Equivalent Necessary and Sufficient Conditions on Noise Sequences for Stochastic Approximation Algorithms

I-Jeng Wang* Edwin K. P. Chong*

Sanjeev R. Kulkarni†

To appear in *Adv. Appl. Prob.*, Sept. 1996

Abstract

We consider stochastic approximation algorithms on a general Hilbert space, and study four conditions on noise sequences for their analysis: Kushner and Clark's condition, Chen's condition, a decomposition condition, and Kulkarni and Horn's condition. We discuss various properties of these conditions. In our main result we show that the four conditions are all equivalent, and are both necessary and sufficient for convergence of stochastic approximation algorithms under appropriate assumptions.

Keywords: stochastic approximation, convergence, equivalent necessary and sufficient conditions, noise sequences

AMS Classification: Primary 62L20; Secondary 62L12, 65D99

*School of Electrical Engineering, Purdue University, West Lafayette, IN 47907-1285. E-mail: {iwang, echong}@ecn.purdue.edu. This research was supported in part by a Purdue Research Foundation Fellowship, and by the National Science Foundation through grants ECS-9410313 and ECS-9501652.

†Department of Electrical Engineering, Princeton University, Princeton, NJ 08544. E-mail: kulkarni@ee.princeton.edu. This research was supported in part by the National Science Foundation under grant IRI-9457645 and by the Army Research Office under grant DAAL03-92-G-0320.

1 Introduction

Since stochastic approximation algorithms were first introduced by Robbins and Monro in [13], they have been widely studied and applied in various areas, including stochastic optimization and adaptive control. Results on the convergence of stochastic approximation algorithms constitute a major part of the research in this field; see, for example, [13], [7], [12], [5], [10], [6]. Various sufficient conditions for the convergence of stochastic approximation algorithms have been proposed over the years, with a view of obtaining increasingly weaker conditions. These include the well known condition of Kushner and Clark [7], which has been extensively applied. Recently, in [2], Chen gives a sufficient condition that is similar to that of Kushner and Clark, but appears weaker.

It is of interest to question whether or not the sufficient conditions used in the literature can be further relaxed. If a sufficient condition is also necessary for convergence, we conclude that the condition cannot be further relaxed. In [5], Clark shows that for a particular choice of the step size, a form of the law of large number on noise sequences is both necessary and sufficient for convergence of stochastic approximation algorithms. Chen et al. give a similar result in [3]. Recently, in [6], Kulkarni and Horn present a condition that is necessary and sufficient for convergence of stochastic approximation algorithms for general step size sequences. They show in [6] that Clark's result [5] can be derived from their result.

In this paper, we study four known conditions on noise sequences for convergence of stochastic approximation algorithms. These are Kushner and Clark's condition, Chen's condition, a decomposition condition, and Kulkarni and Horn's condition. We prove that under standard assumptions, all four conditions are equivalent, and are both necessary and sufficient for convergence of stochastic approximation algorithms (see Theorem 1). Our result establishes that the four conditions above cannot be further weakened, and they are all equally weak in some sense. In our proof we use the convergence theorem of [6].

We consider a general stochastic approximation algorithm for finding the zero of

a function $f: \mathcal{H} \to \mathcal{H}$ on a general Hilbert space $\mathcal{H}$:

$$x_{n+1} = x_n - a_n f(x_n) + a_n e_n + a_n \gamma_n, \tag{1}$$

where $x_n \in \mathcal{H}$ is the estimate of the zero of the function f, a_n is a positive scalar referred to as the *step size*, $e_n \in \mathcal{H}$ represents noise originating from estimation or noisy observation of the function value $f(x_n)$, and $\{\gamma_n\}$ is a sequence converging to zero. We assume that the step size satisfies

$$\lim_{n\to\infty} a_n = 0 \qquad \text{and} \qquad \sum_{n=1}^{\infty} a_n = \infty, \tag{2}$$

which is a standard assumption in the literature. The four conditions we consider are deterministic conditions on the noise sequence $\{e_n\}$. To use our results in the stochastic setting, we simply apply the conditions to sample paths of $\{e_n\}$.

The stochastic approximation algorithm in (1) is identical in form to those considered in the literature (e.g., [7], [12]), although the latter are usually considered in a Euclidean (finite dimensional) setting, rather than on a general Hilbert space $\mathcal{H}$. For the case where $\mathcal{H}$ is finite dimensional, we show that the conditions on $\{e_n\}$ hold if and only if they hold for each coordinate of $\{e_n\}$ with respect to a given basis for $\mathcal{H}$ (see Proposition 7).

In the remainder of the paper, we denote the inner product on $\mathcal{H}$ by $\langle \cdot, \cdot \rangle$ and the corresponding induced norm by $\| \cdot \|$. We use $\mathbb{R}$ to denote the real line, and $\mathbb{N}$ the set of natural numbers $\{1, 2, \ldots\}$.

2 Conditions on the Noise Sequence

We consider four conditions on the noise sequence $\{e_n\}$ that have been discussed in the literature. Actually, we first provide extensions of three of the conditions that will be used in our main result. These extensions are motivated by the form of the fourth condition which was introduced in [6]. The conditions all depend explicitly on the step size sequence $\{a_n\}$, and express the notion that the cumulative effect of

the noise should be "small" relative to that of the step size. We present and discuss the four conditions in turn, and study certain special properties of the conditions. Some of these properties are used in the proof of our main result, while others are of independent interest.

2.1 Kushner and Clark's condition

In their well known book [7], Kushner and Clark propose an important sufficient condition on the noise sequence for convergence of stochastic approximation algorithms. The condition has been adopted in numerous applications for establishing the convergence of stochastic approximation algorithms under specific probabilistic assumptions on the noise; see, for example, [12], [8], [11], [9], [1]. We present an extension of the original condition of Kushner and Clark in the following.

