Weighted Averaging and Stochastic Approximation

I-Jeng Wang [*] Edwin K.P. Chong [†]
Sanjeev R. Kulkarni [‡]

To appear in *Mathematics of Control, Signals, and Systems*

Abstract

We explore the relationship between weighted averaging and stochastic approximation algorithms, and study their convergence via a sample-path analysis. We prove that the convergence of a stochastic approximation algorithm is equivalent to the convergence of the weighted average of the associated noise sequence. We also present necessary and sufficient noise conditions for convergence of the average of the output of a stochastic approximation algorithm in the linear case. We show that the averaged stochastic approximation algorithms can tolerate a larger class of noise sequences than the stand-alone stochastic approximation algorithms.

Keywords: stochastic approximation, weighted averaging, convergence, necessary and sufficient noise conditions, noise sequences.

[*]Institute for Systems Research, University of Maryland, College Park, MD 20742. E-mail: iwang@isr.umd.edu.

[†]School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907-1285. E-mail: echong@ecn.purdue.edu. This research was supported 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 by the National Science Foundation through NYI Grant IRI-9457645.

1. Introduction

There has been significant recent interest in using averaging to "accelerate" convergence of stochastic approximation algorithms; see, for example, [GW, KY1, KY2, L1, PJ, SW, Y, YY]. It has been shown that the simple arithmetic average $\frac{1}{n}\sum_{k=1}^{n} x_k$ of the estimates $\{x_n\}$ obtained from a stochastic approximation algorithm converges to the desired point x^* with optimal rate [KY2, PJ]. Under appropriate assumptions, the choice of the step size does not affect this optimal rate of convergence. Most of the results focus on the asymptotic optimality of stochastic approximation algorithms with various averaging schemes.

The central property of the stochastic approximation procedure is its ability to deal with noise. Therefore, from both theoretical and practical points of view, it is important to characterize the set of all possible noise sequences that a stochastic approximation algorithm can tolerate. In [WCK], Wang *et al.* establish four equivalent necessary and sufficient noise condition for convergence of a standard stochastic approximation algorithms, which include the widely-applied condition by Kushner and Clark [KC, M]. Convergence of the weighted average of the noise sequence has been used as a sufficient condition for convergence of stochastic approximation algorithms in [L2, WZ]. In this paper, we prove that this sufficient condition is equivalent to the four necessary and sufficient conditions studied in [WCK], and hence also necessary for convergence of stochastic approximation algorithms (see Theorems 3 and 4). Moreover, we establish necessary and sufficient noise conditions for the convergence of the averaged output of a stochastic approximation algorithm (see Theorem 5). The established noise conditions for convergence of the averaged stochastic approximation algorithms are considerably weaker than the conditions for convergence of the stand-alone stochastic approximation. This result illustrates an important aspect of the averaging scheme: it allows us to relax conditions on noise sequences for convergence of stochastic approximation algorithms. Our analysis is deterministic—we study the sample-path behavior of the algorithms. Note that analysis of stochastic approximation via a deterministic approach has been reported in other work; see, for example, [S2, WZ, KH, Ch1].

In Section 2, we define the weighted averaging operator and introduce two important properties of the operator: regularity and effectiveness. In Section 3, we establish necessary and sufficient conditions on a sequence for convergence of its average. In Section 4, we apply the results in the previous sections to the analysis of stochastic approximation algorithms. In Section 4.1, we establish the convergence of the weighted average of the noise sequence as a necessary and sufficient condition for convergence of the standard stochastic approximation algorithms. In Section 4.2, we present a necessary and sufficient noise condition for

convergence of the averaged stochastic approximation algorithms in the linear case. Finally, we state some conclusions and remarks in Section 5.

2. Weighted Averaging

We first define what we mean by "weighted averaging." Let $\mathbb{H}$ be a real Hilbert space and $\mathbb{L} = \mathbb{H}^{\mathbb{N}}$ be the vector space containing all sequences on $\mathbb{H}$. We denote the inner product on $\mathbb{H}$ by $\langle \cdot, \cdot \rangle$ and the corresponding norm by $\| \cdot \|$, and assume that the index set for elements in $\mathbb{L}$ is $\mathbb{N} = \{1, 2, \cdots\}$. For a sequence $\boldsymbol{x} \in \mathbb{L}$, we write $(\boldsymbol{x})_n$ to denote the nth element of the sequence $\boldsymbol{x}$, and $\boldsymbol{x} \to c$ to mean that $\boldsymbol{x}$ converges to $c \in \mathbb{H}$.

Definition 1. The *weighted averaging* operator with respect to a positive real sequence $\boldsymbol{a} = \{a_n\}$ is the operator $\mathcal{A}_a : \mathbb{L} \longrightarrow \mathbb{L}$ defined by

$$(\mathcal{A}_a \boldsymbol{x})_n = \begin{cases} a_1 x_1 & \text{if } n = 1, \\ (1 - a_n) (\mathcal{A}_a \boldsymbol{x})_{n-1} + a_n x_n & \text{otherwise,} \end{cases} \tag{1}$$

for $\boldsymbol{x} = \{x_n\} \in \mathbb{L}$. Given $\boldsymbol{x} \in \mathbb{L}$, we call $\mathcal{A}_a \boldsymbol{x}$ the *weighted average* of $\boldsymbol{x}$.

We will refer to the sequence $\boldsymbol{a}$ in the definition as the *averaging sequence* of the corresponding weighted average. It is easy to see that the operator $\mathcal{A}_a$ is linear. The following lemma gives a useful representation for the weighted average defined above. Note that this result was established in earlier work; see for example, [LPW, Pa].

Lemma 1. *Given a real sequence $\boldsymbol{a} = \{a_n\}$ satisfying $a_1 = 1$ and $0 < a_n < 1$ for all $n \geq 2$; define real sequences $\{\beta_n\}$ and $\{\gamma_n\}$ by*

$$\beta_n = \begin{cases} 1 & n = 1, \\ \prod_{k=2}^{n} \frac{1}{1 - a_k} & \textit{otherwise.} \end{cases} \tag{2}$$

$$\gamma_n = a_n \beta_n. \tag{3}$$

Then

1. $\beta_n = \sum_{k=1}^{n} \gamma_k$;
2. $\sum_{n=1}^{\infty} a_n = \infty$ *if and only if* $\lim_{n \to \infty} \beta_n = \infty$; *and*
3. $(\mathcal{A}_a \boldsymbol{x})_n = \frac{1}{\beta_n} \sum_{k=1}^{n} \gamma_k x_k$ *for any* $\boldsymbol{x} = \{x_n\} \in \mathbb{L}$.

Example 1. Lemma 1 makes the notion of "weighted averaging" precise by specifying the "weight" γ_n placed on nth term of a sequence as a function of the averaging sequence $\{a_n\}$. Let us look at a class of weighted averaging operator $\mathcal{A}_a$ with $a_n = \frac{1}{n^\alpha}$, $0 \leq \alpha \leq 1$. When $\alpha = 1$, $(\mathcal{A}_a \boldsymbol{x})_n = \frac{1}{n}\sum_{k=1}^{n} x_k$ is the ordinary arithmetic average. When $\alpha < 1$, we have $\gamma_{n+1} > \gamma_n$ for $n \geq 2$. That is, the weight on each term is always larger than the weight on the previous term. As α approaches 0, this corresponds to the case where $\{a_n\}$ is a fixed number, and more and more "weight" is put on the "tail" of the sequence. Nevertheless, for the case where $\alpha > 0$, $\frac{\gamma_{n+1}}{\gamma_n} \to 1$.

In the sequel, we will assume that the averaging sequence satisfies the assumption in Lemma 1 for simplicity of analysis; this assumption is not crucial to the results. All the results hold as long as there exists $N \in \mathbb{N}$ such that $a_n < 1$ for all $n \geq N$. Note that $\frac{\sum_{k=1}^{n}\gamma_k}{\beta_n} \to 1$ in this case.

