Local convergence theorems for adaptive stochastic approximation schemes

(regression/weighted sums/martingale transform)

T. L. LAI AND HERBERT ROBBINS

Department of Mathematical Statistics, Columbia University, New York, New York 10027; and State University of New York, Stony Brook, New York 11790

Contributed by Herbert Robbins, April 27, 1979

ABSTRACT For the regression model $y = M(x) + \epsilon$, adaptive stochastic approximation schemes of the form $x_{n+1} = x_n - y_n/(nb_n)$ for choosing the levels $x_1, x_2, \dots$ at which $y_1, y_2, \dots$ are observed converge with probability 1 to the unknown root θ of the regression function $M(x)$. Certain local convergence theorems that relate the convergence rate of $x_n - \theta$ to the limiting behavior of the random variables b_n are established.

Let $\epsilon, \epsilon_1, \epsilon_2, \dots$ be independent, identically distributed random variables with $E\epsilon = 0$ and $E\epsilon^2 = \sigma^2 < \infty$. Let $\mathfrak{F}_0 \subset \mathfrak{F}_1 \subset \dots$ be an increasing sequence of σ -fields such that ϵ_i is $\mathfrak{F}_i$ -measurable and is independent of $\mathfrak{F}_{i-1}$ for all $i \geq 1$. Consider the general regression model

$$y_i = M(x_i) + \epsilon_i, \quad i = 1, 2, \dots, \quad [1]$$

in which the regression function $M(x)$ is a Borel function such that

$$M(\theta) = 0, \quad M'(\theta) = \beta > 0 \quad [2]$$

$$\inf_{\delta \leq |x-\theta| \leq 1/\delta} \{M(x)(x-\theta)\} > 0 \quad \text{for all } 0 < \delta < 1, \quad [3]$$

$$|M(x)| \leq c|x| + d \quad \text{for some } c, d > 0 \quad \text{and all } x. \quad [4]$$

The Robbins-Monro stochastic approximation scheme for choosing the levels x_i so that they will eventually approach the unknown root θ of the regression function is defined by

$$x_1 = \mathfrak{F}_0\text{-measurable random variable,}$$

$$x_{i+1} = x_i - y_i/(ib), \quad i = 1, 2, \dots, \quad [5]$$

where b is some positive constant. It is well known that for $b < 2\beta$,

$$\sqrt{n}(x_n - \theta) \rightarrow N(0, \sigma^2/(b(2\beta - b))) \text{ in distribution, and} \quad [6]$$

$$\limsup \sqrt{\frac{n}{2 \log \log n}} |x_n - \theta| = \frac{\sigma}{\sqrt{b(2\beta - b)}} \text{ a.s. (almost surely).} \quad [7]$$

Moreover, if for some $\lambda > 0$

$$M(x) = \beta(x - \theta) + O(|x - \theta|^{1+\lambda}) \quad \text{as } x \rightarrow \theta, \quad [8]$$

then for $b > 2\beta$, there exists a random variable z such that

$$n^{\beta/b}(x_n - \theta) \rightarrow z \text{ a.s.,} \quad [9]$$

and for $b = 2\beta$,

$$\sqrt{\frac{n}{\log n}}(x_n - \theta) \rightarrow N(0, \sigma^2/b^2) \text{ in distribution} \quad [10]$$

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

(cf. ref. 1). Thus, the asymptotically optimal choice of b is β .

In practice, the value $M'(\theta) = \beta$ will rarely be known. It is therefore natural to try estimating β from the data being observed and to use instead of [5] an adaptive stochastic approximation scheme of the form

$$x_1 = \mathfrak{F}_0\text{-measurable random variable}$$

$$x_{i+1} = x_i - y_i/(ib_i), \quad i = 1, 2, \dots, \quad [11]$$

where $b_i > 0$ is some estimate of β based on $x_1, y_1, \dots, x_i, y_i$. The following theorem, which follows from lemma 5 of ref. 1, deals with the a.s. convergence of the stochastic approximation scheme [11].

THEOREM 1. Suppose that in the regression model [1], $M(x)$ satisfies conditions [2] - [4]. Define inductively y_i by [1] and x_{i+1} by [11], where $\{b_i\}$ is a sequence of positive random variables such that

$$b_i \text{ is } \mathfrak{F}_{i-1}\text{-measurable for all } i \geq 1, \quad [12]$$

$$\inf b_i > 0 \quad \text{and} \quad \sup b_i < \infty \text{ a.s.} \quad [13]$$

Then $\lim x_n = \theta$ a.s.

