The Annals of Statistics
1990, Vol. 18, No. 4, 1817–1828

NONLINEAR STOCHASTIC APPROXIMATION PROCEDURES FOR L_p LOSS FUNCTIONS[1]

BY ZHILIANG YING

University of Illinois at Urbana-Champaign

The classical stochastic approximation problem can be regarded as choosing design points so that the responses are close to some target level in the expected squared distance. Motivated by different loss criteria, a family of stochastic approximation algorithms is proposed. This family has the same simplicity as the classical Robbins–Monro procedure does and contains the latter as a special case. Using appropriate representations and martingale limit theorems, we establish asymptotic properties for this family. Using the semiparametric formulation, lower bounds are obtained for estimating the desired parameters under any adaptive design, showing that the proposed algorithms with appropriate scaling are asymptotically efficient.

1. Introduction. The stochastic approximation method was first introduced by Robbins and Monro (1951). Consider a regression model

$$y_n = M(x_n) + \varepsilon_n, \tag{1.1}$$

where M is an unknown function such that for some θ, $M(\theta) = 0$, $M(x) < 0$ for $x < \theta$, and $M(x) > 0$, for $x > \theta$, and where ε_n are i.i.d. mean zero random disturbances with a common distribution function F. At each stage n, one is faced with a decision of choosing x_n, based upon the previous information $\mathcal{F}_{n-1} = \sigma(y_{n-1}, \ldots, y_1, x_{n-1}, \ldots, x_1)$, so that it will stay as close to θ as possible. The class of procedures that Robbins and Monro (1951) proposed is the following simple recursion

$$x_{n+1} = x_n - a_n y_n, \tag{1.2}$$

where a_n is any predetermined sequence of nonnegative constants satisfying the condition,

$$\sum_{n=1}^{\infty} a_n = \infty, \qquad \sum_{n=1}^{\infty} a_n^2 < \infty. \tag{1.3}$$

They showed that x_n, under certain regularity conditions, converge to θ in probability. Their work was further advanced by Kiefer and Wolfowitz (1952), who provided a similar recursion for finding the design point, under random perturbations, such that the regression function reaches its maximum, by Blum (1954), who showed the strong consistency of the Robbins–Monro procedure and by Chung (1954) and Sacks (1958), who studied the asymptotic

Received June 1988; revised December 1989.
[1]Research supported in part by NSF Grant DMS-86-09819.
AMS 1980 *subject classifications*. Primary 62L20, 62L05; secondary 60F17.
Key words and phrases. Stochastic approximation, Robbins–Monro procedure, sequential design, L_p loss, information bound.

normality of the Robbins–Monro estimate. In particular, Chung (1954) and Sacks (1958) showed that, among some reasonable class of $\{a_n\}$, an asymptotically optimal one is $a_n = 1/(n\beta)$, where $\beta = M'(\theta) > 0$ is the slope of M at θ, and that with this $\{a_n\}$,

$$\sqrt{n}\,(x_n - \theta) \to_{\mathscr{D}} N\left(0, \frac{\sigma^2}{\beta^2}\right). \tag{1.4}$$

Lai and Robbins (1979, 1981), on the other hand, showed that if $a_n \sim 1/(n\beta)$ a.s., then (1.4) still holds; moreover, they also showed how this kind of designs can be adaptively constructed.

Although stochastic approximation procedures are generally regarded as extensions of the classical Newton–Raphson tangent method for finding the root of a given function, their applications and implications are far beyond this limitation. Consider y_n's as the quantities of interest. The root θ of M can be viewed as the unique value, among all possible design points x, such that

$$V(x) = E\{y_n^2 | x_n = x\} = M^2(x) + \sigma^2 \tag{1.5}$$

is minimized, where $\sigma^2 = \mathrm{Var}(\varepsilon_n)$. This formulation includes a particularly useful model for multiperiod control problems in the econometrics literature [cf. Zellner (1971), Anderson and Taylor (1976) and Lai and Robbins (1982)]. Another use of the stochastic approximation method is to solve adaptive control problems in linear systems. This appears in the engineering control literature, starting with the important work of Goodwin, Ramadge and Caines (1981).