Suppose that $\boldsymbol{x}$ is a sequence of estimates of an unknown parameter x^*, obtained from some algorithm. There are two motivations behind the application of weighted averaging to the sequence:

1. If $\boldsymbol{x}$ does not converge to x^* but is sufficiently well-behaved, then it may be possible that a weighted average of $\boldsymbol{x}$ converges to x^*.

2. Suppose that $\boldsymbol{x}$ converges to x^* slowly. It may be possible to speed up the convergence by taking the weighted average of $\boldsymbol{x}$.

In other words, weighted averaging serves as a post-filter for the sequence of estimates $\boldsymbol{x}$. In this paper, we focus on the first issue. Specifically, we provide necessary and sufficient conditions on $\boldsymbol{x}$ for convergence of its weighted average. We first define two important properties of a weighted average and give necessary and sufficient conditions for them to hold.

Definition 2. A weighted average $\mathcal{A}_a$ is *regular* if for any sequence $\boldsymbol{x}$ converging to x^*, $\mathcal{A}_a \boldsymbol{x}$ also converges to x^*.

Definition 3. A weighted average $\mathcal{A}_a$ is *effective* if it is regular and $\mathcal{A}_a \boldsymbol{x}$ converges for some non-convergent sequence $\boldsymbol{x}$.

Example 2. A simple example of a regular and effective weighted average is the ordinary arithmetic average, that is, $\mathcal{A}_a$ with $a_n = \frac{1}{n}$. The regularity is straightforward to establish. The effectiveness can be shown by considering the non-convergent sequence $\{x_n\} = \{(-1)^{n+1}\}$, whose average $\mathcal{A}_a \boldsymbol{x}$ converges to 0.

On the other hand, the weighted average with a constant averaging sequence is regular but not effective. To see this, consider the case where the weighted averaging sequence is $\{\epsilon\}$, $0 < \epsilon < 1$. Then for any sequence $\{x_n\}$,

$$x_n = \frac{1}{\epsilon}\bar{x}_n + (1 - \frac{1}{\epsilon})\bar{x}_{n-1},$$

where $\{\bar{x}_n\}$ is the weighted average of $\{x_n\}$. Hence $\{\bar{x}_n\} \to x^*$ implies that $\{x_n\} \to x^*$ Therefore the weighted average is not effective.

The regularity of a weighted averaging operator guarantees that the weighted average of *every* convergent sequence also converges to the same limit of the original sequence—the weighted averaging will not impair convergence. In addition, the effectiveness of a weighted averaging operator makes sure that *some* non-convergent sequence can be made convergent via weighted averaging—the weighted averaging will extend the domain of convergence. We give necessary and sufficient conditions for regularity and effectiveness of weighted averaging in Propositions 1 and 2 below, respectively.

Proposition 1. *A weighted average $\mathcal{A}_a$ with $\boldsymbol{a} = \{a_n\}$ is regular if and only if $\sum_{n=1}^{\infty} a_n = \infty$.*

Proof. ($\Longrightarrow$) If $\sum_{n=1}^{\infty} a_n < \infty$, then from Lemma 1, $\sum_{n=1}^{\infty} \gamma_n = M < \infty$. Let $\boldsymbol{x} = \{x_n\}$ be a sequence on $\mathbb{H}$ defined by

$$x_n = \begin{cases} \eta & \text{if } n = 1, \\ 0 & \text{otherwise} \end{cases}$$

with $\|\eta\| = \frac{M}{2}$. Then $\boldsymbol{x}$ converges to 0, but $\left\|(\mathcal{A}_a\boldsymbol{x})_n\right\| = \beta_n^{-1}\|\eta\| \geq \frac{1}{M}\|\eta\| = \frac{1}{2}$ so that $\mathcal{A}_a$ is not regular.
($\Longleftarrow$) Suppose $\sum_{n=1}^{\infty} a_n = \infty$ and that $\boldsymbol{x} = \{x_n\}$ converges to x^*. Given any $\epsilon > 0$, choose N_1 such that $\left\|x_n - x^*\right\| < \frac{\epsilon}{2}$ for $n \geq N_1$, and choose N_2 such that for all $n \geq N_2$, $\beta_n = \sum_{k=1}^{n} \gamma_k \geq \frac{2}{\epsilon}\left\|\sum_{k=1}^{N_1} \gamma_k(x_k - x^*)\right\|$. Then for all $n \geq \max\{N_1, N_2\}$,

$$\begin{aligned} \left\|\frac{1}{\beta_n}\sum_{k=1}^{n}\gamma_k x_k - x^*\right\| &= \left\|\frac{1}{\beta_n}\sum_{k=1}^{n}\gamma_k(x_k - x^*)\right\| \\ &\leq \frac{1}{\beta_n}\left\|\sum_{k=1}^{N_1}\gamma_k(x_k - x^*)\right\| + \frac{1}{\beta_n}\sum_{k=N_1+1}^{N_2}\gamma_k\|x_k - x^*\| \\ &\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon, \end{aligned}$$

where the first term in the last inequality arises because of the choice of N_2 and the second term arises because of the choice of N_1. □

The next proposition gives us a necessary and sufficient condition for the effectiveness of a weighted average.

Proposition 2. *A regular weighted average $\mathcal{A}_a$ is effective if and only if $\boldsymbol{a}$ has a subsequence converging to 0.*

Proof. ($\Longrightarrow$) Suppose that $\{a_n\}$ does not contain a subsequence that converges to 0. Then there exist $\delta > 0$ and $N < \infty$ such that $a_n > \delta$ for all $n > N$. Let $\boldsymbol{x} = \{x_n\} \in \mathbb{L}$ be a sequence and suppose that $\mathcal{A}_a\boldsymbol{x} = \{\bar{x}_n\}$ converges to x^*. For $n > 1$, we have

$$x_n = \frac{\bar{x}_n - \bar{x}_{n-1}}{a_n} + \bar{x}_{n-1},$$

so that

$$\left\|x_n - x^*\right\| \leq \frac{\left\|\bar{x}_n - \bar{x}_{n-1}\right\|}{a_n} + \left\|\bar{x}_{n-1} - x^*\right\| \to 0.$$

Hence, whenever $\{\bar{x}_n\}$ converges we have that $\{x_n\}$ converges, so that $\mathcal{A}_a$ is not effective. ($\Longleftarrow$) If there is a subsequence of $\{a_n\}$ that converges to zero, we can choose a sequence n_k such that $a_{n_k} < \frac{1}{2^k}$ and $\beta_{n_k} > 4\beta_{n_{k-1}}$ with $\beta_{n_1} > \frac{1}{2}$. Let $\boldsymbol{x} = \{x_n\} \in \mathbb{L}$ satisfy

$$\left\|x_n\right\| = \begin{cases} \frac{1}{2} & \text{if } n = n_k \text{ for some } k, \\ 0 & \text{otherwise.} \end{cases}$$

Clearly $\boldsymbol{x}$ does not converge. However, $\|\bar{x}_{n_1}\| = \frac{1}{\beta_{n_1}} a_{n_1} \|x_{n_1}\| < \frac{1}{2}$ and if $\|x_{n_k}\| < \frac{1}{2^k}$ then

$$\begin{aligned} \|\bar{x}_{n_{k+1}}\| &= \left\|\frac{1}{\beta_{n_{k+1}}}(\beta_{n_k}\bar{x}_{n_k} + \gamma_{n_{k+1}}x_{n_{k+1}})\right\| \\ &\leq \frac{\beta_{n_k}}{\beta_{n_{k+1}}}\left\|\bar{x}_{n_k}\right\| + \frac{1}{2}a_{n_{k+1}} \\ &< \frac{1}{4}\frac{1}{2^k} + \frac{1}{2}\frac{1}{2^{k+1}} \\ &= \frac{1}{2^{k+1}}. \end{aligned}$$

Therefore, by induction $\|\bar{x}_{n_k}\| < \frac{1}{2^k}$ for all k. Finally, for $n_k < n < n_{k+1}$ it is easy to see that $\|\bar{x}_n\| \leq \|\bar{x}_{n_k}\|$. Thus, $\bar{x}_n \to 0$ so that $\mathcal{A}_a$ is effective. □

As an example, we now apply Propositions 1 and 2 to a "forgetting factor" type of averaging that is widely used in the area of parameter estimation.