Definition 1 Fix a sequence of positive real numbers $\{a_n\}$. We say a sequence $\{e_n\}$ on $\mathcal{H}$ satisfies *Kushner and Clark's condition with height* r, $r \geq 0$, (or simply the KC(r) condition) if for every $T > 0$,

$$\limsup_{n\to\infty} \left(\sup_{n\leq p\leq m(n,T)} \left\| \sum_{i=n}^{p} a_i e_i \right\| \right) \leq rT,$$

where

$$m(n,T) \stackrel{\triangle}{=} \max\{k : a_n + \cdots + a_k \leq T\}.$$

When $r = 0$, the KC(r) condition reduces to the regular condition of Kushner and Clark as stated in the literature (see [12], [10, p. 11]):

$$\lim_{n\to\infty} \left(\sup_{n\leq p\leq m(n,T)} \left\| \sum_{i=n}^{p} a_i e_i \right\| \right) = 0 \quad \text{for all } T > 0.$$

For simplicity, we refer to the above KC(0) condition as the KC condition.

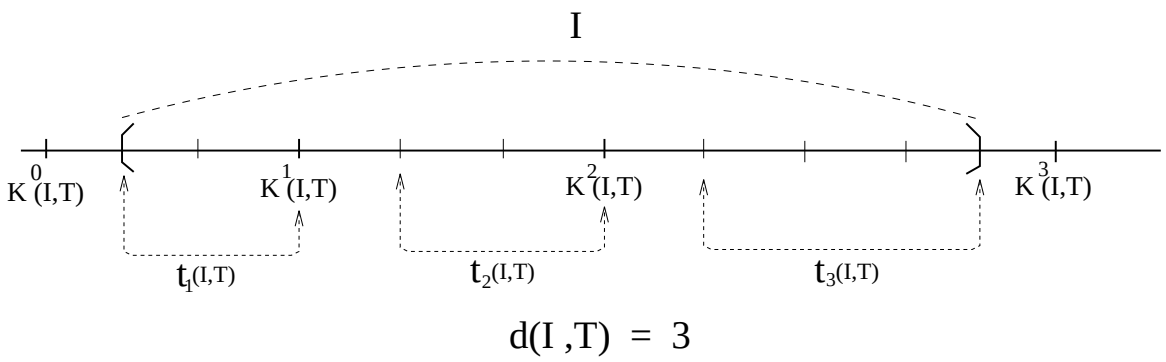

Figure 1: Pictorial description of some notation.

We introduce some notation that will be used in the sequel. We define a class of operations on intervals $I \subset \mathbb{N}$ and $T > 0$ by

$$\begin{aligned} K^0(I,T) &\triangleq \min(I) - 1, \\ K^j(I,T) &\triangleq m\left(K^{j-1}(I,T) + 1, T\right), \quad j \geq 1,\ j \in \mathbb{N}, \\ d(I,T) &\triangleq \min\{j : K^j(I,T) \geq \max(I)\}, \qquad (3) \\ t_j(I,T) &\triangleq \left[K^{j-1}(I,T) + 1, K^j(I,T)\right] \cap I,\ 1 \leq j \leq d(I,T). \qquad (4) \end{aligned}$$

Note that $\{t_j(I,T) : 1 \leq j \leq d(I,T)\}$ is a partition of I that divides the interval I into $d(I,T)$ subintervals, such that the following hold:

$$\begin{aligned} \sum_{n \in t_j(I,T)} a_n &\leq T, \quad 1 \leq j \leq d(I,T), \\ \sum_{n \in t_j(I,T)} a_n + a_{\max(t_j(I,T))+1} &> T, \quad 1 \leq j < d(I,T). \end{aligned}$$

Figure 1 illustrates the above notation. In the sequel, we may drop the arguments I and T of the operators defined above if it involves no confusion.

We are now ready to present a lemma that will be useful later (e.g., in the proof of the main result).

Lemma 1 *Let $\{a_n\}$ be a positive sequence satisfying (2) and let T be a positive real number. Then, for every $\delta \in \mathbb{R}$, $0 < \delta < 1$, there exists $N \in \mathbb{N}$ such that*

$$\sum_{n \in I} a_n > [d(I,T) - 1]\delta T$$

for any interval I on $\mathbb{N}$ with $\min(I) \geq N$.

Proof: Fix $\delta \in \mathbb{R}$, $0 < \delta < 1$, and $T > 0$. Since $a_n \to 0$, there exists $N \in \mathbb{N}$ such that $a_n \leq (1-\delta)T$ for all $n \geq N$. For any interval I with $\min(I) \geq N$, the desired inequality is trivial if $\sum_{n \in I} a_n \leq T$ (since $d(I,T) - 1 = 0$ in this case). Let us assume that $\sum_{n \in I} a_n > T$, i.e., $d(I,T) > 1$. From the definition of t_j and $K^j(I,T)$, we have

$$\sum_{n \in t_j} a_n + a_{\max(t_j)+1} > T,$$

for $1 \leq j \leq d(I,T) - 1$. Furthermore, since $a_n \leq (1-\delta)T$ for all $n \in I$,

$$\sum_{n \in t_j} a_n > T - (1-\delta)T = \delta T.$$

Therefore, we obtain

$$\begin{aligned} \sum_{n \in I} a_n &= \sum_{j=1}^{d} \sum_{n \in t_j} a_n \\ &> \sum_{j=1}^{d-1} \sum_{n \in t_j} a_n \\ &> (d-1)\delta T, \end{aligned}$$

which completes the proof. ∎

The above lemma enables us to simplify the KC(r) condition. The KC(r) condition requires checking that

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

holds for all $T > 0$. In the following proposition we show that it is enough to check that the above holds for a countable number of $T > 0$.

Proposition 1 *Suppose the sequence $\{a_n\}$ satisfies (2). Then, $\{e_n\}$ satisfies the KC(r) condition, $r \geq 0$, if and only if there exist positive sequences $\{T_k\}$ and $\{s_k\}$ converging to 0 such that*

$$\limsup_{n\to\infty} \left(\sup_{n\leq p\leq m(n,T_k)} \left\| \sum_{i=n}^{p} a_i e_i \right\| \right) \leq (r + s_k)T_k$$

for all $k \in \mathbb{N}$.

