppose (A1)-(A3) hold. Then, for the linear recursion (3),*

(i) *If $\mathrm{Real}(\lambda) < -\frac{1}{2}$ for every eigenvalue λ of A, then for some $\delta = \delta(A, \Sigma_\Delta) > 0$,*

$$\mathrm{Cov}\,(\theta_n) = n^{-1}\Sigma_\theta + O(n^{-1-\delta}), \qquad n \ge 0,$$

where $\Sigma_\theta \ge 0$ is the solution to the Lyapunov equation (4). Consequently, the rate of convergence of $\mathsf{E}[\|\widetilde{\theta}_n\|^2]$ is $1/n$.

(ii) *Suppose there is an eigenvalue λ of A that satisfies $-\varrho_0 = \mathrm{Real}(\lambda) > -\frac{1}{2}$. Let $v \ne 0$ denote a corresponding left eigenvector, and suppose that $\Sigma_\Delta v \ne 0$. Then, $\mathsf{E}[|v^{\intercal}\widetilde{\theta}_n|^2]$ converges to 0 at rate $n^{-2\varrho_0}$.* □

The proof proceeds by establishing the convergence rate for each component in (31). Details are in Appendix A.2.

One challenge in extension to nonlinear recursions is that the noise sequence depends on the parameter estimates (recall (2)). This is true even for TD learning with linear function approximation (see discussion surrounding (17)). Extension to these recursions is obtained through coupling: consider the error sequence for a random linear recursion

$$\widetilde{\theta}_{n+1}^{\circ} = \widetilde{\theta}_n^{\circ} + \alpha_{n+1}[A_{n+1}\widetilde{\theta}_n^{\circ} + A_{n+1}\theta^* - b_{n+1}] \tag{35}$$

subject to the following assumptions:

(A4) The sequences $\{A_n, b_n\}$ are functions of the Markov chain:

$$A_n = \mathcal{A}(\Phi_n), \quad b_n = \mathcal{B}(\Phi_n),$$

which satisfy $\|\mathcal{A}^2_{i,j}\|_V < \infty$, $\|\mathcal{B}^2_i\|_V < \infty$ for each $1 \le i, j \le d$. The steady state means are denoted $A = \mathsf{E}_\pi[A_n]$, $b = \mathsf{E}_\pi[b_n]$. Moreover, the matrix A is Hurwitz, and $\theta^* = A^{-1}b$.

Theorem 2.5. *Under A1-A4, if the matrix $\frac{1}{2}I + A$ is Hurwitz, the error bound (5) holds for $\{\widetilde{\theta}^\circ_n\}$ obtained from (35), with $\Delta_{n+1} = A_{n+1}\theta^* - b_{n+1}$.*

To establish coupling with (3), we write (35) in the suggestive form

$$\widetilde{\theta}^\circ_{n+1} = \widetilde{\theta}^\circ_n + \alpha_{n+1}[A\widetilde{\theta}^\circ_n + \Delta_{n+1} + \tilde{A}_{n+1}\widetilde{\theta}^\circ_n] \quad (36)$$

where $\Delta_{n+1} = A_{n+1}\theta^* - b_{n+1}$, $\tilde{A}_{n+1} = A_{n+1} - A$.

With common initial condition Φ_0, the sequence $\{\widetilde{\theta}^\circ_n\}$ is compared with the error sequence $\{\widetilde{\theta}^\bullet_n\}$ for the corresponding linear SA algorithm:

$$\widetilde{\theta}^\bullet_{n+1} = \widetilde{\theta}^\bullet_n + \alpha_{n+1}[A\widetilde{\theta}^\bullet_n + \Delta_{n+1}]$$

The difference sequence $\{\mathcal{E}_n := \widetilde{\theta}^\circ_n - \widetilde{\theta}^\bullet_n\}$ evolves according to (3), with vanishing noise:

$$\mathcal{E}_{n+1} = \mathcal{E}_n + \alpha_{n+1}[A\mathcal{E}_n + (A_{n+1} - A)\widetilde{\theta}^\circ_n] \quad (37)$$

Let $\lambda = -\varrho_0 + ui$ denote an eigenvalue of the matrix A with largest real part.