Instead of using the expected squared deviation (1.5) as the loss, we may consider different variants. In particular, $|\cdot|^p$, $p \geq 1$, are sensible alternatives and in some situations seem more natural. We shall, in Section 2, propose simple recursions which converge to optimal design points. Asymptotic theory is also established there.

A question concerning stochastic approximation and its variants provided later is whether one can do better by using other procedures, however complicated they might be. The answer, of course, depends upon how much we know about the model. When the density of ε_n belongs to some parametric family, one may "recursify" the maximum likelihood estimate and obtain, in general, a more efficient estimate [cf. Anbar (1973)]. Improvements in efficiency can also be achieved when the density of ε_n is symmetric but otherwise arbitrary [cf. Fabian (1973, 1983)]. Without assuming either parametric form or symmetry of the density function of ε_n, we shall, in Section 3, study their asymptotic bounds. In doing so, we show that the efficient Robbins–Monro scheme, and our variants of it in Sections 2, attain these bounds, and therefore, are asymptotically efficient.

2. L_p loss and stochastic approximations. As we have pointed out earlier, the stochastic approximation problem can be viewed as adjusting x_n so that y_n, related to x_n by (1.1), will stay close to some target value y^* in terms

of mean square error. To extend this, consider the problem of minimizing the expected L_p loss, $E|y_n - y^*|^p$, for some $p \geq 1$. Without loss of generality, assume $y^* = 0$. Let $m_p = m_p(F)$ be the value m such that

$$\int_{-\infty}^{\infty} |t - m|^p \, dF(t) \tag{2.1}$$

is minimized. For $p > 1$, (2.1) is a strictly convex function of m. Thus, m_p is unique. For $p = 1$, $m_p(F)$ is the median of F, which is unique if F is strictly increasing at m_p. Write

$$y_n = R_p(x_n) + e_{p,n}, \tag{2.2}$$

where

$$R_p(x) = M(x) + m_p, \qquad e_{p,n} = \varepsilon_n - m_p. \tag{2.3}$$

Let θ_p be the root of R_p, i.e., $R_p(\theta_p) = 0$ [unique under (2.7a) below]. Define

$$z_{p,n} = \begin{cases} |y_n|^{p-1}, & \text{if } y_n > 0, \\ 0, & \text{if } y_n = 0, \\ -|y_n|^{p-1}, & \text{if } y_n < 0, \end{cases} \tag{2.4}$$

which has a decomposition

$$z_{p,n} = W_p(x_n) + \eta_{p,n}, \tag{2.5}$$

where $W_p(x) = E(z_{p,n}|x_n = x)$ and $\eta_{p,n} = z_{p,n} - E(z_{p,n}|x_n)$. Since $E(\eta_{p,n}|\mathcal{F}_{n-1}) = 0$ and $W_p(\theta_p) = 0$, (2.5) is a stochastic approximation model. Thus, analogous to (1.2), we propose the following recursive scheme for approximating θ_p,

$$x_{n+1} = x_n - a_n z_{p,n}, \tag{2.6}$$

where $a_n \in \mathcal{F}_{n-1}$ are nonnegative satisfying (1.3) almost surely. While the Robbins–Monro scheme is generally regarded as some kind of linear filtering, with x_{n+1} being a linear combination of the previous design point x_n and the last observation y_n, (2.6) may be viewed as a nonlinear stochastic approximation method because in general the $z_{p,n}$ are nonlinear functions of y_n. Moreover, by using convex loss functions other than the pth absolute deviation, schemes similar to (2.6) can readily be obtained.

We list below some conditions that will be used later on.

(2.7a) For any $\alpha > 0$, $\inf_{\alpha \leq |x - \theta_p| \leq \alpha^{-1}} (x - \theta_p) R_p(x) > 0$,

(2.7b) $|M(x)|^{p-1} \leq K(|x| + 1)$, for some constant K,

(2.7c) $\int_{-\infty}^{\infty} |t|^{2(p-1)} \, dF(t) < \infty$,

(2.7d) M is differentiable at θ_p.

