Noise Conditions for Prespecified Convergence Rates of Stochastic Approximation Algorithms

Edwin K. P. Chong, *Senior Member, IEEE*,
I-Jeng Wang, *Member, IEEE*, and Sanjeev R. Kulkarni

***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 zero. 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 $f(x) = Ax - b$. Then, under appropriate assumptions, 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: Kushner and Clark's condition, Chen's condition, Kulkarni and Horn's condition, a decomposition condition, and a weighted averaging condition. Our necessary and sufficient condition on $\{e_n\}$ to achieve a convergence rate of $\{\rho_n\}$ is basically that the sequence $\{e_n/\rho_n\}$ satisfies any one of the above five well-known conditions. We provide examples to illustrate our result. In particular, we easily recover the familiar result that if $a_n = a/n$ and $\{e_n\}$ is a martingale difference process with bounded variance, then $x_n - x^* = o(n^{-1/2}(\log(n))^\beta)$ for any $\beta > 1/2$.**

***Index Terms*—Convergence rate, Kiefer–Wolfowitz, necessary and sufficient conditions, noise sequences, Robbins–Monro, stochastic approximation.**

I. Introduction

We consider a general stochastic approximation algorithm for finding the zero of a function $f\colon \mathcal{H} \to \mathcal{H}$, where $\mathcal{H}$ is a general Hilbert space (on the reals $\mathbb{R}$)

$$x_{n+1} = x_n - a_n f_n \tag{1}$$

where $x_n \in \mathcal{H}$ is the estimate of the zero of the function f, a_n is a positive real scalar called the step size, and $f_n \in \mathcal{H}$ represents an estimate of $f(x_n)$. The estimate f_n is often written as $f_n = f(x_n) - e_n$, where $e_n \in \mathcal{H}$ represents the noise in the estimate. Typical convergence results for (1) specify sufficient conditions on the sequence $\{e_n\}$ (see [15] and references therein).

In this paper we are concerned with the rate of convergence of $\{x_n\}$ to x^*, the zero of the function f. To characterize the convergence rate, we consider an arbitrary positive real scalar sequence $\{\rho_n\}$ converging monotonically to zero. We are interested in conditions on the noise for which x_n converges to x^* at rate ρ_n; specifically, $x_n - x^* = o(\rho_n)$. By writing $x_n - x^* = o(\rho_n)$ we mean that $\rho_n^{-1}(x_n - x^*) \to 0$. 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. This approach is in line with recent trends in information theory and statistics where individual sequences have been considered for problems such as source coding, prediction, and regression.

In our main result (Theorem 2), we give necessary and sufficient conditions 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 zero). 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: Kushner and Clark's condition, Chen's condition, Kulkarni and Horn's condition, a decomposition condition, and a weighted averaging condition (see [20] for a study on these conditions and their relationship to convergence of stochastic approximation algorithms).

Classical results on convergence rates are obtained by calculating the asymptotic distribution of the process $\{\rho_n^{-1}(x_n - x^*)\}$, where some appropriate probabilistic assumption on $\{e_n\}$ is imposed. Such an approach was first taken by Chung [4], who studied the asymptotic normality of the sequence $\{\rho_n^{-1}(x_n - x^*)\}$ (for appropriate choice of ρ_n). Sacks [18] later provides an alternative proof of Chung's result. These results provide a means of characterizing the rate of convergence of stochastic approximation algorithms—they basically assert that if $a_n = a/n$, $a > 0$, then the rate of convergence for the stochastic approximation algorithm is essentially $O(n^{-1/2})$. In [6], Fabian provides a more general version of such an asymptotic distribution approach. Using weak convergence methods, Kushner and Huang [12] provide even stronger results along these lines.

Recently, Chen presents results on convergence rates that do not involve calculating asymptotic distributions (see [1] for a summary of results). In particular, he provides rate results of the form $x_n - x^* = o(a_n^\delta)$, where $0 < \delta \le 1$. Chen's characterization of rates are similar to ours, with $\rho_n = a_n^\delta$.