Proof: Necessity is obvious, so we only prove sufficiency. For convenience, we define

$$M_n(T) \triangleq \sup_{n\leq p\leq m(n,T)} \left\| \sum_{i=n}^{p} a_i e_i \right\|, \ T > 0. \tag{5}$$

It suffices to show that for all $s > 0$ and $T > 0$, we have

$$\limsup_{n\to\infty} M_n(T) \leq (r + s)T.$$

Let $s > 0$ and $T > 0$ be given. Fix $\epsilon > 0$. Since $T_k \to 0$ and $s_k \to 0$, we can find a $k \in \mathbb{N}$ such that $T_k \leq \min(T, \epsilon/2(r + s))$ and $s_k \leq s$. Define a sequence of intervals $\{I_n\}$ and a sequence $\{c_n\}$ on $\mathbb{N}$ by

$$\begin{aligned} I_n &\triangleq [n, m(n, T)], \\ c_n &\triangleq d(I_n, T_k). \end{aligned}$$

Let $\delta < 1$ be a given positive real number. By Lemma 1 and the definition of I_n, there exists $N_1 \in \mathbb{N}$ such that for all $n \geq N_1$,

$$T \geq \sum_{i\in I_n} a_i > (c_n - 1)\delta T_k.$$

Therefore,

$$c_n < \frac{T}{\delta T_k} + 1 \quad \text{for all } n \geq N_1.$$

Furthermore, since $\limsup_{n\to\infty} M_n(T_k) \leq (r+s_k)T_k$, there exists $N \geq N_1$ such that for all $n \geq N$,

$$M_n(T_k) < (r+s)T_k + \left[\frac{\delta T_k}{2(T+\delta T_k)}\right]\epsilon.$$

Applying the triangle inequality, we get that for all $n \geq N$,

$$\begin{aligned}
M_n(T) &\leq \sum_{j=0}^{c_n-1} M_{K^j(I_n,T_k)+1}(T_k) \\
&< c_n\left[(r+s)T_k + \frac{\delta T_k \epsilon}{2(T+\delta T_k)}\right] \\
&< \left(\frac{T}{\delta T_k}+1\right)\left[(r+s)T_k + \frac{\delta T_k \epsilon}{2(T+\delta T_k)}\right] \\
&= \frac{(r+s)T}{\delta} + (r+s)T_k + \frac{\epsilon}{2} \\
&\leq \frac{(r+s)T}{\delta} + \epsilon.
\end{aligned}$$

Therefore, we obtain

$$\limsup_{n\to\infty} M_n(T) \leq \frac{(r+s)T}{\delta}.$$

Since the above is true for all positive $\delta < 1$, we conclude that

$$\limsup_{n\to\infty} M_n(T) \leq (r+s)T,$$

which completes the proof. ∎

For the case where $r = 0$ (i.e., the KC condition), we can further strengthen the above proposition. Specifically, we show that it is enough to check

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

for *some* $T > 0$.

Proposition 2 *Suppose the sequence $\{a_n\}$ satisfies (2). Then, $\{e_n\}$ satisfies the KC condition if and only if*

$$\lim_{n\to\infty}\left(\sup_{n\le p\le m(n,T)}\left\|\sum_{i=n}^{p}a_ie_i\right\|\right)=0 \quad \textit{for some } T>0.$$

Proof: Necessity is obvious. To prove sufficiency, suppose $\lim_{n\to\infty} M_n(T)=0$ for some $T>0$, where $M_n(T)$ is defined in (5). Let $\{T_k\}$ be any sequence converging to 0 such that $T_k\le T$ for all $k\in\mathbb{N}$. From the definition of $M_n(\cdot)$ we have that for all $k\in\mathbb{N}$,

$$M_n(T_k)\le M_n(T) \quad \text{for all } n\in\mathbb{N}.$$

Therefore, $\lim_{n\to\infty} M_n(T_k)=0$ for all k. By Proposition 1, $\{e_n\}$ satisfies the KC condition. ∎

We point out that the above proposition cannot be generalized to the $r>0$ case. That is, the condition

$$\limsup_{n\to\infty}\left(\sup_{n\le p\le m(n,T)}\left\|\sum_{i=n}^{p}a_ie_i\right\|\right)\le rT \tag{6}$$

for some $T>0$ does not imply that $\{e_n\}$ satisfies the KC(r) condition. In the following, we give a counterexample. Let $a_n=1/n$ and $e_n=(-1)^n(1+n)$. Clearly $\{a_n\}$ satisfies (2). It is easy to check that in this case, for any $T>0$,

$$\sup_{n\le p\le m(n,T)}\left|\sum_{i=n}^{p}a_ie_i\right|=|a_ne_n|=\frac{1}{n}+1,$$

and hence,

$$\limsup_{n\to\infty}\left(\sup_{n\le p\le m(n,T)}\left|\sum_{i=n}^{p}a_ie_i\right|\right)=1.$$

Therefore, given any fixed $r>0$, equation (6) holds for $T\ge 1/r$ but not for $T<1/r$. Hence, $\{e_n\}$ does not satisfy the KC(r) condition.

2.2 Chen's condition

In [2], Chen proposes a condition on noise sequences similar to Kushner and Clark's for convergence of general stochastic approximation algorithms. We give a version of the condition in the following definition.

Definition 2 Fix a sequence of positive real numbers $\{a_n\}$. We say a sequence $\{e_n\}$ on $\mathcal{H}$ satisfies *Chen's condition with height* r, $r \geq 0$, (or simply the CH(r) condition) if

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

When $r = 0$, the CH(r) condition reduces to a version of the condition as stated by Chen in [2]:

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

The CH(r) condition, though similar to the KC(r) condition, seems weaker because the latter requires that

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

be bounded by r for *all* $T > 0$. We show in the following proposition that these two conditions are in fact equivalent.