Now (2.5) in conjunction with the convergence result for stochastic approximation methods [cf. Robbins and Siegmund (1971)] implies the following theorem.

THEOREM 1. *Let $p \geq 1$ be fixed. Let y_n be defined recursively by* (1.1) *and x_n by* (2.6) *with $a_n \in \mathscr{F}_{n-1}$ satisfying* (1.3) *almost surely. Then, under conditions* (2.7a)–(2.7c),

$$x_n \to \theta_p \quad a.s. \tag{2.8}$$

Different choices of $\{a_n\}$ will result in different rates of convergence. In particular, if we use the asymptotically efficient ones, as being discussed in Lai and Robbins (1979) and in the next section, then we have the following results.

THEOREM 2. *Let $p \geq 1$ be fixed and let x_n and y_n be the same as those of Theorem* 1 *with $a_n \in \mathscr{F}_{n-1}$ to be specified. Suppose conditions* (2.7a)–(2.7d) *are satisfied. For $p < 2$ assume that F is differentiable at m_p, while for $p = 1$ assume further that $F'(m_p) > 0$. Moreover, for $p = 1$, assume*

$$a_n \sim \left[2nM'(\theta_p)F'(m_p)\right]^{-1} \quad a.s. \tag{2.9}$$

and for $p > 1$ assume

$$a_n \sim \left[n(p-1)M'(\theta_p)\int_{-\infty}^{\infty} |t - m_p|^{p-2}\, dF(t)\right]^{-1} \quad a.s. \tag{2.10}$$

Denote

$$\sigma_p = \begin{cases} \left[2M'(\theta_p)F'(m_p)\right]^{-1}, & \text{if } p = 1, \\ \dfrac{\left[\int_{-\infty}^{\infty} |t - m_p|^{2(p-1)}\, dF(t)\right]^{1/2}}{(p-1)M'(\theta_p)\int_{-\infty}^{\infty} |t - m_p|^{p-2}\, dF(t)}, & \text{if } p > 1. \end{cases} \tag{2.11}$$

Then the following weak and strong convergence results hold:

$$\sqrt{n}\,(x_n - \theta_p) \to_{\mathscr{D}} N\left(0, \sigma_p^2\right); \tag{2.12}$$

$$\limsup_{n\to\infty} \sqrt{\frac{n}{2\log\log n}}\, |x_n - \theta_p| = \sigma_p \quad a.s. \tag{2.13}$$

Moveover, for $1 \leq p < 2$,

$$n^{(p/2)-1} \sum_{i=1}^{n} |x_i - \theta_p|^p \to_{\mathscr{D}} \sigma_p^p \int_0^1 \frac{|B(t)|^p}{t^p}\, dt, \tag{2.14}$$

$$n^{(p/2)-1} \sum_{i=1}^{n} |y_i - e_{p,i}|^p \to_{\mathscr{D}} \left|M'(\theta_p)\sigma_p\right|^p \int_0^1 \frac{|B(t)|^p}{t^p}\, dt, \tag{2.15}$$

where $B(\cdot)$ denotes the standard Brownian motion process; for $p = 2$,

$$\frac{1}{\log n}\sum_{i=1}^{n}|x_i-\theta_p|^p \to \sigma_p^p \quad a.s., \tag{2.16}$$

$$\frac{1}{\log n}\sum_{i=1}^{n}|y_i-e_{p,i}|^p \to \left|M'(\theta_p)\sigma_p\right|^p \quad a.s.; \tag{2.17}$$

and for $p > 2$,

$$\sum_{i=1}^{\infty}|x_i-\theta_p|^p < \infty \quad a.s., \tag{2.18}$$

$$\sum_{i=1}^{\infty}|y_i-e_{p,i}|^p < \infty \quad a.s. \tag{2.19}$$

REMARK 1. $\sum_{i=1}^{n}|y_i - e_i|^p$ represents the regret or extra cost due to the ignorance of the parameter θ_p in terms of L_p loss. For $p = 2$, this notion of regret has been discussed in Lai and Wei (1987) in the context of adaptive stochastic control. Note that for $p < 2$, the convergence is in distribution.