In view of [7], it is natural to ask whether x_n converges to θ at the rate

$$x_n - \theta = O(n^{-1/2}(\log \log n)^{1/2}) \quad [14]$$

on the event $\{\limsup b_n < 2\beta\}$. An affirmative answer is given by the following.

THEOREM 2. With the same notations and assumptions as in Theorem 1, [14] holds a.s. on the event $\{\limsup b_n < 2\beta\}$.

Theorem 2 is an immediate corollary of the following deeper result which says that if the assumptions of Theorem 1 are satisfied, then, with probability 1, a sufficiently long string of b_n not exceeding $(2 - \eta)\beta$ leads to a corresponding string of x_n differing from θ by no more than a constant times $n^{-1/2}(\log \log n)^{1/2}$.

THEOREM 3. With the same notations and assumptions as in Theorem 1, there exists an event Ω_0 such that $P(\Omega_0) = 1$ and all sample points $\omega \in \Omega_0$ have the following property: For every given $0 < \eta < 2$, there exist $C > 0$ and positive integers M, k (depending on ω and η) such that at ω , for all $m \geq M$ and $n \geq mk$,

$$\max_{m \leq j \leq n} b_j \leq (2 - \eta)\beta \Rightarrow |x_j - \theta| \leq Cj^{-1/2} \times (\log \log j)^{1/2} \text{ for all } mk \leq j \leq n. \quad [15]$$

We preface the proof of Theorem 3 with the following.

LEMMA 1. Suppose that $\{b_n\}$ is a sequence of random variables such that [12] holds and

$$\inf |b_n| > 0 \text{ a.s.}, \quad \sup |b_n| < \infty \text{ a.s.}$$

Let $V_n = \sigma^2 \sum_{i=1}^n b_i^{-2}$. Let $S(\cdot)$ be the random function defined by $S(0) = 0$ and $S(t) = \sum_{i=1}^n (\epsilon_i/b_i)$ for $V_n \leq t < V_{n+1}$, $n = 1, 2, \dots$. Assume that $\sigma \neq 0$. Then, redefining the random variables on a new probability space if necessary, there exists a standard Wiener process $w(t)$, $t \geq 0$, such that

$$\lim_{t \rightarrow \infty} \frac{|S(t) - w(t)|}{(t \log \log t)^{1/2}} = 0 \text{ a.s.}$$

Consequently,

$$\limsup \frac{\sum_{i=1}^n (\epsilon_i/b_i)}{(2V_n \log \log n)^{1/2}} = 1 \text{ a.s.}$$

Lemma 1 is a generalization of theorem 2 of ref. 2 and can be proved by using essentially the same argument, applying theorem 3.1 of ref. 3, and noting that for $p > 0$ and $\delta > 0$,

$$E\{|\epsilon_j|^p I[b_j^2 \epsilon_j^2 \geq \delta j] | \mathcal{F}_{j-1}\} = \int_{b_j^2 x^2 \geq \delta j} |x|^p dG(x),$$

where G is the distribution function of ϵ .

Proof of Theorem 3: Without loss of generality, assume that $\theta = 0$. By Theorem 1, with probability 1,

$$\lim x_n = 0. \quad [16]$$

Let $S_n = \sum_{i=1}^n (\epsilon_i/b_i)$. By Lemma 1, with probability 1,

$$S_n = O[(n \log \log n)^{1/2}]. \quad [17]$$

Let Ω_0 be the event on which [16] and [17] both hold and

$$\inf b_i > 0, \quad \sup b_i < \infty. \quad [18]$$

Then $P(\Omega_0) = 1$. From [2] and [16], it follows that on Ω_0

$$M(x_n) = (\beta + \xi_n)x_n, \quad \text{where } \lim \xi_n = 0. \quad [19]$$

Let $d_j = (\beta + \xi_j)/b_j$. Then, as shown in the proof of theorem 3 of ref. 1,

$$x_{n+1} = \beta_{m-1,n} x_m - \sum_{i=m}^n \beta_{in} \epsilon_i / (ib_i), \quad [20]$$

where for $i = 0, 1, \dots, n-1$,

$$\beta_{in} = \prod_{j=i+1}^n (1 - j^{-1} d_j), \quad \beta_{nn} = 1. \quad [21]$$