Our results are based on a stochastic approximation algorithm, on a general Hilbert space, of the form (1) with $f_n = A_n x_n - b_n - e_n$ and $f(x) = Ax - b$. Here, A_n and b_n are estimates of A and b, respectively, while e_n is the noise. Note that although f is affine, the form of the stochastic approximation algorithm is not a special case of the usually considered case with $f_n = f(x_n) - e_n$. Indeed, the form of the algorithm we adopt has been widely studied (e.g., [8], [9], [19]; see also [10] for a survey of references on this form of stochastic approximation algorithms). In our main result (Theorem 2), we give necessary and sufficient conditions on $\{e_n\}$ for the algorithm to converge with a prespecified rate ρ_n [i.e., $x_n - x^* = o(\rho_n)$]. Our result therefore provides the tightest possible characterization of the noise for a convergence rate of the form $x_n - x^* = o(\rho_n)$ to be achievable, and has so far not been available. Moreover, ours is the first to provide rate results for the general class of stochastic approximation algorithms with $f_n = A_n x_n - b_n - e_n$. We believe that our approach may be useful for obtaining similarly tight rate results for the nonlinear case with $f_n = f(x_n) - e_n$ (see Section V).

II. Background: A Convergence Theorem

To prove our main result on convergence rates, we first need a convergence theorem for the stochastic approximation algorithm. Consider the algorithm

$$x_{n+1} = x_n - a_n A_n x_n + a_n b_n + a_n e_n \tag{2}$$

Manuscript received September 5, 1997; revised April 9, 1998. This work was supported in part by the National Science Foundation under Grants ECS-9501652 and IRI-9457645.

E. K. P. Chong is with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907-1285 USA.

I-J. Wang is with the Research and Technology Development Center, The Johns Hopkins University Applied Physics Laboratory, Laurel, MD 20723-6099 USA (e-mail: I-Jeng.Wang@jhuapl.edu).

S. R. Kulkarni is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA.

Communicated by P. Moulin, Associate Editor for Nonparametric Estimation, Classification, and Neural Networks.

Publisher Item Identifier S 0018-9448(99)01417-0.

where x_n is the iterate, a_n is 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 $b_n \to b$. Note that the above is a stochastic approximation algorithm of the form (1) with $f_n = A_n x_n - b_n - e_n$ (see also [8], [10], [19] for a treatment of convergence for a similar linear case).

We will need some assumptions for our convergence theorem. First, we assume throughout that the step size sequence $\{a_n\}$ satisfies:

A1) $a_n > 0$, $a_n \to 0$, and $\sum_{n=1}^{\infty} a_n = \infty$.

The above are standard requirements in all stochastic approximation algorithms. Next, we consider conditions on the noise sequence $\{e_n\}$.

A. Noise Conditions

A central issue in studying stochastic approximation algorithms is characterizing the noise sequence $\{e_n\}$ for convergence. Here, we give five conditions on the sequence $\{e_n\}$. Each condition depends not only on $\{e_n\}$, but also on the step size sequence $\{a_n\}$. As we shall point out later, these five conditions are all equivalent and are both necessary and sufficient for convergence of (2). In the statement of the following conditions, we use $\|\cdot\|$ to denote the natural norm on $\mathcal{H}$, and convergence is with respect to this norm.

N1) For some $T > 0$

$$\lim_{n\to\infty} \left(\sup_{n \le p \le m(n,T)} \left\| \sum_{i=n}^{p} a_i e_i \right\| \right) = 0$$

where

$$m(n, T) \triangleq \max\{k: a_n + \cdots + a_k \le T\}.$$

N2)

$$\lim_{T\to 0} \frac{1}{T} \lim_{n\to\infty} \left(\sup_{n \le p \le m(n,T)} \left\| \sum_{i=n}^{p} a_i e_i \right\| \right) = 0.$$

N3) For any $\alpha, \beta > 0$, and any infinite sequence of nonoverlapping intervals $\{I_k\}$ on $\mathbb{N} = \{1, 2, \cdots\}$, there exists $K \in \mathbb{N}$ such that for all $k \ge K$

$$\left\| \sum_{n \in I_k} a_n e_n \right\| < \alpha \sum_{n \in I_k} a_n + \beta.$$

N4) There exist sequences $\{\alpha_n\}$ and $\{\beta_n\}$ such that $e_n = \alpha_n + \beta_n$ for all n, $\alpha_n \to 0$, and $\sum_{k=1}^{n} a_k \beta_k$ converges.