REMARK 2. Without prior knowledge of the slope parameter, the sequence $\{a_n\}$ has to be constructed adaptively. This can be accomplished by using the same truncated least squares estimate of $M'(\theta_p)$, as that introduced in Lai and Robbins (1981). Other methods of construction may also be used [see, for example, Martinsek (1988)].

REMARK 3. It will be interesting to see what will be the "L_p" analogues of multivariate and other stochastic approximation procedures such as those studied in Wei (1987), Walk (1977) and Ruppert (1981).

We preface the proof with the following two lemmas.

LEMMA 1. *Let ξ_n be a martingale difference sequence with respect to a σ-filtration $\mathcal{G}_n$ such that*

$$E\left(\xi_n^2\middle|\mathcal{G}_{n-1}\right) \to A \quad a.s., \tag{2.20}$$

for some nonrandom constant A. Assume a Lindeberg-type condition,

$$\sum_{i=1}^{n}E\left(\xi_i^2 I_{[|\xi_i|^2\ge\delta n]}\middle|\mathcal{G}_{i-1}\right) = o_p(n) \quad \forall\,\delta > 0. \tag{2.21}$$

Denote $U_n = \sum_{i=1}^{n}\xi_i$. Then for any $1 \le p < 2$,

$$n^{p/2-1}\sum_{i=1}^{[nt]}\frac{|U_i|^p}{i^p} \to_{\mathcal{D}[0,1]} A^{p/2}\int_0^t\frac{|B(s)|^p}{s^p}\,ds, \tag{2.22a}$$

$$n^{p/2-1}\sum_{i=1}^{[nt]}i^{-1}\sum_{j=1}^{i}\frac{|U_j|^p}{j^p} \to_{\mathcal{D}[0,1]} A^{p/2}\int_0^t\frac{1}{u}\int_0^u\frac{|B(s)|^p}{s^p}\,ds\,du. \tag{2.22b}$$

Here and in the sequel we use $[a]$ to denote the largest integer less than or equal to a.

PROOF. Denote

$$M(n,\delta,t) = n^{p/2-1}\sum_{i=1}^{[nt]}\frac{|U_i|^p}{i^p}I_{\{i\ge n\delta\}},$$

$$R(n,\delta,t) = n^{p/2-1}\sum_{i=1}^{[nt]}\frac{|U_i|^p}{i^p}I_{\{i< n\delta\}}.$$

To show (2.22a), it suffices to show, in view of Theorem 1.4.2 of Billingsley (1968) and the fact that $\int_\delta^t|B(s)/s|^p\,ds \to \int_0^t|B(s)/s|^p\,ds$ uniformly in t as $\delta\to 0$, that

$$M(n,\delta,t) \to_{\mathscr{D}[0,1]} A^{p/2}\int_\delta^t\frac{|B(s)|^p}{s^p}\,ds \quad \forall\,\delta>0, \tag{2.23}$$

$$\sup_n P\Big(\sup_{0\le t\le 1}R(n,\delta,t)\ge\varepsilon\Big)\to 0 \quad \text{as } \delta\to 0. \tag{2.24}$$

Here and in the sequel, we use the convention $\int_a^b = 0$ if $b<a$. In view of (2.20) and (2.21), (2.23) follows from a straightforward application of a weak invariance principle for martingales [cf. Sen (1981), Theorem 2.4.4].

To show (2.24), let $C>0$. Define stopping time

$$T^c = \inf\{n-1\colon E(\xi_n^2|\mathscr{G}_{n-1})>C\},$$

and define

$$U_n^c = \sum_1^n \xi_i I_{[T^c\ge i]} = \sum_1^n \xi_i^c \quad\text{and}\quad R^c(n,\delta,t) = n^{p/2-1}\sum_1^{[nt]}|U_i^c/i|^p I_{\{i<n\delta\}}.$$

Since $p<2$, by Jensen's inequality,

$$E|U_i^c|^p \le \big(E|U_i^c|^2\big)^{p/2} = \left(E\sum_{j=1}^i E\big(\xi_j^{c2}|\mathscr{G}_{j-1}\big)\right)^{p/2} \le C^{p/2}i^{p/2}.$$

Thus $ER^c(n,\delta,1)\le C^{p/2}n^{p/2-1}\sum_{i=1}^{[n\delta]}(1/i^p)i^{p/2}\to 0$ as $\delta\to 0$. Since $E\sup_{t\le 1}R^c(n,\delta,t)\le ER^c(n,\delta,1)$ and $P(T^c=\infty)\to 1$ as $C\to\infty$, (2.24) follows.