Theorem 2.6. *Under (A1)-(A4),*

(i) $\limsup_{n\to\infty} n^2\mathsf{E}[\|\mathcal{E}_n\|^2] < \infty$ *if* $\varrho_0 > 1$.

(ii) $\limsup_{n\to\infty} n^{2\varrho}\mathsf{E}[\|\mathcal{E}_n\|^2] < \infty$ *for all* $\varrho < \varrho_0$, *provided* $\varrho_0 \le 1$. □

Thm. 2.6 provides a remarkable bound when $\rho_0 > 1$: it immediately implies Thm. 2.5 because the mean square coupling error $\mathsf{E}[\|\mathcal{E}_n\|^2]$ tends to zero at rate no slower than n^{-2}, which is far faster than $\mathsf{E}[\|\widetilde{\theta}^\bullet_n\|^2] \approx \text{trace}\,(\Sigma_\theta)n^{-1}$.

An alert reader may observe that Theorems 2.5 and 2.6 leave out a special case: consider $\rho_0 < \frac{1}{2}$, so that the rate of convergence of $\mathsf{E}[\|\widetilde{\theta}^\bullet_n\|^2]$ is the sub-optimal value $n^{-2\rho_0}$. The bound obtained in Thm. 2.6 remains valuable, in the sense that it combined with Thm. 2.4 (ii) implies the rate of convergence of $\mathsf{E}[\|\widetilde{\theta}^\circ_n]\|^2$ is no slower than $n^{-2\rho_0}$. However, because $\mathsf{E}[\|\mathcal{E}_n\|^2]$ and $\mathsf{E}[\|\widetilde{\theta}^\bullet_n\|^2]$ tend to zero at the same rate, we cannot rule out the possibility that $\widetilde{\theta}^\circ_n = \mathcal{E}_n + \widetilde{\theta}^\bullet_n$ converges to zero much faster. In particular, it remains to prove that if there is an eigenvalue λ of A that satisfies $-\varrho_0 = \text{Real}(\lambda) > -\frac{1}{2}$, and an eigenvector $v \ne 0$ satisfying $\Sigma_\Delta v \ne 0$, then, $\mathsf{E}[|v^\intercal\widetilde{\theta}^\circ_n|^2]$ converges to 0 at rate $n^{-2\varrho_0}$.

Implications Thm. 2.4 indicates that the convergence rate of $\text{Cov}\,(\theta_n)$ is determined jointly by the matrix A, and the martingale difference component of the noise sequence $\{\Delta_n\}$. Convergence of $\{\widetilde{\theta}_n\}$ can be slow if the matrix A has eigenvalues close to zero.

The result also explains the slow convergence of some RL algorithms. For instance, the matrix A in Watkins' Q-learning has at least one eigenvalue with real part greater than or equal to $-(1-\beta)$, where β is the discount factor appearing in the Markov decision process (Watkins, 1989; Devraj and Meyn, 2017a; Devraj, 2019). Since β is usually close to one, Thm. 2.4 implies that the convergence rate of the algorithm is much slower than n^{-1}. Under the assumption that A is Hurwitz, the $1/n$ convergence rate is guaranteed by the use of a modified step-size sequence $\alpha_n = g/n$, with $g > 0$ chosen so that the matrix $\frac{1}{2}I + gA$ is Hurwitz. Corollary 2.7 follows directly from Thm. 2.4 (i).

Corollary 2.7. *Let g be a constant such that $\frac{1}{2}I + gA$ is Hurwitz, and define for $n \ge 0$,*

$$\widetilde{\theta}_{n+1} = \widetilde{\theta}_n + \frac{g}{n+1}[A\widetilde{\theta}_n + \Delta_{n+1}]$$

Then, for some $\delta = \delta(A, g, \Sigma_\Delta) > 0$,

$$\text{Cov}\,(\theta_n) = \mathsf{E}[\widetilde{\theta}_n\widetilde{\theta}_n^\intercal] = n^{-1}\Sigma^g_\theta + O(n^{-1-\delta})$$

where $\Sigma^g_\theta \ge 0$ solves the Lyapunov equation

$$[\tfrac{1}{2}I + gA]\Sigma + \Sigma[\tfrac{1}{2}I + gA]^\intercal + g^2\Sigma_\Delta = 0 \quad \square$$