N5)

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

where

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

The noise conditions above have been widely studied in the literature. The first four are proven to be equivalent in [20], while the fifth is shown to be equivalent also to the first four conditions in [21]. Condition N1) is the well-known condition of Kushner and Clark [13], while N2) is a version of a related condition by Chen [2]. Condition N3) was studied by Kulkarni and Horn in [11], related to a "persistently disturbing" noise condition. Condition N4) was suggested in [13, p. 29] (see also [2], [3], [15, p. 11], and [14] for applications of the condition). Condition N5) is a weighted averaging condition (see [19] and [21]). For convenience and without loss of generality, we assume in N5) that $a_n < 1$ for all n. A special case of this condition (with $a_n = 1/n$) is considered in [5] and [10], where the weighted averaging condition reduces to regular arithmetic averaging (i.e., $\gamma_k = 1$ for all k).

For convenience, we define the following terminology.

Definition 1: Let $\{a_n\}$ be a step size sequence. We say that a noise sequence $\{e_n\}$ is *small with respect to* $\{a_n\}$ if $\{e_n\}$ satisfies any of the conditions N1)–N5) with associated step size sequence $\{a_n\}$.

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 $\inf\{\operatorname{Re}\lambda: \lambda \in \sigma(A)\} > 0$, where $\sigma(A)$ denotes the spectrum of A.

B2)

$$\limsup_{n\to\infty} \left(\sum_{k=1}^{n} \gamma_k \|A_k\| \right) \bigg/ \left(\sum_{k=1}^{n} \gamma_k \right) < \infty.$$

In the above, the sequence $\{\gamma_n\}$ is as defined in condition N5). Assumptions B1) and B2) are standard in results for stochastic approximation algorithms of the type (2); see, for example, [8], [10], [19], [21], where (B1) is expressed as $(\sum_{k=1}^{n} \gamma_k A_k)/(\sum_{k=1}^{n} \gamma_k) \to A$. Note that in B1), the smallness of the sequence of operators $\{A_n - A\}$ is with respect to the induced operator norm (on the space of bounded linear operators). Likewise, in B2), the norm on A_k is the induced operator norm.

B. Convergence Theorem

We are now ready to state the convergence theorem that will be used in the proof of our main result. We use the notation $x^* = A^{-1}b$ for the unique solution to $Ax^* = b$, i.e., the desired zero of the function $f(x) = Ax - b$.

Theorem 1: Let $\{x_n\}$ be generated by the stochastic approximation algorithm (2). Suppose conditions A1), B1), and B2) hold. Then, $x_n \to x^*$ if and only if $\{e_n\}$ is small with respect to $\{a_n\}$.

Proof: The proof follows from, [20, Th. 1] and [21, Th. 4]. ■

III. Convergence Rates

In this section, we state and prove our main result (Theorem 2), which provides conditions on the noise sequence that guarantee a given rate of convergence. Recall that we are considering a stochastic approximation algorithm

$$x_{n+1} = x_n - a_n A_n x_n + a_n b_n + a_n e_n. \tag{3}$$

Let $\{\rho_n\}$ be a given positive real sequence converging monotonically to zero. We are interested in conditions on $\{e_n\}$ for which $x_n \to x^*$ at the prescribed rate of ρ_n; specifically, $x_n - x^* = o(\rho_n)$.

First, we need some assumptions on the sequence $\{\rho_n\}$.

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

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

Assumption G1) can in fact be relaxed (see remarks following the proof of Theorem 2), but suffices in many standard scenarios. Assumption G2) is also fairly weak, requiring that the convergence rate of $\{\rho_n\}$ be related to the step size $\{a_n\}$. In the remarks following the proof of Theorem 2, we indicate how G2) can be relaxed. Note that in the standard case where $a_n = an^{-\alpha}$, $a > 0$, $0 < \alpha \le 1$, and $\rho_n = n^{-\gamma}$, $\gamma > 0$, we have $c = 0$ if $\alpha < 1$, and $c = \gamma/a$ if $\alpha = 1$. To see this

$$\begin{aligned}\frac{\rho_n - \rho_{n+1}}{a_n \rho_n} &= \frac{n^{-\gamma} - (n+1)^{-\gamma}}{an^{-\alpha}n^{-\gamma}}\\ &= \frac{n^\gamma}{an^{-\alpha}(n+1)^\gamma}\left(\left(1+\frac{1}{n}\right)^\gamma - 1\right)\\ &= \frac{n^\gamma}{an^{-\alpha}(n+1)^\gamma}\left(\frac{\gamma}{n} + o\left(\frac{1}{n}\right)\right)\\ &= \left(\frac{n}{n+1}\right)^\gamma\left(\frac{\gamma + o(n^{-1})/n^{-1}}{an^{1-\alpha}}\right).\end{aligned}$$

We need one more assumption on $\{\rho_n\}$. We use the following definition.