Finally, (2.22b) follows from (2.22a) and a tightness argument similar to the proof of (2.24). □

LEMMA 2. *Let ξ_n, $\mathscr{G}_n$ be the same as those of Lemma 1 satisfying* (2.20) *and* (2.21). *Let $\tau_n\in\mathscr{G}_{n-1}$ with $\tau_n-\tau_{n-1}=o(\tau_n/n)$* a.s. *Define $S_n=\sum_{i=1}^n\tau_i\xi_i$. Then*

$$\frac{\tau_{[nt]}^{-1}}{\sqrt{n}}S_{[nt]} \to_{\mathscr{D}[0,1]} \sqrt{A}\,B(t). \tag{2.25}$$

Moreover, if for all $\rho > 0$ *and* $\delta > 0$,

$$\sum_{n=3}^{\infty} P\left(|\xi_n| > \delta\sqrt{n \log\log n} \,\middle|\, \mathcal{G}_{n-1}\right) < \infty \quad a.s., \tag{2.26a}$$

$$\sum_{n=3}^{\infty} \frac{E\left(|\xi_n|^2 I_{\{\rho n(\log\log n)^{-1} < |\xi_n|^2 < \delta n \log\log n\}} \,\middle|\, \mathcal{G}_{n-1}\right)}{n \log\log n} < \infty \quad a.s., \tag{2.26b}$$

then

$$\limsup_{n\to\infty} \left| \frac{\tau_n^{-1}}{\sqrt{2n \log\log n}} S_n \right| = \sqrt{A} \quad a.s. \tag{2.27}$$

PROOF. Let $U_n = \sum_{i=1}^{n} \xi_i$. The Abel transformation gives

$$S_n = \tau_n U_n - \sum_{i=2}^{n} U_{i-1}(\tau_i - \tau_{i-1}).$$

By the weak invariance principle [Sen (1981)] and the law of the iterated logarithm [Philipp and Stout (1986)] for martingales, (2.25) and (2.27) are valid when S_n is replaced by $\tau_n U_n$. Thus it suffices to show

$$\frac{\tau_n^{-1}}{\sqrt{n}} \sum_{i=2}^{n} U_{i-1}(\tau_i - \tau_{i-1}) \to_P 0, \tag{2.28}$$

$$\frac{\tau_n^{-1}}{\sqrt{2n \log\log n}} \sum_{i=2}^{n} U_{i-1}(\tau_i - \tau_{i-1}) \to 0 \quad \text{a.s.} \tag{2.29}$$

By Lemma 1 and the assumption that τ_n is slowly varying [cf. Lai and Robbins (1978)],

$$\sum_{i=3}^{n} |U_{i-1}(\tau_i - \tau_{i-1})| = o\left(\sum_{i=3}^{n} \frac{\sqrt{\log\log i}}{\sqrt{i}} \tau_i \right) = o\left(\sqrt{\log\log n}\, \tau_n \sqrt{n}\right) \quad \text{a.s.}$$

Again using the fact that τ_n (and therefore τ_n^p, too) is slowly varying, we can similarly show (2.28). □

PROOF OF THEOREM 2. We first note that, from the assumptions, $W_p(x)$ is differentiable at θ_p with

$$W_p'(\theta_p) = \begin{cases} 2M'(\theta_p)F'(m_p), & \text{if } p = 1, \\ (p-1)M'(\theta_p)\int_{-\infty}^{\infty} |t - m_p|^{p-2}\, dF(t), & \text{if } p > 1. \end{cases}$$

