Proceedings of the 36th
Conference on Decision & Control
San Diego, California USA • December 1997

On Conditions for Convergence Rates of Stochastic Approximation Algorithms

EDWIN K. P. CHONG[1]
School of Elec. & Comp. Engr.
Purdue University
West Lafayette, IN 47907
echong@ecn.purdue.edu

I-JENG WANG
Institute for Systems Research
University of Maryland
College Park, MD 20742
iwang@isr.umd.edu.

SANJEEV R. KULKARNI[2]
Dept. of Elec. Engineering
Princeton University
Princeton, NJ 08544
kulkarni@ee.princeton.edu

Abstract—We develop deterministic necessary and sufficient conditions on individual noise sequences of a stochastic approximation algorithm for the error of the iterates to converge at a given rate. Specifically, suppose $\{\rho_n\}$ is a given positive sequence converging monotonically to 0. Consider a stochastic approximation algorithm $x_{n+1} = x_n - a_n(A_n x_n - b_n) + a_n e_n$, where $\{x_n\}$ is the iterate sequence, $\{a_n\}$ is the step size sequence, $\{e_n\}$ is the noise sequence, and x^* is the desired zero of the function $f(x) = Ax - b$. We show that $x_n - x^* = o(\rho_n)$ if and only if the sequence $\{e_n\}$ satisfies one of five equivalent conditions. These conditions are based on well known formulas for noise sequences found in the literature.

1 Introduction

We consider a stochastic approximation algorithm for finding the zero of an affine function $f: \mathcal{H} \to \mathcal{H}$, $f(x) = Ax - b$, where $\mathcal{H}$ is a general Hilbert space (on the reals $\mathbb{R}$):

$$x_{n+1} = x_n - a_n A_n x_n + a_n b_n + a_n e_n, \qquad (1)$$

x_n is the estimate of $x^* = A^{-1}b$ (the zero f), a_n is a positive real scalar called the step size, e_n is the noise, $A_n : \mathcal{H} \to \mathcal{H}$ is a bounded linear operator, and $b_n \in \mathcal{H}$ with $b_n \to b$. We assume throughout that the step size sequence $\{a_n\}$ satisfies the following standard conditions: $a_n > 0$, $a_n \to 0$, and $\sum_{n=1}^{\infty} a_n = \infty$. Typical convergence results for (1) specify sufficient conditions on the sequence $\{e_n\}$ (e.g., [1] and [2]).

In this work, we are concerned with the rate of convergence of $\{x_n\}$ to x^*. To characterize the convergence rate, we consider an arbitrary positive real scalar sequence $\{\rho_n\}$ converging monotonically to 0. We are interested in conditions on the noise for which x_n converges to x^* at rate ρ_n; specifically, $x_n - x^* = o(\rho_n)$ (i.e., $\rho_n^{-1}(x_n - x^*) \to 0$). As in some previous work on stochastic approximation, instead of making probabilistic assumptions on the noise, we take a deterministic approach and treat the noise as an individual sequence. Therefore, any dependence of the noise on the iterates is embedded in the condition on the sequence, and is not taken into account explicitly.

In our main result (Theorem 1), we give a necessary and sufficient condition on the sequence $\{e_n\}$ for $x_n - x^* = o(\rho_n)$ to hold. To illustrate our result, consider the special case where $a_n = 1/n$. Then, under appropriate assumptions, $x_n - x^* = o(\rho_n)$ if and only if $(\sum_{k=1}^{n} e_k/\rho_k)/n \to 0$ (i.e., the long-term average of $\{e_n/\rho_n\}$ is 0). In fact, our result provides a set of five equivalent necessary and sufficient conditions on $\{e_n\}$, related to certain familiar conditions found in the literature. Our result provides the tightest possible characterization of the noise for a prespecified rate to be achievable, and has so far not been available. Moreover, ours is the first to provide rate results for the class of stochastic approximation algorithms (1). Note that although the form of the stochastic approximation algorithm is linear, it is not a special case of the usually considered algorithm: $x_{n+1} = x_n - a_n f(x_n) + a_n e_n$. Indeed, the form of the algorithm we adopt is of interest in its own right and has been widely studied (see [1] for a survey of references on this form of stochastic approximation algorithms).

Due to space constraints, in this paper we give only a summary of results and some comments. For detailed proofs, complete references, a discussion of how the assumptions can be relaxed, and examples illustrating the results, contact the authors for an expanded manuscript.

2 Main Result

Our result makes use of a deterministic condition on individual sequences, summarized by the following definition. Let $\{a_n\}$ be a given step size sequence with $a_n < 1$ for all n.

[1]Supported in part by the National Science Foundation under grants ECS-9410313 and ECS-9501652.
[2]Supported in part by the National Science Foundation under grant IRI-9457645.

0-7803-3970-8/97 $10.00 © 1997 IEEE

Definition 1 We say that a sequence $\{s_n\}$ is *small with respect to* $\{a_n\}$ if

$$\lim_{n\to\infty} \frac{\sum_{k=1}^{n} \gamma_k s_k}{\sum_{k=1}^{n} \gamma_k} = 0,$$

where $\{\gamma_n\}$ is a weighting sequence defined by

$$\gamma_n = \begin{cases} a_1 & \text{if } n = 1 \\ a_n \Pi_{k=2}^{n} 1/(1-a_k) & \text{otherwise.} \end{cases}$$

It has been shown in [3] that the above "smallness" condition is equivalent to four other conditions studied in the literature: Kushner and Clark's condition, Chen's condition, Kulkarni and Horn's condition, and a decomposition condition (see also [4]). Therefore, smallness can be tested by any one of five equivalent conditions. Moreover, [3] and [4] establish that the algorithm (1) converges if and only if the noise sequence $\{e_n\}$ is small. Note that the special case with $a_n = 1/n$ is considered in [1] and [5], where the smallness condition reduces to having a zero average ($\gamma_k = 1 \; \forall \; k$).