Definition 2: A scalar sequence $\{a_n\}$ is said to have *bounded variation* if $\sum_{n=1}^\infty |a_{n+1} - a_n| < \infty$.

The additional assumption on $\{\rho_n\}$ is as follows.

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

Assumption G3) is technical. In fact, the assumption cannot be relaxed (see remark following Lemma 1 below). Note that G3) holds for any sequence $\{\rho_n\}$ of the form $\rho_n = n^{-\gamma}$, $\gamma > 0$.

Next, we introduce an additional assumption on the operator A.

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

To satisfy B3) in the case where $A \in \mathbb{R}^{d\times d}$, we need the eigenvalues of A to all exceed c. In cases where $c = 0$, assumption B3) reduces to B1).

With the above assumptions, we are ready to state our main result.

Theorem 2: Let $\{x_n\}$ be generated by the stochastic approximation algorithm (3). Assume that assumptions A1), 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 2 is to express the sequence $\{\rho_n^{-1}(x_n - x^*)\}$ using a recursion that is essentially (3) with noise sequence $\{e_n/\rho_n\}$. The desired result then follows from applying Theorem 1.

To prove Theorem 2, we need the following technical lemma.

Lemma 1: Consider a step size sequence $\{a_n\}$. Let $\{s_n\}$ be a positive sequence such that $\{s_n\}$ and $\{1/s_n\}$ have bounded variation. Define $\hat{a}_n = s_n a_n$. Then, the following hold.

1) If $\{a_n\}$ satisfies A1), then so does $\{\hat{a}_n\}$.

2) A sequence $\{\hat{e}_n\}$ is small with respect to $\{a_n\}$ if and only if $\{\hat{e}_n\}$ is small with respect to $\{\hat{a}_n\}$.

Proof: Note that $\{s_n\}$ and $\{1/s_n\}$ having bounded variation implies that s_n converges to some positive real number. Thus, part 1) is trivial. Part 2) follows from [21, Lemma 3]. ■

Remark: Part 2) can in fact be strengthened as follows. Let L_a be the set of sequences that are small with respect to $\{a_n\}$, and $L_{\hat{a}}$ be the set of sequences that are small with respect to $\{\hat{a}_n\}$. Then, we have $L_a = L_{\hat{a}}$ if and only if $\{s_n\}$ and $\{1/s_n\}$ have bounded variation.

Using the above lemma, we can now prove Theorem 2.

Proof of Theorem 2: Substituting $\hat{x}_n = \rho_n^{-1}(x_n - x^*)$ into (3), we obtain

$$\begin{aligned}\hat{x}_{n+1} &= \frac{\rho_n}{\rho_{n+1}}\left(\hat{x}_n - a_n\rho_n^{-1}A_n(\rho_n\hat{x}_n + x^*) + a_n\frac{b_n}{\rho_n} + a_n\frac{e_n}{\rho_n}\right)\\ &= \hat{x}_n - a_n\frac{\rho_n}{\rho_{n+1}}\left(A_n - \frac{1}{a_n(\rho_n/\rho_{n+1})}\left(\frac{\rho_n}{\rho_{n+1}} - 1\right)\right)\hat{x}_n\\ &\quad + a_n\frac{\rho_n}{\rho_{n+1}}\left((A - A_n)x^* + \frac{e_n}{\rho_n}\right)\\ &\quad + a_n\frac{\rho_n}{\rho_{n+1}}\rho_n^{-1}(b_n - b).\end{aligned}$$