Example 3. Consider a weighted average operator $\mathcal{A}_a : \mathbb{L} \longrightarrow \mathbb{L}$ with $a_n = \frac{1-\lambda}{1-\lambda^n}$, $\lambda > 0$, $\lambda \neq 1$. By Lemma 1 we can write the weighted average of a sequence $\boldsymbol{x} = \{x_n\}$ as $(\mathcal{A}_a \boldsymbol{x})_n = \beta_n^{-1} \sum_{k=1}^n \gamma_k x_k$ with

$$\begin{aligned} \beta_n &= \frac{1-\lambda^n}{(1-\lambda)\lambda^{n-1}}, \\ \gamma_n &= \lambda^{1-n}. \end{aligned}$$

Note that

$$(\mathcal{A}_a \boldsymbol{x})_n = \beta_n^{-1} \sum_{k=1}^n \gamma_k x_k = \frac{1-\lambda}{1-\lambda^n} \sum_{k=1}^n \lambda^{n-k} x_k$$

is a "forgetting factor" type of average if $\lambda < 1$. By Proposition 1, $\mathcal{A}_a$ is regular if and only if $\lambda < 1$. However, by Proposition 2 this average is not effective since $\boldsymbol{a}$ converges to $1 - \lambda > 0$ monotonically. Nevertheless, this weighted average may improve the speed of convergence.

3. Convergence of Weighted Averages

In this section, we study conditions on $\{x_n\}$ for the convergence of its weighted average. Throughout the paper, we assume that the associated weighted averaging is both regular and effective. In other words, the averaging sequence is not summable and has a subsequence converging to 0. Without loss of generality, we assume that the desired limit for sequences of interest is 0, that is, $x^* = 0$. For ease of presentation, we define operators $\mathcal{S}_a : \mathbb{L} \longrightarrow \mathbb{L}$, $\Theta : \mathbb{L} \longrightarrow \mathbb{L}$, and $\Theta^{-1} : \mathbb{L} \longrightarrow \mathbb{L}$ as follows: For a sequence $\boldsymbol{x} = \{x_n\} \in \mathbb{L}$, define

$$\begin{aligned} (\mathcal{S}_a \boldsymbol{x})_n &= \sum_{k=1}^n a_k x_k \\ (\Theta \boldsymbol{x})_n &= \begin{cases} 0 & n = 1, \\ x_{n-1} & \text{otherwise.} \end{cases} \\ \left(\Theta^{-1} \boldsymbol{x}\right)_n &= x_{n+1}. \end{aligned}$$

3.1. First-Order Condition

We now present a necessary and sufficient condition on a sequence for convergence of its weighted average.

Theorem 1. *Let $\boldsymbol{x} = \{x_n\}$ be a sequence on $\mathbb{H}$. The weighted average $\mathcal{A}_a\boldsymbol{x}$ converges to 0 if and only if there exist sequences $\boldsymbol{u} = \{u_n\}$ and $\boldsymbol{v} = \{v_n\}$ such that $\boldsymbol{x} = \boldsymbol{u}+\boldsymbol{v}$, $\lim_{n\to\infty} u_n = 0$, and $\mathcal{S}_a\boldsymbol{v}$ converges.*

Proof. Let $\mathcal{A}_a\boldsymbol{x} = \{\bar{x}_n\}$, and $\{\beta_n\}$, $\{\gamma_n\}$ sequences defined in (2) and (3), respectively.
($\Longleftarrow$) We have

$$\bar{x}_n = \beta_n^{-1}\sum_{k=1}^{n}\gamma_k x_k = \beta_n^{-1}\sum_{k=1}^{n}\gamma_k u_k + \beta_n^{-1}\sum_{k=1}^{n}\gamma_k v_k.$$

The first term converges to 0 since $\mathcal{A}_a$ is regular. The second term converges to 0 by Kronecker's lemma [Lo, p. 250].
($\Longrightarrow$) ¿From (1) we construct desired sequences $\boldsymbol{u}$ and $\boldsymbol{v}$ by

$$\begin{aligned} u_1 &= 0, & u_n &= \bar{x}_{n-1} \quad \text{for } n \geq 2, \\ v_1 &= \frac{\bar{x}_1}{a_1}, & v_n &= \frac{\bar{x}_n - \bar{x}_{n-1}}{a_n} \quad \text{for } n \geq 2. \end{aligned}$$

Then we have $\boldsymbol{x} = \boldsymbol{u} + \boldsymbol{v}$, and both $\boldsymbol{u}$ and $\mathcal{S}_a\boldsymbol{v}$ converge to 0. □

We define the condition stated in Theorem 1 as the first-order decomposition condition, which we will refer to in the subsequent discussion.

Definition 4. Fix a sequence of positive real numbers $\{a_n\}$. We say a sequence $\boldsymbol{x} \in \mathbb{L}$ satisfies the *first-order decomposition condition* (or simply the DC_a condition) if there exist sequences $\boldsymbol{u} = \{u_n\}$ and $\boldsymbol{v} = \{v_n\}$ such that $\boldsymbol{x} = \boldsymbol{u} + \boldsymbol{v}$, $\lim_{n\to\infty} u_n = 0$, and $\mathcal{S}_a\boldsymbol{v}$ converges.

Note that if $\boldsymbol{a}$ does not have a subsequence converging to 0, the DC_a condition reduces to the convergence of $\boldsymbol{x}$. This fact is consistent with Proposition 2. In the subsequent discussion, we may drop the subscript when the associated averaging sequence does not affect the result.

3.2. Convergence with Zero Upper Density

A notion of convergence considered in [ShW, Wa], called *convergence with zero upper density*, is similar in spirit to the DC condition. In particular, Shapiro and Wardi [ShW] show that,

for a class of stochastic optimization problems, a gradient descent algorithm converges to the minimum with zero upper density. In addition, they state that the average of the iteration sequence converges to the minimum. Here, we further explore the relationship between convergence with zero upper density and weighted averaging. We present a generalization of the former notion of convergence and prove the statement made by Shapiro and Wardi in [ShW] for this generalized notion of convergence.

Definition 5. Let J be a subset of $\mathbb{N}$. The *weighted upper density* of J (with respect to $\{a_n\}$), denoted as $\mathrm{ud}_a(J)$, is defined by

$$\mathrm{ud}_a(J) = \limsup_{n\to\infty} \frac{\sum_{k\in J\cap\{1,\cdots,n\}} \gamma_k}{\beta_n},$$

where γ_n and β_n are defined as in Lemma 1.

The weighted upper density $\mathrm{ud}_a(J)$ of a set J is a measure of the (asymptotic) "density" of the set J as a subset of the positive integers $\{1, 2, 3, \cdots\}$, weighted according to the weighting sequence $\{\gamma_n\}$. For example, if $a_n = \frac{1}{n}$ ($\gamma_n = 1$ for all n) and J is the set of even numbers, then $\mathrm{ud}_a(J) = \frac{1}{2}$. Note that in the case where $a_n = \frac{1}{n}$, the weighted upper density reduces to the upper density considered in [ShW, Wa]. In this case,

$$\mathrm{ud}_a(J) = \limsup_{n\to\infty} \frac{\sum_{k\in J\cap\{1,\cdots,n\}} 1}{n} = \limsup_{n\to\infty} \frac{|J\cap\{1,\cdots,n\}|}{n},$$

where $|\cdot|$ denotes the cardinality of a set.

Based on the above definition of weighted upper density, we define the notion of convergence with zero weighted upper density.

Definition 6. A sequence $\{x_n\}$ on $\mathbb{H}$ is said to *converge with zero weighted upper density* to x^* (with respect to $\{a_n\}$) if there exists a set $J \subset \mathbb{N}$ with $\mathrm{ud}_a(J) = 0$ such that

$$\lim_{n\to\infty, n\notin J} x_n = x^*.$$

Now we prove that the weighted averaging of a bounded sequence converges if the sequence converges with zero weighted upper density.