We can also ensure the $1/n$ convergence rate by using a matrix gain. Provided A is invertible, and if it is known beforehand, $\alpha_n = -A^{-1}/n$ is the optimal matrix step-size sequence (in terms of minimizing the asymptotic covariance) (Benveniste et al., 1990; Kushner and Yin, 1997; Devraj et al., 2020). The SQNR algorithm of Ruppert (1985) and the Zap-SNR algorithm of Devraj and Meyn (2017a); Devraj (2019) provide general approaches to recursively estimate the optimal matrix gain.

Finer Error Bound With $\hat{f}$ a zero-mean solution to (27), let $\hat{\hat{f}}$ be the zero-mean solution to the second Poisson equation

$$\mathsf{E}[\hat{\hat{f}}\,(\Phi_{k+1}) \mid \Phi_k = z] = \hat{\hat{f}}(z) - \hat{f}(z) \tag{38}$$

We then write, for $n \geq 1$,

$$Z_n = \widehat{\Delta}^m_{n+1} + \widehat{Z}_n - \widehat{Z}_{n+1} \tag{39}$$

where $\widehat{Z}_n = \hat{\hat{f}}(\Phi_n)$, and $\widehat{\Delta}^m_{n+1} = \widehat{Z}_{n+1} - \mathsf{E}[\widehat{Z}_{n+1} \mid \mathcal{F}_n]$ is a martingale difference sequence.

The type of decomposition discussed below (29) can be applied to $\widetilde{\theta}^{(2)}_n$ in (30) for $n \geq 2$:

$$\widetilde{\theta}^{(2)}_n = \widetilde{\theta}^{(2,1)}_n + \widetilde{\theta}^{(2,2)}_n + \widetilde{\theta}^{(2,3)}_n \tag{40}$$

The first two sequences evolve as SA recursions:

$$\begin{aligned}\widetilde{\theta}^{(2,1)}_{n+1} = \widetilde{\theta}^{(2,1)}_n &+ \alpha_{n+1}\big[A\widetilde{\theta}^{(2,1)}_n \\ &- \alpha_n[I + A]\widehat{\Delta}^m_{n+2}\big]\end{aligned} \tag{41a}$$

$$\begin{aligned}\widetilde{\theta}^{(2,2)}_{n+1} = \widetilde{\theta}^{(2,2)}_n &+ \alpha_{n+1}\big[A\widetilde{\theta}^{(2,2)}_n \\ &+ \alpha_{n-1}\alpha_n[2I + A][I + A]\widehat{Z}_{n+1}\big]\end{aligned} \tag{41b}$$

with initial conditions $\widetilde{\theta}^{(2,1)}_1 = Z_1, \widetilde{\theta}^{(2,2)}_2 = -\frac{1}{2}[I + A]\widehat{Z}_2$, and $\widetilde{\theta}^{(2,3)}_n = \alpha_{n-1}\alpha_n[I+A]\widehat{Z}_{n+1}$. Therefore, $\widetilde{\theta}_n$ for $n \geq 2$ can be decomposed as:

$$\widetilde{\theta}_n = \widetilde{\theta}^{(1)}_n + \widetilde{\theta}^{(2,1)}_n + \widetilde{\theta}^{(2,2)}_n + \widetilde{\theta}^{(2,3)}_n + \widetilde{\theta}^{(3)}_n \tag{42}$$

The error bound (6) is obtained from (42). The proof is in Appendix A.3.