Proposition 3 *Suppose the sequence* $\{a_n\}$ *satisfies (2). Then,* $\{e_n\}$ *satisfies the KC*(r) *condition,* $r \geq 0$*, if and only if it satisfies the CH*(r) *condition.*

Proof: As mentioned above, necessity is clear. To prove sufficiency, we appeal to Proposition 1. Suppose $\{e_n\}$ satisfies the CH(r) condition. Let $\{s_k\}$ be any given positive sequence converging to 0. By CH(r), for each k there exists $\bar{T}_k > 0$ such that for all $T \in (0, \bar{T}_k]$,

$$\frac{1}{T} \limsup_{n\to\infty} \left(\sup_{n\leq p\leq m(n,T)} \left\| \sum_{i=n}^{p} a_i e_i \right\| \right) \leq r + s_k.$$

Let $T_k = \min(\bar{T}_k, \bar{T}_1/k)$, $k \in \mathbb{N}$. By construction, we have $T_k > 0$, $T_k \to 0$, and

$$\limsup_{n\to\infty} \left(\sup_{n\le p\le m(n,T_k)} \left\| \sum_{i=n}^{p} a_i e_i \right\| \right) \le (r + s_k) T_k$$

for all $k \in \mathbb{N}$. By Proposition 1, $\{e_n\}$ satisfies the KC(r) condition. ∎

The above proof suggests that to check the CH(r) condition, it suffices to check the "$\limsup_{T\to 0}$" along some sequence $T_k \to 0$, as stated in the following.

Proposition 4 *Suppose the sequence $\{a_n\}$ satisfies (2). Then, $\{e_n\}$ satisfies the CH(r) condition, $r \ge 0$, if and only if there exists a positive sequence $\{T_k\}$, $T_k \to 0$, such that*

$$\limsup_{k\to\infty} \frac{1}{T_k} \limsup_{n\to\infty} \left(\sup_{n\le p\le m(n,T_k)} \left\| \sum_{i=n}^{p} a_i e_i \right\| \right) \le r.$$

Proof: The above condition is equivalent to the condition stated in Proposition 1, which is equivalent to KC(r) and, hence also CH(r), by Proposition 3. ∎

2.3 Decomposition condition

We now consider a condition on the noise sequence that involves a special decomposition.

Definition 3 Fix a sequence of positive real numbers $\{a_n\}$. We say a sequence $\{e_n\}$ on $\mathcal{H}$ satisfies the *decomposition condition with height r*, $r \ge 0$, (or simply the DC(r) condition) if there exist sequences $\{f_n\}$ and $\{g_n\}$ with $e_n = f_n + g_n$ for all n such that

$$\sum_{k=1}^{n} a_k f_k \text{ converges, and } \limsup_{n\to\infty} \|g_n\| \le r.$$

A special case of the above condition (when $r = 0$) is suggested in [7, p. 29] (see also [4], [2], [10, p. 11], [8] for applications of the above condition for the case $r = 0$). This condition is related to the one presented in [5] and [3], where the particular step

size sequence $a_n = \frac{1}{n}$ was used. Note that for the case $r = 0$, the condition on g_n simplifies to $g_n \to 0$.

We point out that the decomposition in the DC(r) condition, if it exists, is not unique. For example, it is easy to establish the following equivalence.

Proposition 5 *The sequence $\{e_n\}$ satisfies the DC(r) condition if and only if there exist sequences $\{f'_n\}$ and $\{g'_n\}$ with $e_n = f'_n + g'_n$ for all n such that*

$$\lim_{n\to\infty} \sum_{k=1}^{n} a_k f'_k = 0 \text{ and } \limsup_{n\to\infty} \|g'_n\| \leq r.$$

Proof: Sufficiency is trivial. To prove necessity, suppose $\{e_n\}$ satisfies the DC(r) condition; that is, there exist $\{f_n\}$ and $\{g_n\}$ such that $e_n = f_n + g_n$, $\sum_{k=1}^{\infty} a_k f_k$ exists, and $\limsup_{n\to\infty} \|g_n\| \leq r$. Define sequence $\{f'_n\}$ and $\{g'_n\}$ by

$$\begin{aligned} f'_1 &= f_1 - \frac{1}{a_1} \sum_{k=1}^{\infty} a_k f_k, \\ g'_1 &= g_1 + \frac{1}{a_1} \sum_{k=1}^{\infty} a_k f_k, \end{aligned}$$

and, for all $n \geq 2$, $f'_n = f_n$ and $g'_n = g_n$. It is easy to see that the sequences $\{f'_n\}$ and $\{g'_n\}$ satisfy the required conditions. ∎

2.4 Kulkarni and Horn's condition

Recently, in [6], Kulkarni and Horn establish a necessary and sufficient condition for convergence of stochastic approximation algorithms. We state a version of the condition in [6] in the following.

Definition 4 Fix a sequence of positive real numbers $\{a_n\}$. We say a sequence $\{e_n\}$ on $\mathcal{H}$ satisfies *Kulkarni and Horn's condition with height r*, $r \geq 0$, (or simply the KH(r) condition) if for any $\alpha > r$, $\beta > 0$, and any infinite sequence of non-overlapping

intervals $\{I_k\}$ on $\mathbb{N}$ there exists $K \in \mathbb{N}$ such that for all $k \geq K$,

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

The negation of the above condition is what is described in [6] as a "persistently disturbing" condition on the noise sequence. That is, $\{e_n\}$ is said to be *persistently disturbing* of height r if there exist constants $\alpha > r$, $\beta > 0$, and an infinite set of non-overlapping intervals $\{I_k\}$ such that for all k,

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

The intuition behind the KH(r) condition (i.e., not persistently disturbing) is that the cumulative effect of the noise $\sum_{n \in I} a_n e_n$ over any finite interval I should not be too large relative to the cumulative effect of the step size $\sum_{n \in I} a_n$ over the same interval.

Under appropriate assumptions on the function f, Kulkarni and Horn [6] show that stochastic approximation algorithms converge if and only if the noise sequence satisfies KH(r), where r is related to the minimum "height" of the function and the jump of the function at its zero. We make use of their convergence theorem in the proof of our main result in the next section.

We first give a lemma that slightly extends Kulkarni and Horn's result in [6]. The lemma will be used in the proof of our main result.