Proposition 3. *Let $\{x_n\}$ be a bounded sequence on $\mathbb{H}$. Assume that $\{x_n\}$ converges with zero weighted upper density to x^*. If $\mathcal{A}_a$ is regular, then $\mathcal{A}_a\boldsymbol{x} \to x^*$.*

Proof. By the definition of convergence with zero weighted upper density, there exists a subset J of $\mathbb{N}$, with $\mathrm{ud}_a(J) = 0$, such that $\lim_{n\to\infty, n\notin J} x_n = x^*$. Define two sequences $\boldsymbol{u} = \{u_n\}$ and $\boldsymbol{v} = \{v_n\}$ on $\mathbb{H}$ by

$$\begin{aligned} u_n &= \begin{cases} x^* & \text{if } n \in J, \\ x_n & \text{otherwise;} \end{cases} \\ v_n &= x_n - u_n. \end{aligned}$$

Since $\mathcal{A}_a$ is regular and $\boldsymbol{u} \to x^*$, we have $\mathcal{A}_a\boldsymbol{u} \to x^*$. Now let $\|x_n - x^*\| \le M$. Then, we have

$$\begin{aligned} \limsup_{n\to\infty} \left\| \frac{1}{\beta_n} \sum_{k=1}^{n} \gamma_k v_k \right\| &= \limsup_{n\to\infty} \left\| \frac{1}{\beta_n} \sum_{k\in J\cap\{1,\cdots,n\}} \gamma_k (x_k - x^*) \right\|, \\ &\le \limsup_{n\to\infty} M \frac{\sum_{k\in J\cap\{1,\cdots,n\}} \gamma_k}{\beta_n}, \\ &= 0. \end{aligned}$$

Therefore $\mathcal{A}_a\boldsymbol{x} = \mathcal{A}_a\boldsymbol{u} + \mathcal{A}_a\boldsymbol{v} \to x^*$. □

Note that convergence with zero weighted upper density for a bounded sequence is a stronger condition than the convergence of its weighted average. This fact is illustrated by the example we discussed earlier, where $x_n = (-1)^{n+1}$ has a convergent average but does not converge with zero upper density.

3.3. Second-Order Condition

We now study the situation where a second weighted average is needed to obtain a convergent sequence. We present a necessary and sufficient condition on the sequence for convergence of its "second-order weighted average." To establish the result, we need the following lemma.

Lemma 2. *For a sequence $\boldsymbol{x} \in \mathbb{L}$, the following identity holds:*

$$\mathcal{A}_a\boldsymbol{x} = \mathcal{S}_a\boldsymbol{x} - \mathcal{A}_a(\Theta(\mathcal{S}_a\boldsymbol{x})). \tag{4}$$

Proof. Let $\boldsymbol{x} = \{x_n\}$, $\mathcal{A}_a\boldsymbol{x} = \{\bar{x}_n\}$ and $\mathcal{S}_a\boldsymbol{x} = \{y_n\}$. Then by Lemma 1

$$\bar{x}_n = \beta_n^{-1} \sum_{k=1}^{n} \gamma_k x_k,$$

where $\{\beta_n\}$ and $\{\gamma_n\}$ are defined by (2) and (3). Since $x_n = \frac{(y_n - y_{n-1})}{a_n}$, we have

$$\begin{aligned}
\bar{x}_n &= \frac{1}{\beta_1}\gamma_1 y_1 + \frac{1}{\beta_n}\sum_{k=2}^{n}\frac{\gamma_k}{a_k}(y_k - y_{k-1}) \\
&= \frac{1}{\beta_n}\sum_{k=1}^{n-1}(\beta_k - \beta_{k+1})y_k + y_n \\
&= y_n - \frac{1}{\beta_n}\sum_{k=2}^{n}\gamma_k y_{k-1},
\end{aligned}$$

and the desired identity follows. □

We need an additional assumption on the behavior of the averaging sequence $\{a_n\}$ to establish the second-order condition (Theorem 2). We define the notion of *bounded variation* of a sequence that will be used to state our assumption.

Definition 7. A sequence $\{a_n\}$ is said to have *bounded variation* if $\sum_{n=1}^{\infty}|a_{n+1} - a_n| < \infty$.

The set of sequences with bounded variation is a fairly large class of sequences. For example, any bounded and eventually monotone scalar sequence has bounded variation. To establish the second-order condition we need the following lemmas, Lemmas 3 and 4, which concern the relationship among weighted averages with different averaging sequences.

Lemma 3. *Let $\boldsymbol{a} = \{a_n\}$ and $\boldsymbol{b} = \{b_n\}$ be given sequences. Suppose that the sequence $\left\{\frac{a_n}{b_n}\right\}$ has bounded variation. If $\boldsymbol{x}$ satisfies the DC_b condition, then it also satisfies the DC_a condition.*

Proof. Suppose that $\boldsymbol{x} = \{x_n\}$ satisfies the DC_b condition with decomposition $\boldsymbol{x} = \boldsymbol{u} + \boldsymbol{v}$, where $\boldsymbol{u} \to 0$ and $\mathcal{S}_b\boldsymbol{v}$ converges. Let $S_n = \sum_{k=n}^{\infty} b_k v_k$, $n \in \mathbb{N}$. Since $\mathcal{S}_b\boldsymbol{v}$ converges, $\lim_{n\to\infty} S_n = 0$. For $n \geq 1$,

$$\begin{aligned}
(\mathcal{S}_a\boldsymbol{v})_n &= \sum_{k=1}^{n} a_k v_k, \\
&= \sum_{k=1}^{n}\frac{a_k}{b_k}(S_k - S_{k+1}), \\
&= \frac{a_1}{b_1}S_1 + \sum_{k=2}^{n}\left(\frac{a_k}{b_k} - \frac{a_{k-1}}{b_{k-1}}\right)S_k - \frac{a_n}{b_n}S_{n+1}. \qquad (5)
\end{aligned}$$

The first term in (5) is a constant. The third term in (5) converges to 0 since $\left\{\frac{a_n}{b_n}\right\}$ is bounded and $\{S_n\}$ converges to 0. To show that the second term in (5) converges, we prove

that the corresponding sequence is Cauchy:
Let $\sum_{n=1}^{\infty}\left|\frac{a_{n+1}}{b_{n+1}}-\frac{a_n}{b_n}\right| = M$. Given $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that $\left\|S_n\right\| \leq \frac{\epsilon}{M}$ for all $n > N$. For $m \geq n > N$, we obtain

$$\begin{aligned}\left\|\sum_{k=n}^{m}\left(\frac{a_k}{b_k}-\frac{a_{k-1}}{b_{k-1}}\right)S_k\right\| &\leq \sum_{k=n}^{m}\left|\frac{a_k}{b_k}-\frac{a_{k-1}}{b_{k-1}}\right|\frac{\epsilon}{M}\\ &\leq M\frac{\epsilon}{M}\\ &= \epsilon.\end{aligned}$$

Thus $\mathcal{S}_a\boldsymbol{v}$ converges. Therefore $\boldsymbol{x}$ satisfies the DC_a condition. □

Since the DC_a condition is necessary and sufficient for convergence of $\mathcal{A}_a\boldsymbol{x}$, Lemma 3 relates the convergence of weighted averages of a sequence with different averaging sequences. As a direct corollary of the above lemma, we obtain the following useful result.

Lemma 4. *Suppose the sequence* $\left\{\frac{a_{n+1}}{a_n}\right\}$ *has bounded variation. If* $\boldsymbol{x} \in \mathbb{L}$ *satisfies the* DC_a *condition, then* $\Theta\boldsymbol{x}$ *also satisfies the* DC_a *condition.*

Proof. Replacing sequences $\{a_n\}$ and $\{b_n\}$ in Lemma 3 by $\{a_{n+1}\}$ and $\{a_n\}$, respectively, gives us the result. □

Similarly, we can obtain the analogous result for Θ^{-1}:

Lemma 5. *Suppose the sequence* $\left\{\frac{a_n}{a_{n+1}}\right\}$ *has bounded variation. If* $\boldsymbol{x} \in \mathbb{L}$ *satisfies the* DC_a *condition, then* $\Theta^{-1}\boldsymbol{x}$ *also satisfies the* DC_a *condition.*

Note that under the same assumption, Lemmas 4 and 5 can be extended to "nth-order shifts". For example, Lemma 4 (via induction) implies that $\Theta^n\boldsymbol{x}$ satisfies the DC_a condition if $\boldsymbol{x}$ satisfies the DC_a condition.