Write $\hat{a}_n = a_n(\rho_n/\rho_{n+1})$, $\hat{e}_n = (A - A_n)x^* + e_n/\rho_n$, $\hat{b}_n = \rho_n^{-1}(b_n - b)$, and

$$\hat{A}_n = A_n - \frac{\rho_n - \rho_{n+1}}{a_n\rho_n}I.$$

Then, we have

$$\hat{x}_{n+1} = \hat{x}_n - \hat{a}_n\hat{A}_n\hat{x}_n + \hat{a}_n\hat{b}_n + \hat{a}_n\hat{e}_n$$

where $\hat{b}_n \to 0$ by assumption G1). Note that by Lemma 1-1), $\{\hat{a}_n\}$ satisfies A1). Also, note that by assumption G2), the sequence $\{\hat{A}_n - (A - cI)\}$ is small with respect to $\{a_n\}$ (using [20, Lemma 2]), and hence also with respect to $\{\hat{a}_n\}$, by virtue of assumption G3) and Lemma 1-2). Thus, by assumption B3), $\{\hat{A}_n\}$ satisfies condition B1) with respect to $\{\hat{a}_n\}$ [where the operator A in B1) is taken to mean $A - cI$ here]. Moreover, by writing

$$\|\hat{A}_n\| \le \|A_n\| + \frac{\rho_n - \rho_{n+1}}{a_n\rho_n}$$

we see that $\{\hat{A}_n\}$ satisfies condition B2) with respect to $\{\hat{a}_n\}$. Hence, by Theorem 1, we conclude that $\hat{x}_n \to 0$ if and only if $\{\hat{e}_n\}$ is small with respect to $\{\hat{a}_n\}$, which holds if and only if $\{e_n/\rho_n\}$ is small with respect to $\{\hat{a}_n\}$, by definition of $\hat{e}_n$ and assumption B1) (we use the fact that difference between two small sequences is also a small sequence). By assumption G3) and Lemma 1-2), we obtain the desired result. ■

Remark:

- Assumption G1) can be relaxed to $\{\rho_n^{-1}(b_n - b)\}$ is small with respect to $\{a_n\}$. The proof above can be easily modified to accommodate this relaxed assumption by including $\rho_n^{-1}(b_n - b)$ into $\hat{e}_n$. The same line of argument as above can then be used.
- Similarly, assumption G2) can be relaxed to $\{(\rho_n - \rho_{n+1})/(a_n\rho_n) - c\}$ is small with respect to $\{a_n\}$. Note that the condition $\lim_{n\to\infty}(\sum_{k=1}^n \gamma_k(\rho_k - \rho_{k+1})/(a_k\rho_k))/(\sum_{k=1}^n \gamma_k) < \infty$ is automatically implied. Essentially the same proof as above goes through in this case.

IV. Examples

In this section, we provide examples to illustrate our main result, Theorem 2. We assume throughout this section that B1)–B3), and G1) hold.

A. Random Noise Sequence

We give an example where $\{e_n\}$ is a random process.

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

Proof: We apply Theorem 2 with $\rho_n = n^{-(\alpha-1/2)+\epsilon}$, $\epsilon > 0$. Note that $\{a_n\}$ satisfies A1) and $\{\rho_n\}$ satisfies G2) and G3).

Now, the process $\{\sum_{k=1}^n a_k e_k/\rho_k\}$ is a martingale (with respect to the filtration generated by $\{e_1, \cdots, e_n\}$). Moreover, we have

$$E\left(\sum_{k=1}^n a_k e_k/\rho_k\right)^2 = \sum_{k=1}^n a_k^2 E(e_k^2)/\rho_k^2 \le \sigma^2\sum_{k=1}^n k^{-(1+2\epsilon)} < \infty.$$

Hence, by the martingale convergence theorem, we conclude that $\sum_{k=1}^n a_k e_k/\rho_k$ converges. Thus, $\{e_n\}$ satisfies condition N4), and the desired result follows from Theorem 2. ■

For the case where $\alpha = 1$ in Proposition 1, we have that $x_n - x^* = o(n^{-1/2+\epsilon})$ for any $\epsilon > 0$, which is consistent with the early rate results of [4] and [18]. We conjecture that the result in Proposition 1 can be sharpened to a rate of the form $o(n^{-\alpha/2+\epsilon})$, with appropriate assumptions on $\{e_n\}$.

