Proceedings of the 35th
Conference on Decision and Control
Kobe, Japan • December 1996

Weighted Averaging and Stochastic Approximation

I-JENG WANG * EDWIN K.P. CHONG *
School of Electrical and Computer Engineering
Purdue University, West Lafayette, IN 47907-1285
{iwang, echong}@ecn.purdue.edu

SANJEEV R. KULKARNI†
Department of Electrical Engineering
Princeton University, Princeton, NJ 08544
kulkarni@ee.princeton.edu

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.

1. Introduction

Recently, there has been significant interest in using averaging to "accelerate" convergence of stochastic approximation algorithms; see, for example, [2, 5, 6, 11, 17, 22, 23]. 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 [6, 11]. 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 [19], Wang *et al.* establish four equivalent necessary and sufficient noise condition for convergence of a standard stochastic approximation algorithms. Convergence of the weighted average of the noise sequence has been used as a sufficient condition for convergence of stochastic approximation algorithms in [8, 18]. In this paper, we prove that this sufficient condition is equivalent to the four necessary and sufficient conditions studied in [19], 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. We believe that applications of the weighted averaging techniques presented here are not limited to the area of stochastic approximation. The results will be useful in general parameter estimation problems with uncertainty.

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. Specifically, 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 averages." 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

*This research was supported by the National Science Foundation through grants ECS-9410313 and ECS-9501652.
†This research was supported by the National Science Foundation through NYI Grant IRI-9457645.

0-7803-3590-2/96 $5.00

$x \to c$ to mean that x converges to $c \in \mathbb{H}$.

Definition 1. The *weighted averaging* operator with respect to a positive real sequence $\boldsymbol{a} = \{a_n\}$ (with $a_1 = 1$) is the operator $\mathcal{A}_a : \mathbb{L} \longrightarrow \mathbb{L}$ defined by

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

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. Note that we assume that $a_1 = 1$ only for simplicity of analysis; the assumption is not crucial to the results. 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, [7, 9].

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*

$$\beta_n = \begin{cases} 1 & n = 1, \\ \prod_{k=2}^{n} \frac{1}{1-a_k} & \text{otherwise.} \end{cases} \quad (2)$$

$$\gamma_n = a_n \beta_n. \quad (3)$$

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* $\boldsymbol{x} = \{x_n\} \in \mathbb{L}$.

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 a 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 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 $\boldsymbol{x}$.

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 improve 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 $\sum_{n=1}^{\infty} a_n = \infty$.*

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.*

Note that the notion of *regularity* is an important concept in *summability theory*; see, for example, [3, 12, 21]. A necessary and sufficient condition for the regularity of a general summability method is provided by the Toeplitz Limit Theorem [3, 12]. The above two propositions can in fact be proven by applications of this theorem.

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 an operators $\mathcal{S}_a : \mathbb{L} \longrightarrow \mathbb{L}$, as follows: For a sequence $\boldsymbol{x} = \{x_n\} \in \mathbb{L}$, defined

$$(\mathcal{S}_a \boldsymbol{x})_n = \sum_{k=1}^{n} a_k x_k.$$

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. See [20]. □

We define the condition stated in Theorem 1 as the first-order decomposition condition that 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.

To see how averaging can improve the convergence property of a sequence, consider the following classical example: Suppose $a_n = \frac{1}{n}$ and $x_n = (-1)^{n+1}$. Then $\mathcal{A}_a\boldsymbol{x} \to 0$ although $\boldsymbol{x}$ oscillates between -1 and 1. This situation may correspond to the case where the estimate wanders around the desired parameter value but does not converges to it.

3.2. 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." 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 5. 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 2 and 3, which concern the relationship among weighted averages with different averaging sequences.