Lemma 2 *Let $\{a_n\}$ be a sequence of positive real numbers. Suppose*

$$\epsilon_n = e_n + \gamma_n, \qquad \lim_{n \to \infty} \gamma_n = 0.$$

Then, for $r \geq 0$, $\{\epsilon_n\}$ satisfies the KH(r) condition if and only if $\{e_n\}$ satisfies the KH(r) condition.

Proof: We first prove sufficiency. Suppose $\{\epsilon_n\}$ does not satisfy KH(r). Then, there exist $\alpha > r$, $\beta > 0$ and an infinite sequence of non-overlapping intervals $\{I_k\}$ such

that for all k,

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

Since $\gamma_n \to 0$, there exists $N \in \mathbb{N}$ such that for all $n > N$, we have $\|\gamma_n\| \leq \alpha - r - \delta$ with $0 < \delta < \alpha - r$. Let K be any natural number such that $\min(I_K) > N$ and define an infinite subsequence $\{I'_k\}$ of $\{I_k\}$ by $I'_k = I_{K+k-1}$. Then, for all k,

$$\begin{aligned} \left\| \sum_{n \in I'_k} a_n e_n \right\| + \sum_{n \in I'_k} a_n \|\gamma_n\| &\geq \left\| \sum_{n \in I'_k} a_n (e_n + \gamma_n) \right\| \\ &= \left\| \sum_{n \in I'_k} a_n \epsilon_n \right\| \\ &\geq \alpha \sum_{n \in I'_k} a_n + \beta. \end{aligned}$$

Hence,

$$\begin{aligned} \left\| \sum_{n \in I'_k} a_n e_n \right\| &\geq \alpha \sum_{n \in I'_k} a_n - \sum_{n \in I'_k} a_n \|\gamma_n\| + \beta \\ &\geq (\alpha - (\alpha - r - \delta)) \sum_{n \in I'_k} a_n + \beta \\ &= (r + \delta) \sum_{n \in I'_k} a_n + \beta. \end{aligned}$$

Thus, $\{e_n\}$ does not satisfy the KH(r) condition.

To prove necessity, write $e_n = \epsilon_n - \gamma_n$ and apply the proof of the sufficiency part above with γ_n replaced by $-\gamma_n$. ∎

Lemma 2 implies that we can add any sequence that converges to 0 to the noise sequence $\{e_n\}$ without changing the KH(r) property of $\{e_n\}$. This fact is useful in applications where there is a natural decomposition of the noise sequence. Using a similar argument as in the proof of Lemma 2, we can easily show that the same result holds for KC(r), CH(r), and DC(r) as well.

2.5 The finite dimensional case

A special case of interest is where $\mathcal{H}$ is finite dimensional, since this is the case typically studied for stochastic approximations. In this case, we show that to check the conditions previously discussed, it is equivalent to check that they hold when restricted to one dimension. This result is analogous to a result shown in [6] for KH(r).

Proposition 6 *Suppose $\mathcal{H}$ is finite dimensional. Then, $\{e_n\}$ satisfies the KC(r) condition if and only if $\{\langle e_n, v\rangle\}$ satisfies the KC(r) condition for each $v \in \mathcal{H}$, $\|v\| = 1$. The same result holds for CH(r), DC(r), and KH(r).*

Proof: To prove necessity, let $\|v\| = 1$ and suppose that $\{e_n\}$ satisfies the KC(r) condition. By Cauchy-Schwarz's inequality,

$$\left|\sum_{i=n}^{p} a_i \langle e_i, v\rangle\right| = \left|\left\langle \sum_{i=n}^{p} a_i e_i, v \right\rangle\right| \leq \left\|\sum_{i=n}^{p} a_i e_i\right\|.$$

Therefore, for each $T > 0$,

$$\limsup_{n\to\infty}\left(\sup_{n\leq p\leq m(n,T)}\left|\sum_{i=n}^{p} a_i \langle e_i, v\rangle\right|\right) \leq \limsup_{n\to\infty}\left(\sup_{n\leq p\leq m(n,T)}\left\|\sum_{i=n}^{p} a_i e_i\right\|\right) \leq rT.$$

To prove sufficiency, we use contraposition and an argument from [6]. Suppose there exist $T > 0$ and $\delta > 0$ such that

$$\sup_{n\leq p\leq m(n,T)}\left\|\sum_{i=n}^{p} a_i e_i\right\| \geq rT + \delta$$

for infinitely many n. Let I be the infinite set of such n. Let

$$p_n = \arg\sup_{n\leq p\leq m(n,T)}\left\|\sum_{i=n}^{p} a_i e_i\right\|$$

and

$$w_n = \left(\sum_{i=n}^{p_n} a_i e_i\right) \Big/ \left\|\sum_{i=n}^{p_n} a_i e_i\right\|.$$

We have

$$\begin{aligned}\left|\sum_{i=n}^{p_n} a_i \langle e_i, w_n\rangle\right| &= \left|\left\langle \sum_{i=n}^{p_n} a_i e_i, w_n \right\rangle\right| \\ &= \left\|\sum_{i=n}^{p_n} a_i e_i\right\| \\ &\geq rT + \delta\end{aligned}$$

for all $n \in I$. Choose $\epsilon > 0$ such that $(1-\epsilon)(rT+\delta) \geq rT + \frac{\delta}{2}$. Let $C_n = \{v : \|v - w_n\| < \epsilon\}$. Note that for each $n \in I$ and each $v \in C_n$,

$$\begin{aligned}\sup_{n\leq p\leq m(n,T)} \left|\sum_{i=n}^{p} a_i \langle e_i, v\rangle\right| &\geq \left|\left\langle \sum_{i=n}^{p_n} a_i e_i, v \right\rangle\right| \\ &\geq \left|\left\langle \sum_{i=n}^{p_n} a_i e_i, w_n \right\rangle\right| - \left|\left\langle \sum_{i=n}^{p_n} a_i e_i, v - w_n \right\rangle\right| \\ &\geq (1-\epsilon)\left\|\sum_{i=n}^{p_n} a_i e_i\right\| \\ &\geq (1-\epsilon)(rT+\delta) \\ &\geq rT + \frac{\delta}{2}. \qquad (7)\end{aligned}$$

Since $\{w_n : n \in I\}$ is bounded, the Bolzano-Weierstrass Theorem ensures that there exists $N \in I$ such that $w_n \in C_N$ for infinitely many $n \in I$. Therefore, $w_N \in C_n$ for infinitely many $n \in I$. From (7), we have that

$$\sup_{n\leq p\leq m(n,T)} \left|\sum_{i=n}^{p} a_i \langle e_i, w_N\rangle\right| \geq rT + \frac{\delta}{2}$$

for infinitely many n, which completes the proof.