By partial summation,

$$\left| \sum_{i=m}^n \frac{\beta_{in} \epsilon_i}{ib_i} \right| \leq \left\{ \sum_{i=m}^{n-1} \left| \frac{\beta_{in}}{i} - \frac{\beta_{i+1,n}}{i+1} \right| |S_i| \right\} + \left| \frac{\beta_{nn}}{n} S_n \right| + \left| \frac{\beta_{m-1,n}}{m} S_{m-1} \right|. \quad [22]$$

Let $\omega \in \Omega_0$ and let $0 < \eta < 1$. Choose $1 > \rho > 0$ such that

$$(1 - \rho)/(2 - \eta) = p > \frac{1}{2}. \quad [23]$$

Because $d_j = [\beta + o(1)]/b_j$ at ω by [19], there exists j_0 (depending on ω) such that at ω , for all $j \geq j_0$,

$$d_j/j \leq \frac{1}{2}, \quad (1 - \rho)\beta/b_j \leq d_j \leq 2\beta/b^*, \quad [24]$$

where $b^* = \inf b_i$. By [21] and [24], for $n > i \geq j_0$,

$$\left| \frac{\beta_{in}}{i} - \frac{\beta_{i+1,n}}{i+1} \right| \leq \frac{\beta_{in}}{i^2} + \frac{1}{i+1} \left| \beta_{i+1,n} - \beta_{in} \right| \leq i^{-2} \beta_{i+1,n} (1 + 2\beta/b^*) \text{ at } \omega. \quad [25]$$

From [21], [23], and [24], it follows that at ω

$$\begin{aligned} n > m > j_0 \text{ and } \max_{m \leq j \leq n} b_j &\leq (2 - \eta)\beta \Rightarrow \\ 0 < \beta_{in} &\leq \prod_{j=i+1}^n \left(1 - \frac{1 - \rho}{(2 - \eta)j} \right) \\ &= \gamma_n / \gamma_i \quad \text{for } n > i \geq m - 1, \quad [26] \end{aligned}$$

where

$$\gamma_n = \prod_{j=j_0}^n \left(1 - \frac{p}{j} \right) \sim Dn^{-p} \quad [27]$$

for some $D > 0$. Because $p > 1/2$, we can choose a positive integer k such that

$$p(1 - k^{-1}) > \frac{1}{2}. \quad [28]$$

From [16] - [18], [20], [22], and [25] - [28], it then follows that, by choosing $M(\geq j_0)$ and C sufficiently large, the desired implication [15] holds at ω for all $m \geq M$ and $n \geq m^k$.

COROLLARY 1. With the same notations and assumptions as in Theorem 1, let $\bar{x}_n = n^{-1} \sum_{i=1}^n x_i$. Then almost all sample points ω have the following property: For every given $0 < \eta < 2$, there exist $C > 0$ and positive integers M, k (depending on ω and η) such that at ω , for all $m \geq M$ and $n \geq m^k$,

$$\begin{aligned} \max_{m \leq j \leq n} b_j &\leq (2 - \eta)\beta \Rightarrow \\ |\bar{x}_j - \theta| &\leq C j^{-1/2} (\log \log j)^{1/2} \text{ for all } m^k \leq j \\ &\leq n + \sqrt{n}. \quad [29] \end{aligned}$$

Proof: With probability 1, because $x_n \rightarrow \theta$ by Theorem 1,

$$\sum_{i=1}^m (x_i - \theta) = o(m), \quad \sum_{n \leq i \leq n + \sqrt{n}} (x_i - \theta) = o(\sqrt{n}),$$

and therefore the desired conclusion follows from Theorem 3 by choosing $k \geq 2$.

An obvious choice for b_n in the stochastic approximation scheme [11] is the usual least squares estimate

$$\hat{\beta}_n = \frac{\sum_{i=1}^n (x_i - \bar{x}_n) y_i}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}, \quad [30]$$

at least in the linear case $M(x) = \beta(x - \theta)$. Note that $\hat{\beta}_n$ is $\mathcal{F}_n$ -measurable but not $\mathcal{F}_{n-1}$ -measurable. Hence, to satisfy [12] we should take $b_n = \hat{\beta}_{n-1}$ instead. It seems artificial to have to drop the last observation (x_n, y_n) in estimating β at stage n , and it is therefore desirable to remove the condition [12]. In the sequel we shall replace [12] by the following

Condition C: There exist random variables b'_n and u_n such that (i) $\sum_{i=1}^n u_i^2 < \infty$ a.s., (ii) b_n and u_n are both $\mathcal{F}_{n-1}$ -measurable for all $n \geq 1$, and

$$(iii) \quad P[|b_n - b'_n| = 0 | u_n \epsilon_n] \text{ as } n \rightarrow \infty = 1.$$

THEOREM 4. Suppose that in Theorems 1, 2, and 3 we replace the condition [12] by Condition C. Then the conclusions of Theorems 1, 2, and 3 still hold.