Theorem 2.8. *Suppose Assumptions (A1)-(A3) hold, and moreover Real*$(\lambda) < -1$ *for every eigenvalue* λ *of* A*. Then, for* (3)*,*

$$\mathrm{Cov}\,(\theta_n) = n^{-1}\Sigma_\theta + n^{-2}\Sigma_{\theta,2} + O(n^{-2-\delta})$$

where $\delta = \delta(I + A, \Sigma_\Delta) > 0$*,* $\Sigma_{\theta,2} = \Sigma_\sharp + \Sigma_Z - \mathsf{E}_\pi[\Delta^m_n \widehat{Z}^\intercal_n] - \mathsf{E}_\pi[\widehat{Z}_n(\Delta^m_n)^\intercal]$*, and* $\Sigma_\sharp$ *is the unique solution to the Lyapunov equation:*

$$\begin{aligned}0 =&[I + A][\Sigma - \mathrm{Cov}_\pi(\widehat{\Delta}^m_n, \Delta^m_n)] \\ &+ [\Sigma - \mathrm{Cov}_\pi(\Delta^m_n, \widehat{\Delta}^m_n)][I + A]^\intercal \\ &+ A\Sigma_\theta A^\intercal - \Sigma_\Delta\end{aligned} \tag{43}$$

3 Conclusions

Performance bounds for recursive algorithms are challenging outside of the special cases surveyed in the introduction. The general framework developed in this paper provides tight finite time performance for linear stochastic recursions under mild conditions on the Markovian noise, and we are confident that the techniques will extend to obtain similar bounds for nonlinear SA provided that the linearization (2) is meaningful.

The bound (5) implies that, for some constant b_θ and all n,

$$\mathsf{E}[\|\widetilde{\theta}_n\|^2] \leq n^{-1}\mathrm{trace}\,(\Sigma_\theta) + n^{-1-\delta}b_\theta\,.$$

It may be argued that we have not obtained a finite-n bound, because a bound on the constant b_θ is lacking. Our response is that the precision of the dominant term is most important. We have tested the bound in numerous experiments in which the empirical mean square error is obtained from multiple independent trials, and the resulting histogram is compared to what is predicted by the Central Limit Theorem with covariance Σ_θ. It is found that the Central Limit Theorem is highly predictive of finite-n performance in most cases (Devraj and Meyn, 2017a; Devraj, 2019; Devraj et al., 2020). While it is hoped that further research will provide bounds on b_θ, it seems likely that any bound will involve high-order statistics of the Markov chain; evidence of this is the complex coefficient of n^{-2} in (6) for the special case $\delta = 1$.

Current research concerns these topics, as well as algorithm design for RL in various settings.

Acknowledgements Financial supports from ARO grant W911NF1810334, EPCN 1935389 & CPS 1646229, French National Research Agency grant ANR-16-CE05-0008, and University of Florida Informatics Institute are gratefully acknowledged.

References

Benveniste, A., Métivier, M., and Priouret, P. (1990). *Adaptive algorithms and stochastic approximations*, volume 22 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin. Translated from the French by Stephen S. Wilson.

Benveniste, A., Métivier, M., and Priouret, P. (2012). *Adaptive algorithms and stochastic approximations*. Springer.

Bhandari, J., Russo, D., and Singal, R. (2018). A finite time analysis of temporal difference learning with linear function approximation. *arXiv preprint arXiv:1806.02450*.

Blum, J. R. (1954). Multidimensional stochastic approximation methods. *The Annals of Mathematical Statistics*, pages 737–744.

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

Borkar, V. S. and Meyn, S. P. (2000). The ODE method for convergence of stochastic approximation and reinforcement learning. *SIAM J. Control Optim.*, 38(2):447–469. (see also *IEEE CDC*, 1998).

Chen, S., Devraj, A. M., Bušić, A., and Meyn, S. (2019a). Zap Q Learning with nonlinear function approximation. Submitted for publication and arXiv e-prints.

Chen, Z., Zhang, S., Doan, T., Maguluri, S., and Clarke, J. (2019b). Performance of q-learning with linear function approximation: Stability and finite-time analysis. *arXiv preprint arXiv:1905.11425*.

Chung, K. L. et al. (1954). On a stochastic approximation method. *The Annals of Mathematical Statistics*, 25(3):463–483.

Dalal, G., Szorenyi, B., Thoppe, G., and Mannor, S. (2017). Finite sample analysis of two-timescale stochastic approximation with applications to reinforcement learning. *arXiv preprint arXiv:1703.05376*.