An alternative stronger rate result can in fact be obtained

$$x_n - x^* = o(n^{-(\alpha-1/2)}(\log(n))^{\beta}) \qquad \text{a.s.}$$

where $\alpha > 1/2$ (as before) and $\beta > 1/2$. Here, it is again straightforward to show that $\{\rho_n\} = \{n^{-(\alpha-1/2)}(\log(n))^{\beta}\}$ satisfies G2) and G3). Following the argument in the proof above, we have

$$\begin{aligned} E\left(\sum_{k=1}^{n} a_k e_k/\rho_k\right)^2 &= \sum_{k=1}^{n} a_k^2 E(e_k^2)/\rho_k^2 \\ &\le \sigma^2 \sum_{k=1}^{n} \frac{1}{n(\log(n))^{2\beta}} < \infty \end{aligned}$$

(see [16, Th. 3.29]), from which the result follows. For the $\alpha = 1$ case, we obtain $x_n - x^* = o(n^{-1/2}(\log(n))^{\beta})$, which is consistent with almost sure rate results obtained by the law of the iterated logarithm (see, e.g., [7, p. 845] for the Robbins–Monro case, and [9, p. 120] for a linear multivariable algorithm for adaptive filtering).

B. Kiefer–Wolfowitz Algorithm

The Kiefer–Wolfowitz algorithm is a stochastic approximation algorithm designed for minimization using finite differences. Specifically, suppose $J: \mathbb{R}^d \to \mathbb{R}$ is a function with minimizer x^*. The Kiefer–Wolfowitz algorithm for finding x^* is given by

$$x_{n+1} = x_n - a_n \frac{J_n^+ - J_n^-}{2c_n}$$

where c_n is a “perturbation” parameter such that $c_n \to 0$, and J_n^+ and J_n^- are vectors of noisy measurements of the function J at perturbed points. Specifically, the ith components of J_n^+ and J_n^- are defined by

$$\begin{aligned} J_n^+(i) &= J(x_n + c_n\mu_i) - e_n^+(i) \\ J_n^-(i) &= J(x_n - c_n\mu_i) - e_n^-(i) \end{aligned}$$

where μ_i is the unit vector in the ith coordinate direction.

In the literature on Kiefer–Wolfowitz algorithms, it is well known (e.g., [13, p. 252]) that the rate of convergence with $a_n = an^{-1}$ and $c_n = cn^{-1/6}$ is essentially $O(n^{-1/3})$, obtained using asymptotic distribution calculations with appropriate probabilistic assumptions (see also [7] and [17] for almost sure versions of the rate result based on the law of the iterated logarithm). Here, we show that our rate result is consistent with the rate of $n^{-1/3}$ above.

Proposition 2: Suppose J is a positive definite quadratic function, and $\{e_n\}$ satisfies the same conditions as in Proposition 1. Let $a_n = n^{-\alpha}$ with $\alpha > 1/2$, and $c_n = n^{-\gamma}$ with $\alpha - \gamma - 1/2 > 0$ and $\alpha - 3\gamma - 1/2 \le 0$. Then, for the Kiefer–Wolfowitz algorithm above, we have $x_n - x^* = o(n^{-(\alpha-\gamma-1/2)+\epsilon})$ a.s. for any $\epsilon > 0$.

Proof: For convenience, let f_n be the vector whose ith component is $(J(x_n + c_n\mu_i) - J(x_n - c_n\mu_i))/(2c_n)$, e_n the vector whose ith component is $e_n^+(i) - e_n^-(i)$, and $b_n = \nabla J(x_n) - f_n$. Then, we can write the Kiefer–Wolfowitz algorithm as

$$x_{n+1} = x_n - a_n \nabla f(x_n) + a_n b_n + a_n \frac{e_n}{2c_n}.$$

For J a positive definite quadratic function, ∇f is linear and B1) and B2) hold. Moreover, it is straightforward to show that $b_n \to 0$. To complete the proof, we set $\rho_n = n^{-(\alpha-\gamma-1/2)+\epsilon}$ and proceed as in the proof of Proposition 1. ■

From the above result, we see that the best rate is achieved with $\gamma = (\alpha - 1/2)/3$, whence $\rho_n = n^{-2/3\alpha+1/3+\epsilon}$. For the case where $\alpha = 1$, we have that $x_n - x^* = o(n^{-1/3+\epsilon})$ for any $\epsilon > 0$, which is consistent with the well-known rate results as mentioned earlier.

As in the last section, we can strengthen the above result by taking the rate sequence $\rho_n = n^{-(\alpha-\gamma-1/2)}(\log(n))^{\beta}$ to obtain

$$x_n - x^* = o(n^{-(\alpha-\gamma-1/2)}(\log(n))^{\beta}) \qquad \text{a.s.}$$

with $\beta > 1/2$. For the $\alpha = 1$ case, this result is consistent with the almost sure rate results in [17, p. 186] and [7, p. 848] based on the law of the iterated logarithm for a one-dimensional Kiefer–Wolfowitz algorithm.