With the above lemmas, we prove the following theorem that establishes the necessary and sufficient condition for convergence of the "second-order" average.

Theorem 2. *Suppose that the sequences* $\left\{\frac{b_{n+1}}{b_n}\right\}$ *and* $\left\{\frac{a_n}{b_n}\right\}$ *have bounded variation. Then, for* $\boldsymbol{x} \in \mathbb{L}$*, the following are equivalent:*

1. $\mathcal{A}_a\boldsymbol{x}$ *satisfies the* DC_b *condition;*
2. *There exist sequences* $\boldsymbol{u}$ *and* $\boldsymbol{v}$ *such that* $\boldsymbol{x} = \boldsymbol{u} + \boldsymbol{v}$*, and* $\boldsymbol{u}$ *and* $\mathcal{S}_a\boldsymbol{v}$ *satisfy the* DC_b *condition; and*

3. $\mathcal{A}_b(\mathcal{A}_a\boldsymbol{x})$ *converges to* 0.

Proof. With Theorem 1, $(1 \iff 3)$ follows directly. We prove the equivalence between conditions 1 and 2 below.
$(1 \implies 2)$ Suppose that $\mathcal{A}_a\boldsymbol{x}$ satisfies the DC_b condition. Let $\mathcal{A}_a\boldsymbol{x} = \{\bar{x}_n\}$. By (1) we have

$$x_n = \begin{cases} \bar{x}_1 & n = 1, \\ \frac{\bar{x}_n - \bar{x}_{n-1}}{a_n} + \bar{x}_{n-1} & \text{otherwise.} \end{cases}$$

Define $\boldsymbol{u}$ and $\boldsymbol{v}$ by

$$\begin{aligned} u_1 &= 0, & u_n &= \bar{x}_{n-1} \quad \text{for } n \geq 2, \\ v_1 &= \frac{\bar{x}_1}{a_1}, & v_n &= \frac{\bar{x}_n - \bar{x}_{n-1}}{a_n} \quad \text{for } n \geq 2. \end{aligned}$$

Then, $\boldsymbol{x} = \boldsymbol{u} + \boldsymbol{v}$ and $\boldsymbol{u} = \Theta(\mathcal{A}_a\boldsymbol{x})$. By Lemma 4, $\boldsymbol{u}$ satisfies the DC_b condition. Moreover, $\mathcal{S}_a\boldsymbol{v} = \mathcal{A}_a\boldsymbol{x}$ also satisfies the DC_b condition.
$(1 \impliedby 2)$ Suppose that $\boldsymbol{x} = \boldsymbol{u} + \boldsymbol{v}$ where $\boldsymbol{u}$ and $\mathcal{S}_a\boldsymbol{v}$ satisfy the DC_b condition. Since $\mathcal{A}_a$ is linear,

$$\mathcal{A}_a\boldsymbol{x} = \mathcal{A}_a\boldsymbol{u} + \mathcal{A}_a\boldsymbol{v}.$$

Since $\boldsymbol{u}$ satisfies the DC_b condition, $\mathcal{A}_a\boldsymbol{u}$ converges to 0 by Lemma 3 and Theorem 1. By Lemma 2 we have

$$\mathcal{A}_a\boldsymbol{v} = \mathcal{S}_a\boldsymbol{v} - \mathcal{A}_a(\Theta(\mathcal{S}_a\boldsymbol{v})).$$

By Lemma 4, Lemma 3, and Theorem 1, $\mathcal{A}_a(\Theta(\mathcal{S}_a\boldsymbol{v}))$ converges to 0. Hence

$$\mathcal{A}_a\boldsymbol{x} = [\mathcal{A}_a\boldsymbol{u} - \mathcal{A}_a(\Theta(\mathcal{S}_a\boldsymbol{v}))] + \mathcal{S}_a\boldsymbol{v},$$

and $[\mathcal{A}_a\boldsymbol{u} - \mathcal{A}_a(\Theta(\mathcal{S}_a\boldsymbol{v}))]$ converge to 0. Therefore, $\mathcal{A}_a\boldsymbol{x}$ satisfies the DC_b condition. □

We state the condition 2 on $\boldsymbol{x}$ in Theorem 2 in the next definition for later reference.

Definition 8. Fix sequences of positive real numbers $\{a_n\}$ and $\{b_n\}$. We say a sequence $\boldsymbol{x} \in \mathbb{L}$ satisfies the *second-order decomposition condition* (or simply the DC^2_{ab} condition) if there exist sequences $\boldsymbol{u}$ and $\boldsymbol{v}$ such that $\boldsymbol{x} = \boldsymbol{u} + \boldsymbol{v}$, and $\boldsymbol{u}$ and $\mathcal{S}_a\boldsymbol{v}$ satisfy the DC_b condition.

Again, we may omit the subscripts when the associated averaging sequences are clear from the context.

In the next section, we explore the close relationship between weighted averaging and stochastic approximation, and present a necessary and sufficient noise condition for convergence of the averaged stochastic approximation for a class of linear problems.

4. Stochastic Approximation and Averaging

In [WCK], Wang *et al.* show that the DC condition on the noise sequence is necessary and sufficient for convergence of a stochastic approximation algorithm under appropriate assumptions. This result, together with Theorem 1 in the previous section, establishes a form of equivalence between weighted averaging and stochastic approximation in terms of convergence. More precisely, we show that the convergence of weighted average of the noise sequence is necessary and sufficient for convergence of stochastic approximation algorithms (Theorems 3 and 4). Based on this equivalence, we further show that the DC^2 condition on the sequence is a necessary and sufficient condition for convergence of the averaged stochastic approximation algorithm (Theorem 5). These results illustrate an important aspect of the averaging scheme: it allows us to relax the condition on noise sequences for convergence of stochastic approximation algorithms. We prove that, with a weighted averaging, stochastic approximation can tolerate a larger class of noise.

4.1. Weighted Averaging as a Noise Condition

The close relationship between stochastic approximation and weighted averaging has been reported in the literature. In [L2], Ljung shows that convergence of the weighted average of the noise sequence, with the step size being the averaging sequence, is sufficient for convergence of a stochastic approximation algorithm. Walk and Zsidó prove a similar result for a class of linear problems in [WZ]. In [Ke, R1, S2], it is shown that the stochastic approximation algorithm can be represented as a weighted average of the noise sequence when it converges. In the case where the step size $a_n = c/n$, Clark proves in [Cl] that the convergence of the true average of the noise sequence is necessary and sufficient for convergence of Robbins-Monro algorithms. Here, we generalize Clark's result to general step size sequences by applying results in the last section and [WCK].

In [WCK], Wang *et al.* show that the DC_a condition on the noise sequence $\{e_n\}$ is necessary and sufficient for convergence of the stochastic approximation algorithm described

by

$$x_{n+1} = x_n - a_n f(x_n) + a_n e_n + a_n b_n, \tag{6}$$

where $b_n \in \mathbb{H}$ with $b_n \to 0$, and $f\colon \mathbb{H} \to \mathbb{H}$ satisfies

(A) There exists $x^* \in \mathbb{H}$ such that for all $\delta > 0$, there exists $h_\delta > 0$ such that

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

Assumption (A) is a Lyapunov-type condition. In fact, on a Euclidean space $\mathbb{R}^n$, (A) implies a condition used in [De] that

$$\max_{x \in C} \langle f(x), \nabla V(x) \rangle > 0$$

for any compact set $C \subset \mathbb{R}^n \setminus \{x^*\}$ with $V(x) = \frac{1}{2}\|x - x^*\|^2$. Note also that functions satisfying assumption (A) are not necessarily continuous. For example, the function defined by

$$f(x) = \begin{cases} c(x - x^*) + h & \text{if } x \geq x^*, \\ c(x - x^*) & \text{otherwise} \end{cases}$$

with $c > 0$ satisfies assumption (A) but is not continuous at x^*.