Lemma 2. *Let $\boldsymbol{a} = \{a_n\}$ and $\boldsymbol{b} = \{b_n\}$ be given sequences. Suppose that the sequence $\{\frac{a_n}{b_n}\}$ has bounded variation. If $\boldsymbol{x}$ satisfies the DC_b condition, then it also satisfies the DC_a condition.*

Since the DC_a condition is necessary and sufficient for convergence of $\mathcal{A}_a\boldsymbol{x}$, Lemma 2 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 3. *Let $\boldsymbol{a} = \{a_n\}$ and $\boldsymbol{b} = \{b_n\}$ be given sequences. Suppose that the sequence $\{\frac{a_n}{b_n}\}$ has bounded variation. If $\mathcal{A}_b\boldsymbol{x} \to 0$, $\mathcal{A}_a\boldsymbol{x} \to 0$.*

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

Theorem 2. *Suppose that the sequences $\{\frac{b_{n+1}}{b_n}\}$ and $\{\frac{a_n}{b_n}\}$ 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. See [20]. □

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

Definition 6. 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 [19], 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). As mentioned in the beginning, there has been significant interest in using averaging to accelerate convergence of stochastic approximation algorithm. The result here illustrates another 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 [8], 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 [18]. In [4, 13, 16], 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 [1] 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 [19].

In [19], 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

$$x_{n+1} = x_n - a_n f(x_n) + a_n e_n + a_n b_n, \qquad (4)$$

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

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

This 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 (4) 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 $\mathcal{A}_a e \to 0$.*

Proof. See [20]. □

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

$$x_{n+1} = x_n - a_n A_n x_n + a_n b_n + a_n e_n, \qquad (5)$$

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 [18], 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 [18], letting $x'_n = x_n - A^{-1}b$ and $b'_n = b_n - A_n A^{-1} b$, we can rewrite (5) 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 (5) in considering the convergence of the stochastic approximation algorithm. In other words, we can simply focus on the algorithm described by

$$x_{n+1} = x_n - a_n A_n x_n + a_n e_n. \qquad (6)$$

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 (5). Note that the sufficiency is proved by Walk and Zsidó in [18].

Theorem 4. *Suppose that assumptions (A1–3) hold. Then $\{x_n\}$ defined by (5) converges to $A^{-1}b$ if and only if $\mathcal{A}_a e$ converges to 0.*

Proof. See [20]. □

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 [10] and [14], 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 step size. Since then, other authors have further explored the benefits of using averaging for stochastic approximation; see, for example, [2, 6, 5, 11, 15, 17, 22, 23]. Most of the results focus on the asymptotic optimality of stochastic approximation algorithms with various averaging schemes. Except for results in [15, 17], 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

$$x_{n+1} = x_n - a_n A x_n + a_n b_n + a_n e_n, \qquad (7)$$

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

$$\bar{x}_n = (1 - a_n)\bar{x}_{n-1} + a_n x_n. \qquad (8)$$

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

Theorem 5. *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 x and $\{\bar{x}_n\}$ defined by (7) and (8), the following are equivalent:*

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

Proof. See [20]. □

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

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 2, we can establish a sufficient noise condition for convergence.

Corollary 1. *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 x defined by (7), $\mathcal{A}_c x$ converges to $A^{-1}b$ if $\mathcal{A}_a^2 e$ converges to 0.*

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 x oscillates with increasing magnitude, we have $\mathcal{A}_a^2 x \to 0$. Note that $\mathcal{A}_a x = \{(-1)^{n+1}\}$ 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.

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.

Although averaging has been applied to accelerate convergence in [5, 6, 11], it is not clear that averaging can always guarantee a speedup of convergence in the deterministic setting adopted in this paper. Consider the case where $a_n = \frac{1}{n}$ and $x_n, e_n \in \mathbb{R}$, if $\frac{x_n}{a_n} = n x_n \to 0$, 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 more detailed analysis is needed to characterize situations where a speedup can be achieved.

References

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

[2] 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.

[3] G. H. Hardy, *Divergent Series.* Oxford University Press, 1949.

[4] 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.

[5] 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.

[6] 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.

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

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

[9] 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.

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

[11] 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.

[12] R. E. Powell and S. M. Shah, *Summability Theory and Applications.* Van Nostrand Reinhold Company, 1972.

[13] 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.

[14] 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.

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

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

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

[18] 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.

[19] 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*, Sept. 1996, to appear.

[20] I.-J. Wang, E. K. Chong, and S. R. Kulkarni, "Weighted averaging and stochastic approximation," *submitted to Mathematics of Control, Signals, and Systems*, 1996.

[21] A. Wilansky, *Summability Through Functional Analysis.* North-Holland, 1984.

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

[23] 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.