A similar argument applies to CH(r), DC(r), and KH(r). ∎

For the case where $r = 0$, we show that to check the four conditions, it suffices to check them in each coordinate (with respect to a given basis), that is, it is enough to check several one-dimensional conditions in each basis direction. In $\mathbb{R}^{\mathbb{K}}$, for example,

it suffices to check the conditions in each component of $\{e_n\}$.

Proposition 7 *Suppose $\mathcal{H}$ is finite dimensional. Let $e_n^1, \ldots, e_n^K$ be the coordinates of e_n with respect to a given basis for $\mathcal{H}$. Then, $\{e_n\}$ satisfies the KC condition if and only if $\{e_n^k\}$ satisfies the KC condition for each $k = 1, \ldots, K$. The same result holds for CH(0), DC(0), and KH(0).*

Proof: Let $\{v_1, \ldots, v_K\}$ be the given basis for $\mathcal{H}$, that is, $e_n = e_n^1 v_1 + \cdots + e_n^K v_K$.

To prove necessity, note that each e_i^k is a linear combination of $\langle e_i, v_1 \rangle, \ldots, \langle e_i, v_K \rangle$. Thus, the desired result follows easily from Proposition 6.

To prove sufficiency, note that by the triangle inequality,

$$\begin{aligned} \left\| \sum_{i=n}^{p} a_i e_i \right\| &= \left\| \sum_{i=n}^{p} a_i \sum_{k=1}^{K} e_i^k v_k \right\| \\ &= \left\| \sum_{k=1}^{K} \left(\sum_{i=n}^{p} a_i e_i^k \right) v_k \right\| \\ &\leq \sum_{k=1}^{K} \left| \sum_{i=n}^{p} a_i e_i^k \right| \|v_k\|. \end{aligned}$$

Therefore, for each $T > 0$,

$$\lim_{n \to \infty} \left(\sup_{n \leq p \leq m(n,T)} \left\| \sum_{i=n}^{p} a_i e_i \right\| \right) \leq \sum_{k=1}^{K} \|v_k\| \lim_{n \to \infty} \left(\sup_{n \leq p \leq m(n,T)} \left| \sum_{i=n}^{p} a_i e_i^k \right| \right) = 0,$$

which is the desired result.

A similar argument applies to CH(0), DC(0), and KH(0). ∎

3 Equivalence and Convergence Theorem

The four conditions discussed in the last section all express the requirement that the cumulative effect of the noise be small relative to that of the step size. Kushner and Clark's condition, Chen's condition, and the decomposition condition have all

been used as sufficient conditions for convergence of stochastic approximation algorithms. Of the four conditions, Kushner and Clark's is the most widely known and applied, and is often thought to be among the weakest sufficient conditions available. Nonetheless, it has not been resolved whether or not the condition can be weakened any further. Our main result in this section establishes that all four conditions in the previous section are in fact equivalent, and that they are all necessary and sufficient for convergence of stochastic approximation algorithms. In doing so, we confirm that the well used condition of Kushner and Clark cannot be further weakened, generalizing the result of [14].

We consider a class of functions $f: \mathcal{H} \to \mathcal{H}$ that satisfy the following assumptions:

(A1) f is bounded on $\mathcal{H}$; and

(A2) There exist $x^* \in \mathcal{H}$ and $r \geq 0$ such that for all $\delta > 0$, there exists $h_\delta > 0$ such that

$$\|x - x^*\| \geq \delta \quad \text{implies} \quad \langle f(x), x - x^* \rangle \geq (r + h_\delta)\|x - x^*\|.$$

The above assumptions define a fairly general class of functions, quite standard in stochastic approximation settings. The constant $r \geq 0$ in (A2) is related to the minimum "height" of the function and the "height" of the jump of the function at x^*. The case where $r = 0$ includes functions that are continuous at x^*.

We are now ready to present our main result.

Theorem 1 *Consider the stochastic approximation algorithm*

$$x_{n+1} = x_n - a_n f(x_n) + a_n e_n + a_n \gamma_n, \tag{8}$$

where $\{x_n\}$, $\{e_n\}$, *and* $\{\gamma_n\}$ *are sequences on* $\mathcal{H}$, $f: \mathcal{H} \to \mathcal{H}$, $\{a_n\}$ *is a sequence of positive real numbers satisfying (2), and* $\lim_{n\to\infty} \gamma_n = 0$. *The following are equivalent:*

1. *For all* f *satisfying (A1–A2) and all* $x_1 \in \mathcal{H}$, $\lim_{n\to\infty} x_n = x^*$.
2. *For some* f *satisfying* $\lim_{\epsilon\to 0} \sup_{\|u\|<\epsilon} \|f(x^* + u)\| \leq r$ *and some* $x_1 \in \mathcal{H}$, $\lim_{n\to\infty} x_n = x^*$.

3. $\{e_n\}$ *satisfies Kushner and Clark's condition with height* r.

4. $\{e_n\}$ *satisfies Chen's condition with height* r.

5. $\{e_n\}$ *satisfies the decomposition condition with height* r.

6. $\{e_n\}$ *satisfies Kulkarni and Horn's condition with height* r.

Proof: We already have ($3 \iff 4$) by Proposition 3, and ($1 \implies 2$) is obvious because the set of functions satisfying (A1–A2) and $\lim_{\epsilon \to 0} \sup_{\|u\|<\epsilon} \|f(x^* + u)\| \leq r$ is nonempty (see the second remark after the proof). Therefore, it remains to prove the following implications: ($1 \iff 6$), ($4 \implies 6$), ($2 \implies 5$), ($5 \implies 3$).