The above convergence result, together with Theorem 1, gives us the following theorem that establishes the desired equivalence.

Theorem 3. *Let $\{a_n\}$ satisfy $\sum_{n=1}^{\infty} a_n = \infty$ and $a_n \to 0$. Suppose that $\{x_n\}$ is generated according to the algorithm (6) and $\{f(x_n)\}$ is bounded. Then $x_n \to x^*$ for all f satisfying (A) and all $x_1 \in \mathbb{H}$ if and only if $\mathcal{A}_a \boldsymbol{e} \to 0$.*

Theorem 3 not only shows that convergence of the weighted average of the noise sequence is necessary and sufficient for convergence of a stochastic approximation algorithm; it also establishes the equivalence between this condition and other noise conditions studied in [WCK], which include the well-known condition by Kushner and Clark [KC, M], a modification of Kushner and Clark's condition introduced by Chen in [Ch2], a deterministic condition presented recently by Kulkarni and Horn in [KH], and the decomposition condition.

Note that the boundedness of the sequence $\{f(x_n)\}$ assumed in Theorem 3 can be guaranteed by either assuming that f is Lipschitz continuous and has linear growth rate as in

[S2], or modifying the algorithm (6) to include projections with increasing bounds as in [CLG, Ch1].

We now establish the same equivalence for a class of linear problems, without the assumption that $\{f(x_n)\}$ be bounded. Consider the problem of recursively estimating the zero of an unknown linear function $Ax - b$, $A\colon \mathbb{H} \to \mathbb{H}$ and $b \in \mathbb{H}$, via the following stochastic approximation algorithm

$$x_{n+1} = x_n - a_n A_n x_n + a_n b_n + a_n e_n, \tag{7}$$

where $x_1 \in \mathbb{H}$ is arbitrary; A_n and b_n are estimates of A and b, respectively; and $\{e_n\}$ is the noise sequence. We assume that the step size $\{a_n\}$ is a sequence of nonnegative real number with $a_1 = 1$, $a_n < 1$ for $n \geq 2$, $a_n \to 0$, and $\sum_{n=1}^{\infty} a_n = \infty$. Furthermore we assume that $A_n\colon \mathbb{H} \to \mathbb{H}$ is a sequence of bounded linear operators, and $\{b_n\}$ and $\{e_n\}$ are sequences on the Hilbert space $\mathbb{H}$. Following Walk and Zsidó [WZ], we assume that A_n and b_n satisfy the following assumptions throughout:

(A1) A is a bounded linear operator with $\inf\{\operatorname{Re}\lambda\colon \lambda \in \sigma(A)\} > 0$, where $\sigma(A)$ denotes the spectrum of A.

(A2) $\limsup_{n\to\infty} \frac{1}{\beta_n} \sum_{k=1}^{n} \gamma_k \|A_k\| < \infty$;

(A3) $\left\| \frac{1}{\beta_n} \sum_{k=1}^{n} \gamma_k A_k - A \right\| \to 0$;

(A4) $\left\| \frac{1}{\beta_n} \sum_{k=1}^{n} \gamma_k b_k - b \right\| \to 0$.

Assumption (A1) guarantees that A is invertible. Assumption (A2) is a technical condition that will be used in the proof of convergence. Following Walk and Zsidó [WZ], letting $x'_n = x_n - A^{-1}b$ and $b'_n = b_n - A_n A^{-1} b$, we can rewrite (7) as

$$x'_{n+1} = x'_n - a_n x'_n + a_n b'_n + a_n e_n.$$

Assumptions (A3) and (A4) imply that $\frac{1}{\beta_n} \sum_{k=1}^{n} \gamma_k b'_k$ converges to 0. Therefore we can assume that $b = 0$ without loss of generality. In fact, by Assumption (A4) and the linearity of $\mathcal{A}_a$, we can ignore the term b_n in (7) in considering the convergence of the stochastic approximation algorithm. In other words, we can simply focus on the algorithm described by

$$x_{n+1} = x_n - a_n A_n x_n + a_n e_n. \tag{8}$$

This will be clear when we present our convergence results (Theorem 4 and 5) later.

In the following, we show that convergence of the weighted average of the noise is necessary and sufficient for convergence of the algorithm described by (7). The sufficiency is proved by Walk and Zsidó in [WZ]; we only show the necessity here.

Theorem 4. *Suppose that assumptions (A1–3) hold. Then* $\{x_n\}$ *defined by* (7) *converges to* $A^{-1}b$ *if and only if* $\mathcal{A}_a\boldsymbol{e}$ *converges to* 0.

Proof. We prove the result for (8). Direct applications of Assumption (A4) and the linearity of $\mathcal{A}_a$ yield the desired result.
($\Longleftarrow$) See Walk and Zsidó [WZ].
($\Longrightarrow$) Suppose that $\{x_n\}$ converges to 0 for (8). From (8) we obtain

$$e_n = \frac{x_{n+1} - x_n}{a_n} + A_n x_n.$$

Since $\sum_{k=1}^{n} a_n \frac{x_{k+1}-x_k}{a_n} = x_{n+1} - x_1$ converges, $\{\frac{x_{n+1}-x_n}{a_n}\}$ satisfies the DC_a condition. Hence $\mathcal{A}_a\{\frac{x_{n+1}-x_n}{a_n}\} \to 0$ by Theorem 1. Furthermore, by assumption (A2) and the convergence of $\{x_n\}$, we have $\mathcal{A}_a\{A_n x_n\} \to 0$. Therefore $\mathcal{A}_a\{e_n\} = \mathcal{A}_a\{\frac{x_{n+1}-x_n}{a_n}\} + \mathcal{A}_a\{A_n x_n\} \to 0$. □

The noise condition $\mathcal{A}_a\boldsymbol{e} \to 0$ presented in Theorems 3 and 4 is a deterministic condition. We provide two examples here to illustrate how the condition can be verified for stochastic noise. We assume that the noise sequence $\{e_n\}$ is a random process defined on an appropriate probability space in the following two examples.

Example 4. Consider an independent process $\{e_n\}$. Suppose that $\mathcal{A}_a\{m_n\} \to 0$ and $\sum_{k=1}^{n} a_k^2\sigma_n^2 < \infty$, where m_n and σ_n denote the mean and variance of e_n, respectively. Then the process $\{\sum_{k=1}^{n} a_k(e_k - m_k)\}$ is a martingale (with respect to the filtration generated by $\{e_1, \cdots, e_n\}$). Moreover, we have

$$E\left(\sum_{k=1}^{n} a_k(e_k - m_k)\right)^2 = \sum_{k=1}^{n} a_k^2\sigma_k^2 < \infty.$$

Hence, by the martingale convergence theorem, we have that $\sum_{k=1}^{n} a_k(e_k - m_k)$ converges almost surely. Thus, $\mathcal{A}_a\{e_n - m_n\} \to 0$ by Theorem 1. Therefore, we can conclude that $\mathcal{A}_a\boldsymbol{e} = \mathcal{A}_a\{e_n - m_n\} + \mathcal{A}_a\{m_n\} \to 0$ almost surely.

Example 5. Suppose that $\{e_n\}$ is stationary and ergodic, and $E(e_n) = 0$. Then, in the case where $a_n = \frac{1}{n}$, $\mathcal{A}_a\boldsymbol{e} \to 0$ almost surely by the ergodic theorem [Du]. Note that since any stationary mixing process is ergodic, the condition holds for these processes.

For general weighted average, additional assumptions are needed for $\mathcal{A}_a \boldsymbol{e} \to 0$ to hold. For example, in [Y] Yin considers a stationary φ-mixing process $\{e_n\}$ satisfying $E(e_n) = 0$ and $E|e_n|^{2+\delta} < 0$ for some $\delta > 0$. For $m > 0$, the mixing measure is defined by

$$\varphi(m) = \sup_{B \in \mathcal{F}^{n+m}} |P(B|\mathcal{F}_n) - P(B)|_{(2+\delta)/(1+\delta)},$$

where $\mathcal{F}_n = \sigma\{e_k : k \le n\}$ and $\mathcal{F}^n = \sigma\{e_k : k \ge n\}$. Assuming that $\sum_{m=1}^{\infty} \varphi^{\delta/(1+\delta)}(m) < \infty$, Yin shows that $\sum_{k=1}^{n} k^{-\alpha} e_k$ converges almost surely for $\frac{1}{2} < \alpha \le 1$. Therefore, by Theorem 1, $\mathcal{A}_a \boldsymbol{e} \to 0$ almost surely for $a_n = n^{-\alpha}$.