Devraj, A. M. (2019). *Reinforcement Learning Design with Optimal Learning Rate*. PhD thesis, University of Florida.

Devraj, A. M., Bušić, A., and Meyn, S. (2019). Zap Q-Learning – a user's guide. In *Proc. of the Fifth Indian Control Conference*.

Devraj, A. M., Bušić, A., and Meyn, S. (2020). Fundamental design principles for reinforcement learning algorithms. In *Handbook on Reinforcement Learning and Control*. Springer.

Devraj, A. M. and Meyn, S. P. (2017a). Fastest convergence for Q-learning. *ArXiv e-prints*.

Devraj, A. M. and Meyn, S. P. (2017b). Zap Q-learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*.

Duffy, K. and Meyn, S. (2014). Large deviation asymptotics for busy periods. *Stochastic Systems*, 4(1):300–319.

Duffy, K. R. and Meyn, S. P. (2010). Most likely paths to error when estimating the mean of a reflected random walk. *Performance Evaluation*, 67(12):1290–1303.

Gerencser, L. (1999). Convergence rate of moments in stochastic approximation with simultaneous perturbation gradient approximation and resetting. *IEEE Transactions on Automatic Control*, 44(5):894–905.

Glynn, P. W. and Ormoneit, D. (2002). Hoeffding's inequality for uniformly ergodic Markov chains. *Statistics and Probability Letters*, 56:143–146.

Golub, G. and Van Loan, C. (1996). Matrix computations. 3rd. edn. ed.

Hu, B. and Syed, U. A. (2019). Characterizing the exact behaviors of temporal difference learning algorithms using markov jump linear system theory. *arXiv preprint arXiv:1906.06781*.

Karimi, B., Miasojedow, B., Moulines, E., and Wai, H.-T. (2019). Non-asymptotic analysis of biased stochastic approximation scheme. In

Conference on Learning Theory, pages 1944–1974.

Kiefer, J. and Wolfowitz, J. (1952). Stochastic estimation of the maximum of a regression function. *Ann. Math. Statist.*, 23(3):462–466.

Kushner, H. and Yin, G. G. (2003). *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media.

Kushner, H. J. and Yin, G. G. (1997). *Stochastic approximation algorithms and applications*, volume 35 of *Applications of Mathematics (New York)*. Springer-Verlag, New York.

Lakshminarayanan, C. and Szepesvari, C. (2018). Linear stochastic approximation: How far does constant step-size and iterate averaging go? In *International Conference on Artificial Intelligence and Statistics*, pages 1347–1355.

Metivier, M. and Priouret, P. (1984). Applications of a Kushner and Clark lemma to general classes of stochastic algorithms. *IEEE Transactions on Information Theory*, 30(2):140–151.

Meyn, S. P. (2007). *Control Techniques for Complex Networks*. Cambridge University Press. Pre-publication edition available online.

Meyn, S. P. and Tweedie, R. L. (2009). *Markov chains and stochastic stability*. Cambridge University Press, Cambridge, second edition. Published in the Cambridge Mathematical Library. 1993 edition online.

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

Polyak, B. T. and Juditsky, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.*, 30(4):838–855.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407.

Ruppert, D. (1985). A Newton-Raphson version of the multivariate Robbins-Monro procedure. *The Annals of Statistics*, 13(1):236–245.

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

Srikant, R. and Ying, L. (2019). Finite-time error bounds for linear stochastic approximation and TD learning. *CoRR*, abs/1902.00923.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Mach. Learn.*, 3(1):9–44.

Tadić, V. B. (2006). Asymptotic analysis of temporal-difference learning algorithms with constant step-sizes. *Machine learning*, 63(2):107–133.

Tsitsiklis, J. N. and Van Roy, B. (1997). An analysis of temporal-difference learning with function approximation. *IEEE Trans. Automat. Control*, 42(5):674–690.

Venter, J. et al. (1967). An extension of the robbins-monro procedure. *The Annals of Mathematical Statistics*, 38(1):181–190.

Watkins, C. J. C. H. (1989). *Learning from Delayed Rewards*. PhD thesis, King's College, Cambridge, Cambridge, UK.