($\mathbf{1 \iff 6}$):
Kulkarni and Horn established this result in [6] for the case where there is no extra term γ_n. The present result follows directly from the theorem in [6] by applying Lemma 2.

($\mathbf{4 \implies 6}$):
Suppose $\{e_n\}$ satisfies the CH(r) condition. Fix $\alpha > r$, $\beta > 0$, and an infinite sequence of non-overlapping intervals $\{I_k\}$. Choose $\epsilon > 0$ such that $\epsilon < \alpha - r$. Let $M_n(T)$ be defined as in (5). By the CH(r) condition, that is,

$$\limsup_{T \to 0} \frac{1}{T} \limsup_{n \to \infty} M_n(T) \leq r,$$

there exists $T < \frac{\beta}{r+\epsilon}$ such that

$$\limsup_{n \to \infty} M_n(T) \leq \left(r + \frac{\epsilon}{2}\right) T.$$

Therefore, there exists $N \in \mathbb{N}$ such that for all $n > N$

$$M_n(T) < (r + \epsilon)T. \tag{9}$$

By Lemma 1, there exists $K \geq \min\{k: \min(I_k) > N\}$ such that for all $k \geq K$,

$$\sum_{n \in I_k} a_n > [d(I_k, T) - 1] \left(\frac{r + \epsilon}{\alpha} \right) T. \tag{10}$$

Then, for all I_k with $k \geq K$, we have the following two possibilities:

(i) $\sum_{n \in I_k} a_n \leq T$:

By (9) we have

$$\begin{aligned} \left\| \sum_{n \in I_k} a_n e_n \right\| &\leq M_{\min(I_k)}(T) \\ &< (r + \epsilon)T \\ &< \beta \\ &< \alpha \sum_{n \in I_k} a_n + \beta. \end{aligned}$$

(ii) $\sum_{n \in I_k} a_n > T$:

By (9), we have

$$\left\| \sum_{n \in t_j} a_n e_n \right\| \leq M_{\min(t_j)}(T) < (r + \epsilon)T < \beta, \ 1 \leq j \leq d, \tag{11}$$

where d and t_j are as defined in (3) and (4). From (10), (11), and an application of the triangle inequality, we obtain

$$\begin{aligned} \left\| \sum_{n \in I_k} a_n e_n \right\| &\leq \sum_{j=1}^{d-1} \left\| \sum_{n \in t_j} a_n e_n \right\| + \left\| \sum_{n \in t_d} a_n e_n \right\| \\ &< \sum_{j=1}^{d-1} M_{\min(t_j)}(T) + \beta \\ &< (d-1)(r + \epsilon)T + \beta \\ &< \alpha \sum_{n \in I_k} a_n + \beta. \end{aligned}$$

By (i) and (ii), we have

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

for all $k \geq K$. Therefore, $\{e_n\}$ satisfies the KH(r) condition.

($\mathbf{2} \Longrightarrow \mathbf{5}$):

We prove this part by explicitly constructing the required decomposition. Suppose $x_n \to x^*$ for some $x_1 \in \mathcal{H}$ and some f satisfying $\lim_{\epsilon \to 0} \sup_{\|u\| < \epsilon} \|f(x^* + u)\| \leq r$. Following [6], we define sequences

$$\begin{aligned} f_n &\stackrel{\triangle}{=} e_n - f(x_n) = \frac{x_{n+1} - x_n}{a_n}, \\ g_n &\stackrel{\triangle}{=} f(x_n). \end{aligned}$$

Clearly, we have $f_n + g_n = e_n$. Furthermore,

$$\lim_{n \to \infty} \sum_{k=1}^{n} a_n f_n = \lim_{n \to \infty} (x_n - x_1) = x^* - x_1$$

and

$$\limsup_{n \to \infty} \|g_n\| = \limsup_{n \to \infty} \|f(x_n)\| \leq r.$$

($\mathbf{5} \Longrightarrow \mathbf{3}$):

Fix $\epsilon > 0$ and $T > 0$, and suppose $\{e_n\}$ satisfies the DC(r) condition; that is, there exist sequences $\{f_n\}$ and $\{g_n\}$ on $\mathcal{H}$ with $e_n = f_n + g_n$ for all n such that $\sum_{k=1}^{n} a_k f_k$ converges and $\limsup_{n \to \infty} \|g_n\| \leq r$. Then, there exists $N \in \mathbb{N}$ such that for all $m \geq n \geq N$,

$$\left\| \sum_{i=n}^{m} a_i f_i \right\| < \frac{\epsilon}{2}$$

and

$$\|g_n\| < r + \frac{\epsilon}{2T}.$$

Therefore, for all $n \geq N$,

$$\begin{aligned}
\sup_{n \leq p \leq m(n,T)} \left\| \sum_{i=n}^{p} a_i e_i \right\| &\leq \sup_{n \leq p \leq m(n,T)} \left\| \sum_{i=n}^{p} a_i f_i \right\| + \sup_{n \leq p \leq m(n,T)} \left\| \sum_{i=n}^{p} a_i g_i \right\| \\
&< \frac{\epsilon}{2} + \sup_{n \leq p \leq m(n,T)} \sum_{i=n}^{p} a_i \left(r + \frac{\epsilon}{2T} \right) \\
&\leq \frac{\epsilon}{2} + T \left(r + \frac{\epsilon}{2T} \right) \\
&= rT + \epsilon.
\end{aligned}$$

Therefore, $\{e_n\}$ satisfies the KC(r) condition. ∎

Remarks

- In the above theorem, we view all the sequences as deterministic sequences. To apply the theorem to the case of stochastic sequences, we simply interpret the theorem in a sample path fashion. For example, we can state the theorem in the stochastic setting for almost sure convergence of $\{x_n\}$ by having the conditions on $\{e_n\}$ hold almost surely.