4.2. Averaged Stochastic Approximation

Recently, Polyak and Ruppert independently proposed the idea of speeding up convergence of stochastic approximation by means of averaging in [Po] and [R2], respectively. They show that the average of the output of a stochastic approximation algorithm, $\frac{1}{n}\sum_{k=1}^{n} x_k$, converges with the optimal rate, together with the optimal asymptotic covariance matrix. The optimality can be achieved with a slowly varying step size, and is independent of the design constant for the step size. Since then, other authors have further explored the benefits of using averaging for stochastic approximation; see for example [GW, KY2, KY1, PJ, S1, SW, Y, YY]. Most of the results focus on the asymptotic optimality of stochastic approximation algorithms with various averaging schemes. Except for results in [S1, SW], a probabilistic approach is used in the analyses.

In this paper, we explore a different aspect of the averaging scheme. We show that the averaging technique, if properly designed, allows us to relax the noise condition for convergence of stochastic approximation. Specifically, we establish a necessary and sufficient noise condition for convergence of an averaged stochastic approximation algorithm in the linear case. This condition is substantially weaker than the known necessary and sufficient noise conditions for convergence of the standard stochastic approximation without averaging. Our analysis is deterministic.

We consider the algorithm described by

$$x_{n+1} = x_n - a_n A x_n + a_n b_n + a_n e_n, \tag{9}$$

and study the convergence of the weighted average $\{\bar{x}_n\}$ of $\{x_n\}$, where

$$\bar{x}_n = (1 - a_n)\bar{x}_{n-1} + a_n x_n. \tag{10}$$

In the following theorem, we present a necessary and sufficient noise condition for convergence of the weighted average of $\{x_n\}$. We will use $\mathcal{A}_a^2\boldsymbol{x} = \mathcal{A}_a(\mathcal{A}_a\boldsymbol{x})$ to denote the second-order weighted averaging of a sequence $\boldsymbol{x}$ with the same averaging sequence $\{a_n\}$ for both averagings.

Theorem 5. *Let $\{a_n\}$ satisfy $\sum_{n=1}^{\infty} a_n = \infty$ and $a_n \to 0$. Suppose that $A\colon \mathbb{H} \to \mathbb{H}$ satisfies assumption (A1), assumption (A4) holds, and $\left\{\frac{a_{n+1}}{a_n}\right\}$ and $\left\{\frac{a_n}{a_{n+1}}\right\}$ have bounded variation. Then, for $\boldsymbol{x}$ and $\{\bar{x}_n\}$ defined by (9) and (10), the following are equivalent:*

1. *$\{\bar{x}_n\} = \mathcal{A}_a\boldsymbol{x}$ converges to $A^{-1}b$*
2. *$\mathcal{A}_a^2\boldsymbol{e}$ converges to 0.*
3. *$\{e_n\}$ satisfies the DC_a^2 condition.*

Proof. We already have $(2 \iff 3)$ by Theorem 2. Therefore we only need to prove $(1 \iff 2)$.

With the help of Lemma 1, we write a recursion for $(\mathcal{A}_a\boldsymbol{x})_n$:

$$(\mathcal{A}_a\boldsymbol{x})_{n+1} = (\mathcal{A}_a\boldsymbol{x})_n - a_{n+1}A(\mathcal{A}_a\boldsymbol{x})_n + a_{n+1}(\mathcal{A}_a\boldsymbol{b})_n + a_{n+1}(\mathcal{A}_a\boldsymbol{e})_n.$$

Since $\mathcal{A}_a$ is regular, $\mathcal{A}_a^2\boldsymbol{b} = \mathcal{A}_a(\mathcal{A}_a\boldsymbol{b}) \to b$. Hence $\mathcal{A}_a\boldsymbol{x} \to 0$ if and only if $\mathcal{A}_{\Theta^{-1}a}(\mathcal{A}_a\boldsymbol{e}) \to 0$ by Theorem 4. By Lemmas 4 and 5, $\mathcal{A}_{\Theta^{-1}a}(\mathcal{A}_a\boldsymbol{e}) \to 0$ if and only if $\mathcal{A}_a^2\boldsymbol{e}$ converges to 0. This completes the proof. □

Note that the assumptions on the step size stated in Theorem 5 hold for the step sizes of the form $\frac{c}{n^\alpha}$, $0 < \alpha \leq 1$.

In the case where different sequences are used for stochastic approximation and weighted averaging, a tight result analogous to Theorem 5 is not easy to obtain. However, with the help of Lemma 3, we can establish a sufficient noise condition for convergence.

Corollary 1. *Let $\{a_n\}$ and $\{c_n\}$ satisfy $\sum_{n=1}^{\infty} a_n = \infty$ and $a_n \to 0$, $\sum_{n=1}^{\infty} c_n = \infty$, respectively. Suppose that $A\colon \mathbb{H} \to \mathbb{H}$ satisfies assumption (A1), assumption (A4) holds, and $\left\{\frac{a_{n+1}}{a_n}\right\}$ and $\left\{\frac{c_n}{a_n}\right\}$ have bounded variation. Then, for $\boldsymbol{x}$ defined by (9), $\mathcal{A}_c\boldsymbol{x}$ converges to $A^{-1}b$ if $\mathcal{A}_a^2\boldsymbol{e}$ converges to 0.*

Note that the sufficient condition in Corollary 1 might be loose when the step size $\{a_n\}$ converges 0 very slowly, that is, when $\mathcal{A}_a$ is nearly ineffective. In this case, the condition $\mathcal{A}_a^2\boldsymbol{e} \to 0$ is almost as strong as the convergence of $\{e_n\}$ to 0. Note also that Theorem 5 and Corollary 1 does not apply when the step size $\{a_n\}$ does not have a subsequence converging

to 0. A specific case of this situation is studied by Györfi and Walk in [GW], where a constant step size ($a_n = \epsilon$) for stochastic approximation and the ordinary arithmetic averaging ($c_n = 1/n$) are considered. Györfi and Walk show that the averaged output $\frac{1}{n}\sum_{k=1}^{n} x_k$ of a stochastic approximation described by

$$x_{n+1} = x_n - \epsilon(A_{n+1}x_n - b_{n+1})$$

converges for small ϵ under the assumptions that the sequence $\{(A_n, b_n)\}$ is stationary and ergodic with $E\|A_n\| < \infty$, $E\|b_n\| < \infty$, and $[E(A_n)]^{-1}$ exists, and the random variable $\epsilon\sum_{n=1}^{\infty} \|(I - \epsilon A_n)\cdots(I - \epsilon A_1)(b_0 - A_0 x_0)\|$ is integrable. Note that the above assumptions imply that Assumptions (A1–4) with $a_n = \frac{1}{n}$ hold almost surely, and hence imply the convergence of stochastic approximation algorithm with the step size $a_n = \frac{1}{n}$. This result by Györfi and Walk illustrates an interesting phenomenon that the convergence property of the averaged stochastic approximation is "dominated" by the faster averaging when the chosen step size is not "effective" (the corresponding averaging operator is not effective).

Theorem 5 and Corollary 1 assert that a stochastic approximation algorithm with averaging can tolerate any noise sequence that satisfies the DC^2 condition. Due to the regularity and effectiveness of weighted averaging, it is clear that the second-order averaging $\mathcal{A}_a^2$ is more "powerful" than the first-order averaging $\mathcal{A}_a$, in the sense that the former can transform a larger class of sequences into convergent sequences. In fact, it is straightforward to establish the inclusion relation: $DC_a \subset DC_a^2$, where we abuse the notation by adopting DC_a and DC_a^2 to denote the sets of sequences satisfying the corresponding conditions. Consider an example where $a_n = \frac{1}{n}$ and $x_n = (-1)^{n+1}(2n-1)$. Although the sequence $\boldsymbol{x}$ oscillates with increasing magnitude, we have $\mathcal{A}_a^2\boldsymbol{x} \to 0$. Note that $\mathcal{A}_a\boldsymbol{x} = \{(-1)^{n+1}\}$ does not converge. Since the DC_a condition is necessary and sufficient for convergence of stochastic approximation, the fact that weighted averaging relaxes the noise condition is evident by Theorem 5 and Corollary 1.