C. Chen's Rate Condition

In [1], Chen gives sufficient conditions for the convergence rate to be a_n^{δ}, where $0 < \delta \le 1$, i.e., $x_n - x^* = o(a_n^{\delta})$. In particular, Chen imposes the following condition on the noise sequence $\{e_n\}$.

CH) There exists α_n and β_n such that $e_n = \alpha_n + \beta_n$ for all n, $\alpha_n = o(a_n^{\delta})$, and $\sum_{k=1}^{n} a_k^{1-\delta}\beta_k$ converges.

Using Theorem 2, we now show that condition CH) above is in fact both necessary and sufficient for $x_n - x^* = o(a_n^{\delta})$. Note that Chen's result in [1] applies to algorithms with $f_n = f(x_n) - e_n$ with general nonlinear f (in contrast to ours, which is of the form $f_n = A_n x_n - b_n - e_n$).

Proposition 3: Suppose the assumptions of Theorem 2 hold. Then, $x_n - x^* = o(a_n^{\delta})$ with $0 < \delta \le 1$ if and only if condition CH) holds.

Proof: Setting $\rho_n = a_n^{\delta}$ and applying Theorem 2 with condition N4), we find that $x_n - x^* = o(a_n^{\delta})$ if and only if there exists α'_n and β'_n such that $a_n^{-\delta}e_n = \alpha'_n + \beta'_n$ for all n, $\alpha'_n \to 0$, and $\sum_{k=1}^{n} a_k\beta'_k$ converges. Setting $\alpha_n = a_n^{\delta}\alpha'_n$ and $\beta_n = a_n^{\delta}\beta'_n$, the desired result follows from observing that

$$\alpha'_n \to 0 \qquad \Leftrightarrow \qquad \alpha_n = o(a_n^{\delta})$$

and

$$\sum_{k=1}^{n} a_k\beta'_k \text{ converges} \qquad \Leftrightarrow \qquad \sum_{k=1}^{n} a_k^{1-\delta}\beta_k \text{ converges.}$$

■

V. Final Remarks

We have provided a tight characterization on the noise sequence of a stochastic approximation algorithm for the convergence to achieve a prespecified rate sequence. Our result applies in a general Hilbert space setting, with affine f. It is of interest to do the same with general f in the usual setting, i.e., with $f_n = f(x_n) - e_n$. We believe that our approach will be useful in dealing with this case as well. Specifically, we will need to modify the proof of Theorem 2 along the following lines. First, consider the algorithm

$$x_{n+1} = x_n - a_n f(x_n) + a_n b_n + a_n e_n$$

where $b_n \to 0$. Then, the derivation of the recursion involving $\hat{x}_n = \rho_n^{-1}(x_n - x^*)$ in the proof of Theorem 2 will proceed as

$$\begin{aligned} \hat{x}_{n+1} &= \frac{\rho_n}{\rho_{n+1}}\left(\hat{x}_n - a_n\rho_n^{-1}f(\rho_n\hat{x}_n + x^*) + a_n\frac{b_n}{\rho_n} + a_n\frac{e_n}{\rho_n}\right) \\ &= \hat{x}_n - a_n\frac{\rho_n}{\rho_{n+1}}\left(\rho_n^{-1}f(\rho_n\hat{x}_n + x^*) \right. \\ &\qquad\qquad \left. - \frac{1}{a_n(\rho_n/\rho_{n+1})}\left(\frac{\rho_n}{\rho_{n+1}} - 1\right)\hat{x}_n\right) \\ &\quad + a_n\frac{\rho_n}{\rho_{n+1}}b_n + a_n\frac{\rho_n}{\rho_{n+1}}\frac{e_n}{\rho_n}. \end{aligned}$$

Similar to before, we write $\hat{a}_n = a_n(\rho_n/\rho_{n+1})$, $\hat{e}_n = e_n/\rho_n$, $\hat{b}_n = \rho_n^{-1}b_n$, and

$$f_n(\hat{x}_n) = \rho_n^{-1}f(\rho_n\hat{x}_n + x^*) - \frac{\rho_n - \rho_{n+1}}{a_n\rho_n}\hat{x}_n$$

so that we have

$$\hat{x}_{n+1} = \hat{x}_n - \hat{a}_n f_n(\hat{x}_n) + \hat{a}_n\hat{b}_n + \hat{a}_n\hat{e}_n.$$

We then apply a convergence theorem for a stochastic approximation algorithm of the above form and follow the argument in the proof of

Theorem 2. Note that the sequence of functions $\{f_n\}$ above converges (pointwise) to the linear function $f(x) = (A - cI)x$. We are not currently aware of any convergence result for the above form of stochastic approximation algorithm. However, once such a result is obtained, our rate result will then apply in the general nonlinear case, using the above argument.