- The set of functions satisfying (A1–A2) and $\lim_{\epsilon \to 0} \sup_{\|u\| < \epsilon} \|f(x^* + u)\| \leq r$ is nonempty. An example of such a function is

$$f(x) = \begin{cases} (r + \tanh(\|x - x^*\|))(x - x^*)/\|x - x^*\| & \text{if } x \neq x^* \\ 0 & \text{if } x = x^*. \end{cases}$$

- For the special case $r = 0$, the proof of the equivalence of KC(0), CH(0), KH(0), and DC(0), and their necessity and sufficiency for convergence of $\{x_n\}$, can be considerably simplified. Indeed, that DC(0) implies KC(0), CH(0), and KH(0) is easy to establish, and the necessity of DC(0) for convergence is as shown in ($2 \Longrightarrow 5$) above. Therefore, from the sufficiency of KC(0), CH(0), or KH(0) for convergence, we conclude that all four conditions are equivalent, and are necessary and sufficient for convergence.

- Note that the function f does not appear in any of the four noise conditions. Hence, their equivalence holds without assumptions (A1) or (A2). These assumptions are relevant only in establishing the sufficiency of the four conditions for convergence.

- Although assumption (A1) is somewhat restrictive, for the $r = 0$ case we can eliminate (A1) as follows. In [6], (A1) is needed only to prove sufficiency of KH(0) for convergence. Sufficiency without (A1) has been proved for KC(0) and CH(0) (e.g., [2], [7]). Therefore, by the equivalence of these conditions, we have sufficiency without (A1). For necessity, (A1) is not needed in our proof. Hence, for $r = 0$, the result of Theorem 1 holds without (A1). However, we should point out that without (A1), other additional assumptions may be needed. For example, sufficiency results for KC(0) (e.g., [7]) typically require the assumption that the sequence $\{x_n\}$ is bounded. Alternatively, projections are often incorporated into the algorithm, or growth rate conditions are imposed on f (e.g., [2], [7]).

  For the $r > 0$ case, the only convergence result we know of is that of [6]. In [6], (A1) is used to prove sufficiency of KH(r) for convergence. The proof in [6] goes through if (A1) is replaced by boundedness of $\{x_n\}$. We also believe that, as in the $r = 0$ case, this assumption can be relaxed if an explicit projection mechanism is incorporated, or if growth rate conditions are imposed on f.

- In the convergence theorem of Kushner and Clark in [7, p. 28], the condition on the noise sequence involves the convergence *in probability* of a particular sequence. Specifically, the condition is that for all $\epsilon > 0$ and some $T > 0$,

$$\lim_{n\to\infty} P\left\{\sup_{k\geq n} \sup_{k\leq p\leq m(k,T)} \left\|\sum_{i=k}^{p} a_i e_i\right\| \geq \epsilon\right\} = 0, \tag{12}$$

  where P is the given probability measure. Since almost sure convergence implies convergence in probability, we conclude that if the KC condition holds almost

surely, then for all $\epsilon > 0$ and some $T > 0$,

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

which in turn implies (12). Therefore, the condition of Kushner and Clark in [7, p. 28] is necessary for convergence under the same assumptions as in Theorem 1. In fact, as pointed out in [7, p. 29], the condition in (12) is equivalent to the KC condition holding almost surely.

4 Conclusion

We discussed four conditions on noise sequences that are used in the analysis of stochastic approximation algorithms. We proved that they are all equivalent, and are necessary and sufficient for convergence of stochastic approximation algorithms. Our result establishes that the well known condition of Kushner and Clark is not only sufficient but also necessary for convergence, and therefore cannot be further weakened.

We should point out that to verify these conditions for stochastic noise, application of appropriate devices from probability theory may be necessary. For Kushner and Clark's condition, martingale arguments and application of Doob's type of inequalities constitute a popular approach [12]. With respect to the condition of Kulkarni and Horn, a simple application of Markov's inequality and the Borel-Cantelli lemma is effective for several cases, as pointed out in [6]. Although Kushner and Clark's condition has been well known for some time, the other three conditions we discussed are not as well known. It remains to be seen if they are easier to apply than the condition of Kushner and Clark.

Acknowledgment: We thank an anonymous reviewer for helpful comments.

References

[1] J. D. Bartusek and A. M. Makowski, "On stochastic approximations driven by sample averages: Convergence results via the ODE method," manuscript, Electrical Engineering Department and Institute for Systems Research, University of Maryland, College Park, MD 20742, 1994.

[2] H.-F. Chen, "Stochastic approximation and its new applications," in *Proceedings of 1994 Hong Kong International Workshop on New Directions of Control and Manufacturing*, pp. 2–12, 1994.

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

[4] 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.

[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] S. R. Kulkarni and C. Horn, "Necessary and sufficient conditions for convergence of stochastic approximation algorithms under arbitrary disturbances," manuscript, Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, 1994.

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

[8] T. L. Lai, "Stochastic approximation and sequential search for optimum," in *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer* (L. LeCam and R. A. Olshen, eds.), vol. 2, (Monterey, CA), pp. 557–577, Wadsworth, 1985.

[9] P. L'Ecuyer and P. W. Glynn, "Stochastic optimization by simulation: Convergence proofs for the GI/G/1 queue in steady-state," *Management Science*, vol. 40, no. 11, pp. 1562–1578, 1994.

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

[11] D.-J. Ma, A. M. Makowski, and A. Schwartz, "Stochastic approximations for finite state markov chains," *Stochastic Processes and their Applications*, vol. 35, pp. 27–45, 1990.

[12] M. Metivier and P. Priouret, "Applications of a Kushner and Clark lemma to general classes of stochastic algorithms," *IEEE Transactions on Information Theory*, vol. IT-30, no. 2, pp. 140–151, Mar. 1984.

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

[14] I.-J. Wang, E. K. P. Chong, and S. R. Kulkarni, "Necessity of Kushner–Clark condition for convergence of stochastic approximation algorithms," in *Proceedings of the 32nd Annual Allerton Conference on Communication, Control, and Computing*, pp. 167–175, 1994.