To prove Theorem 4, we shall need

LEMMA 2. Let $\{u_n\}$ be a sequence of random variables such

that $\sum_1^\infty u_n^2 < \infty$ a.s. and u_n is $\mathcal{F}_{n-1}$ -measurable for all $n \geq 1$. Then

- (a) $\sum_1^\infty u_n \epsilon_n$ converges a.s.,
- (b) $\sum_1^\infty u_n^2 \epsilon_n^2 < \infty$ a.s., and therefore
- (c) $\sum_1^n |u_i| \epsilon_i^2 = O(\sqrt{n})$ a.s.

Proof of Theorem 4: By Lemma 2(a), $u_n \epsilon_n \rightarrow 0$ a.s., and therefore Condition C(iii) implies that $b_n - b'_n \rightarrow 0$ a.s. Hence, by [13],

$$\inf b'_i > 0 \quad \text{and} \quad \sup b'_i < \infty \text{ a.s.} \quad [31]$$

We note that

$$\sum_1^n \frac{\epsilon_i}{ib_i} = \sum_1^n \frac{\epsilon_i}{ib'_i} + \sum_1^n \frac{\epsilon_i}{i} \left(\frac{1}{b_i} - \frac{1}{b'_i} \right). \quad [32]$$

Because b'_i is $\mathcal{F}_{i-1}$ -measurable for all $i \geq 1$ and because [31] holds,

$$\sum_1^\infty \epsilon_i / (ib'_i) \text{ converges a.s.} \quad [33]$$

Because $\sum_1^\infty \epsilon_i^2 / i^2 < \infty$ and $\sum_1^\infty u_i^2 \epsilon_i^2 < \infty$ a.s. by Lemma 2(b), it then follows from [13], [31], Condition C(iii), and the Schwarz inequality that

$$\sum_1^\infty \frac{|\epsilon_i| |b_i - b'_i|}{ib_i b'_i} = \sum_1^\infty O\left(|u_i \epsilon_i| \frac{|\epsilon_i|}{i}\right) < \infty \text{ a.s.} \quad [34]$$

By [32], [33], and [34], $\sum_1^\infty \epsilon_i / (ib_i)$ converges a.s., and therefore by lemma 5 of ref. 1,

$$x_n \rightarrow \theta \text{ a.s.} \quad [35]$$

Let $S_n = \sum_1^n \epsilon_i / b_i$, $S'_n = \sum_1^n \epsilon_i / b'_i$. By [13], [31], Condition C(iii), and Lemma 2(c),

$$|S_n - S'_n| = \sum_1^n O(|u_i| \epsilon_i^2) = O(\sqrt{n}) \text{ a.s.} \quad [36]$$

Because $S'_n = O[(n \log \log n)^{1/2}]$ by Lemma 1, it then follows from [36] that

$$S_n = O[(n \log \log n)^{1/2}] \text{ a.s.} \quad [37]$$

In view of [35] and [37], the same argument as that used to prove Theorem 3 shows that the conclusion of Theorem 3 still holds in the present more general case, and therefore so does the conclusion of Theorem 2.

As a corollary of Theorem 4, we obtain

COROLLARY 2. Suppose that in Corollary 1 we replace the condition [12] by Condition C; then the conclusion of Corollary 1 still holds.

In practice, although β is unknown, it is usually possible to give prior bounds $B > B' > 0$ such that $B' < \beta < B$. In this case, a natural choice for b_n is $\hat{\beta}_n$ truncated below by B' and above by B , where $\hat{\beta}_n$ is the least squares estimate of β defined by [30]. Obviously the b_n so chosen satisfies [13], and it can be shown that Condition C also holds. Therefore Theorem 4 and Corollary 2 are applicable, and they provide very useful tools for showing that such b_n indeed converge to β a.s. and for analyzing the asymptotic properties of the adaptive stochastic approximation scheme [11]. The details will be given elsewhere.

This research was supported by the National Science Foundation.

1. Lai, T. L. & Robbins, H. (1979) *Ann. Stat.* **7**, in press.
2. Lai, T. L. & Robbins, H. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 1068-1070.
3. Jain, N. C., Jogdeo, K. & Stout, W. F. (1975) *Ann. Probab.* **3**, 119-145.