We give an example of stochastic noise for which the deterministic noise conditions in Theorem 5 hold almost surely.

Example 6. Consider a stochastic process $\{e_n\}$ defined by

$$e_n = \begin{cases} \xi_1 & \text{if } n = 1, \\ n(\xi_n - \xi_{n-1}) & \text{otherwise,} \end{cases}$$

where $\{\xi_n\}$ is an independent process with $E(\xi_n) = 0$ and $E(\xi_n^2) < \infty$. Suppose that $a_n = \frac{1}{n}$.

¿From Example 4, $\mathcal{A}_a\{\xi_n\} \to 0$ almost surely. Since $\sum_{k=1}^n a_k e_k = \xi_n$, $\mathcal{A}_a^2 \boldsymbol{e} \to 0$ almost surely by Theorem 2. Note that $\{e_n\}$ has increasing variance and is neither independent nor stationary.

5. Conclusion

In this paper, we study properties of weighted averaging and present necessary and sufficient conditions on a sequence for convergence of its average. We view the weighted averaging as a means to weaken the noise condition for convergence of stochastic approximation and present a necessary and sufficient noise condition for convergence of stochastic approximation with averaging for a special linear case. Note that all the results concerning convergence of weighted average in Sections 2 and 3 go through for a Banach space.

Although averaging has been applied to accelerate convergence in [KY1, KY2, L1, PJ], it is not clear that averaging can always guarantee a speedup of convergence in the deterministic setting adopted in this paper. Under a stochastic framework, the averaging allows the use of larger step size and leads to the best scaling as well as the "smallest" covariance. To some extent, the accelerating effect of the averaging is a consequence of central limit theorem. In the deterministic setting, or on each sample path, the situation is quite different. Consider the case where $a_n = \frac{1}{n}$ and $x_n, e_n \in \mathbb{R}$, if $\frac{x_n}{a_n} = nx_n \to 0$, that is, $x_n = o(a_n)$, and $\liminf_{n\to\infty} \| \sum_{k=1}^n x_k \| > 0$, then $\frac{\bar{x}_n}{x_n} \to \infty$. In other words, averaging actually slows down convergence in this situation. Another situation where averaging cannot speed up convergence is when $\{x_n\}$ monotonically decreases to 0. From the equation

$$\frac{1}{n}\sum_{k=1}^{n} x_k = \frac{1}{n}\sum_{k=1}^{n-1} k(x_k - x_{k+1}) + x_n,$$

we see that the average $\{\bar{x}_n\}$ does not converge faster than $\{x_n\}$ since the first term at the right-hand side is always positive. A similar situation is observed by Spall and Cristion in [SC], where averaging is applied to a stochastic control problem, with the intent to improve convergence of a parameter estimate, and results in inferior performance. Since the sample-path performance is the main concern in many real-world applications, it is important to evaluate the averaging scheme under the deterministic framework. A more detailed analysis is needed to characterize situations where a speedup can be achieved.

Acknowledgment: We thank anonymous reviewers for helpful comments and suggestions.

References

[Ch1] H. F. Chen, "Recent developments in stochastic approximation," in *Proceedings of 1996 IFAC World Congress*, pp. 375–380, June 1996.

[Ch2] 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.

[CLG] 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.

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

[De] B. Delyon, "General results on the convergence of stochastic algorithms," *IEEE Transactions on Automatic Control*, vol. 41, no. 9, pp. 1245–1255, Sept. 1996.

[Du] R. Durrett, *Probability: Theory and Examples.* Wadsworth & Brooks/Cole, 1991.

[GW] L. Györfi and H. Walk, "On the averaged stochastic approximation for linear regression," *SIAM J. Control and Optimization*, vol. 34, no. 1, pp. 31–61, Jan. 1996.

[Ke] G. Kersting, "Almost sure approximation of the Robbins-Monro process by sums of independent random variables," *The Annals of Probability*, vol. 5, no. 6, pp. 954–965, 1977.

[KH] S. R. Kulkarni and C. S. Horn, "An alternative proof for convergence of stochastic approximation algorithms," *IEEE Transactions on Automatic Control*, vol. 41, no. 3, pp. 419–424, Mar. 1996.

[KY1] H. J. Kushner and J. Yang, "Stochastic approximation with averaging and feedback: Rapidly convergent "on-line" algorithms," *IEEE Transactions on Automatic Control*, vol. 40, no. 1, pp. 24–34, 1995.

[KY2] H. J. Kushner and J. Yang, "Stochastic approximation with averaging of the iterates: Optimal asymptotic rate of convergence for general processes," *SIAM J. Contr. Opt.*, vol. 31, no. 4, pp. 1045–1062, 1993.

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

[L1] L. Ljung, "Aspects on accelerated convergence in stochastic approximation schemes," in *Proceedings of the 33rd Conference on Decision and Control*, (Lake Buena Vista, FL), pp. 1649–1652, IEEE, 1994.

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

[L2] L. Ljung, "Strong convergence of a stochastic approximation algorithm," *The Annals of Statistics*, no. 3, pp. 680–696, 1978.

[Lo] M. Loève, *Probability Theory I*. Springer-Verlag, 4th ed., 1977.

[M] 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.

[Pa] A. Pakes, "Some remarks on the paper by Theodorescu and Wolff: "Sequential estimation of expectations in the presence of trend"," *Austral. J. Statist.*, vol. 24, pp. 89–97, 1982.

[Po] B. T. Polyak, "New method of stochastic approximation type," *Automat Remote Control*, vol. 51, pp. 937–946, 1990.

[PJ] B. T. Polyak and A. B. Juditsky, "Acceleration of stochastic approximation by averaging," *SIAM J. Contr. Opt.*, vol. 30, no. 4, pp. 838–855, July 1992.

[R1] D. Ruppert, "Almost sure approximations to the Robbins-Monro and Kiefer-Wolfowitz processes with dependent noise," *The Annals of Probability*, vol. 10, no. 1, pp. 178–187, 1982.

[R2] D. Ruppert, "Efficient estimators from a slowly convergent Robbins-Monro process," Tech. Rep., School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY, 1988.

[S1] R. Schwabe, "Stability results for smoothed stochastic approximation procedures," *Z. angew. Math. Mech.*, vol. 73, pp. 639–643, 1993.

[S2] R. Schwabe, "Strong representation of an adaptive stochastic approximation procedure," *Stochastic Processes and their Applications*, vol. 23, pp. 115–130, 1986.

[SW] R. Schwabe and H. Walk, "On a stochastic approximation procedure based on averaging," *Metrika*, 1996, to appear.

[ShW] A. Shapiro and Y. Wardi, "Convergence analysis of gradient descent stochastic algorithms," Manuscript, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, 1995.

[SC] J. C. Spall and J. A. Cristion, "Model-free control of nonlinear stochastic systems in discrete time," in *Proceedings of the 34th IEEE Conference on Decision and Control*, pp. 2199–2204, 1995.

[WZ] 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.

[WCK] 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.

[Wa] Y. Wardi, "Stochastic algorithms with Armijo step sizes for minimization of functions," *Journal of Optimization Theory and Applications*, vol. 64, pp. 399–417, 1990.

[Y] G. Yin, "On extensions of Polyak's averaging approach to stochastic approximation," *Stochastics and Stochastics Reports*, vol. 36, pp. 245–264, 1991.

[YY] G. Yin and K. Yin, "Asymptotically optimal rate of convergence of smoothed stochastic recursive algorithms," *Stochastics and Stochastics Reports*, vol. 47, pp. 21–46, 1994.