Acknowledgment

The authors would like to thank the anonymous reviewers for their helpful comments. They especially thank the reviewer who provided the suggestion for using rate sequences of the form $\rho_n = n^{-\gamma}(\log(n))^{\beta}$ in Sections IV-A and IV-B.

References

[1] H.-F. Chen, "Recent developments in stochastic approximation," in *Proc. IFAC World Congr.,* June 1996, pp. 375–380.

[2] ——, "Stochastic approximation and its new applications," in *Proc. Hong Kong Int. Workshop New Directions of Control and Manufacturing,* 1994, pp. 2–12.

[3] H.-F. Chen and Y.-M. Zhu, "Stochastic approximation procedures with randomly varying truncations," *Scientia Sinica (Series A),* vol. 29, no. 9, pp. 914–926, 1986.

[4] K. L. Chung, "On a stochastic approximation method," *Ann. Math. Stat.,* vol. 25, pp. 463–483, 1954.

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

[6] V. Fabian, "On asymptotic normality in stochastic approximation," *Ann. Math. Stat.,* vol. 39, no. 4, pp. 1327–1332, 1968.

[7] V. F. Gaposhkin and T. P. Krasulina, "On the law of the iterated logarithm in stochastic approximation processes," *Theory of Prob. Its Appl.,* vol. 19, pp. 844–850, 1974.

[8] L. Györfi, "Adaptive linear procedures under general conditions," *IEEE Trans. Inform. Theory,* vol. 30, pp. 262–267, 1984.

[9] A. Heunis, "Rates of convergence for an adaptive filtering algorithm driven by stationary dependent data," *SIAM J. Contr. Optim.,* vol. 32, pp. 116–139, 1994.

[10] M. A. Kouritzin, "On the convergence of linear stochastic approximation procedures," *IEEE Trans. Inform. Theory,* vol. 42, pp. 1305–1309, July 1996.

[11] S. R. Kulkarni and C. S. Horn, "An alternative proof for convergence of stochastic approximation algorithms," *IEEE Trans. Automat. Contr.,* vol. 41, pp. 419–424, Mar. 1996.

[12] H. J. Kushner and H. Huang, "Rates of convergence for stochastic approximation type algorithms," *SIAM J. Contr. Optim.,* vol. 17, no. 5, pp. 607–617, Sept. 1979.

[13] H. K. Kushner and D. S. Clark, *Stochastic Approximation Methods for Constrained and Unconstrained Systems.* New York: Springer, 1978.

[14] T. L. Lai, "Stochastic approximation and sequential search for optimum," in *Proc. Berkeley Conf. Honor of Jerzy Neyman and Jack Kiefer,* L. LeCam and R. A. Olshen, Eds. Monterey, CA: Wadsworth, 1985, vol. 2, pp. 557–577.

[15] L. Ljung, G. Pflug, and H. Walk, *Stochastic Approximation and Optimization of Random Systems.* Boston, MA: Birkhäuser, 1992.

[16] W. Rudin, *Principles of Mathematical Analysis.* New York: McGraw-Hill, 1976.

[17] D. Ruppert, "Almost sure approximations to the Robbins–Monro and Kiefer–Wolfowitz processes with dependent noise," *Ann. Prob.,* vol. 10, pp. 178–187, 1982.

[18] J. Sacks, "Asymptotic distribution of stochastic approximation procedures," *Ann. Math. Stat.,* vol. 29, pp. 373–405, 1958.

[19] H. Walk and L. Zsidó, "Convergence of Robbins–Monro method for linear problems in a Banach space," *J. Math. Analysis and Appl.,* vol. 139, pp. 152–177, 1989.

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

[21] ——, "Weighted averaging and stochastic approximation," *Math. Contr., Sig., and Syst.,* vol. 10, pp. 41–60, 1997.