Let $\{\rho_n\}$ be a given positive real sequence converging monotonically to 0, satisfying:

(G1) $\rho_n^{-1}(b_n - b) \to 0$;

(G2) $(\rho_n - \rho_{n+1})/(a_n\rho_n) \to c$, where $c \in \mathbb{R}$;

(G3) The sequences $\{\rho_{n+1}/\rho_n\}$ and $\{\rho_n/\rho_{n+1}\}$ have bounded variation.

Assumptions (G1) and (G2) are fairly weak. Note that in the standard case where $a_n = an^{-\alpha}$, $a > 0$, $0 < \alpha \leq 1$, and $\rho_n = n^{-\gamma}$, $\gamma > 0$, we have $c = 0$ if $\alpha < 1$, and $c = \gamma/a$ if $\alpha = 1$. Assumption (G3) is technical (see [3] for the definition of *bounded variation*). Note that (G3) holds for any sequence $\{\rho_n\}$ of the the form $\rho_n = n^{-\gamma}$, $\gamma > 0$. For our main result, (G1) and (G2) can be relaxed, but (G3) cannot.

Next, we assume that the sequence $\{A_n\}$ satisfies:

(B1) $\{A_n - A\}$ is small with respect to $\{a_n\}$, where $A : \mathcal{H} \to \mathcal{H}$ is a bounded linear operator with $\inf\{Re\lambda : \lambda \in \sigma(A)\} > 0$, where $\sigma(A)$ denotes the spectrum of A;

(B2) $\lim_{n\to\infty}(\sum_{k=1}^{n} \gamma_k \|A_k\|)/(\sum_{k=1}^{n} \gamma_k) < \infty$;

(B3) $\inf\{Re\lambda : \lambda \in \sigma(A - cI)\} > 0$, where I is the identity operator and $\sigma(A-cI)$ denotes the spectrum of $A - cI$.

Assumptions (B1–B3) are standard in results for stochastic approximation algorithms of the type (1); see, for example, [1], [2], and [3].

We are ready to state our main result.

Theorem 1 *Let $\{x_n\}$ be generated by the stochastic approximation algorithm (1). Assume that (B1–B3) and (G1–G3) hold. Then, $x_n - x^* = o(\rho_n)$ if and only if $\{e_n/\rho_n\}$ is small with respect to $\{a_n\}$.*

The basic idea of the proof of Theorem 1 is to express the sequence $\{\rho_n^{-1}(x_n - x^*)\}$ using a recursion that is essentially (1) with noise sequence $\{e_n/\rho_n\}$. The desired result then follows from applying the convergence results of [3] and [4].

To illustrate our result, we give an example where $\{e_n\}$ is a random process.

Proposition 1 *Suppose $\{e_n\}$ is a martingale difference process with $E(e_n^2) \leq \sigma^2$, $\sigma^2 \in \mathbb{R}$. Let $a_n = n^{-\alpha}$ with $\alpha > 1/2$. Then, $x_n - x^* = o(n^{-(\alpha-1/2)+\epsilon})$ for any $\epsilon > 0$.*

From the above, we easily recover the familiar result that if $a_n = a/n$, then $x_n - x^* = o(n^{-1/2+\epsilon})$ for any $\epsilon > 0$. With similar assumptions, we can obtain an analogous familiar rate result for the Kiefer-Wolfowitz algorithm: $x_n - x^* = o(n^{-1/3+\epsilon})$ for any $\epsilon > 0$.

In [6], Chen gives a sufficient condition for the convergence rate to be a_n^{δ}, where $0 < \delta \leq 1$; i.e., $x_n - x^* = o(a_n^{\delta})$. Using Theorem 1, we can show that Chen's sufficient condition is in fact also necessary.

References

[1] M. A. Kouritzin, "On the convergence of linear stochastic approximation procedures," *IEEE Transactions on Information Theory*, vol. 42, no. 4, pp. 1305–1309, July 1996.

[2] H. Walk and L. Zsidó, "Convergence of Robbins-Monro method for linear problems in a Banach space," *Journal of Mathematical Analysis and Applications*, vol. 139, pp. 152–177, 1989.

[3] I.-J. Wang, E. K. Chong, and S. R. Kulkarni, "Weighted averaging and stochastic approximation," *Mathematics of Control, Signals, and Systems*, 1997, to appear.

[4] I.-J. Wang, E. K. P. Chong, and S. R. Kulkarni, "Equivalent necessary and sufficient conditions on noise sequences for stochastic approximation algorithms," *Advances in Applied Probability*, vol. 28, pp. 784–801, 1996.

[5] D. S. Clark, "Necessary and sufficient conditions for the Robbins-Monro method," *Stochastic Processes and Their Applications*, vol. 17, pp. 359–367, 1984.

[6] H.-F. Chen, "Recent developments in stochastic approximation," in *Proceedings of 1996 IFAC World Congress*, pp. 375–380, June 1996.