Since, by Theorem 1, $x_n \to \theta_p$ a.s., it follows from Theorem 3 of Lai and Robbins (1979) that

$$x_{n+1} - \theta_p = n^{-1}\tau_n^{-1}\left(\sum_{i=1}^{n} \tau_i \frac{\eta_{p,i}}{d_i} + \rho_0 \right), \tag{2.30}$$

where ρ_0 is some random variable, $\tau_n, d_n \in \mathscr{F}_{n-1}$ with $\tau_n - \tau_{n-1} = o(\tau_n/n)$ a.s. and $d_n \to W_p'(\theta_p)$ a.s. Moreover, $x_n \to \theta_p$ a.s. also implies that

$$E\left(\eta_{p,n}^2 \middle| \mathscr{F}_{n-1}\right) \to \int_{-\infty}^{\infty} |t - m_p|^{2(p-1)}\, dF(t) \quad \text{a.s.} \tag{2.31}$$

Let $\xi_n = \eta_n/d_n$, $S_n = n\tau_n(x_{n+1} - \theta_p)$. It follows from (2.7c) and the standard arguments that, for all $\rho > 0$ and $\delta > 0$,

$$\begin{gathered}\sum_{i=1}^{n} E\left(|\varepsilon_i|^{2(p-1)} I_{\{|\varepsilon_i|^{2(p-1)} \geq \delta n\}}\right) = o(n), \\ \sum_{n=3}^{\infty} P\left\{|\varepsilon_n|^{(p-1)} > \delta\sqrt{n \log\log n}\right\} < \infty \quad \text{a.s.} \\ \sum_{n=3}^{\infty} \frac{E\left(|\varepsilon_n|^{2(p-1)} I_{\{\rho n(\log\log n)^{-1} < |\varepsilon_n|^{2(p-1)} < \delta n \log\log n\}}\right)}{n \log\log n} < \infty \quad \text{a.s.}\end{gathered} \tag{2.32}$$

Since $x_n \to \theta_p$, (2.32) implies (2.21), (2.26a) and (2.26b). Hence, by Lemma 2 and (2.30), (2.12) and (2.13) hold.

Now for $1 \leq p < 2$, again by (2.30),

$$\begin{aligned} & n^{(p/2)-1} \sum_{i=1}^{n} |x_i - \theta_p|^p \\ & = n^{(p/2)-1} \sum_{i=1}^{n-1} \left| i^{-1}\tau_i^{-1} \sum_{j=1}^{i} \tau_j \frac{\eta_{p,j}}{d_j} \right|^p + o(1) \quad \text{a.s.} \end{aligned} \tag{2.33}$$

Applying the summation by parts and the fact that $\tau_j - \tau_{j-1} = o(\tau_j/j)$ a.s.,

$$\begin{aligned} & n^{(p/2)-1} \sum_{i=1}^{n} |x_i - \theta_p|^p \\ & = n^{(p/2)-1} \sum_{i=1}^{n-1} \left| \frac{U_i}{i} + o(1) i^{-1} |\tau_i^{-1}| \sum_{j=1}^{i} \left| \frac{U_j}{j} \tau_j \right| \right|^p + o(1) \quad \text{a.s.} \end{aligned} \tag{2.34}$$

For $1 < p < 2$, let $q = p/(p-1)$. Then by Hölder's inequality and Lemma 1,

$$\begin{aligned} n^{(p/2)-1} \sum_{i=1}^{n} i^{-p} |\tau_i|^{-p} \left| \sum_{j=1}^{i} \frac{U_j}{j} \tau_j \right|^p &\leq n^{(p/2)-1} \sum_{i=1}^{n} i^{-p} |\tau_i|^{-p} \sum_{j=1}^{i} \left| \frac{U_j}{j} \right|^p \left(\sum_{j=1}^{i} |\tau_j|^q \right)^{p/q} \\ &= O_p(1) n^{(p/2)-1} \sum_{i=1}^{n} i^{-1} \sum_{j=1}^{i} \left| \frac{U_j}{j} \right|^p \\ &= O_p(1). \end{aligned} \tag{2.35}$$

For $p = 1$, a similar argument using the Cauchy–Schwarz inequality results in (2.35), noting that $\eta_{p,i}$ are uniformly bounded. In view of Lemma 2, (2.14) follows from (2.8), (2.32), (2.34) and (2.35). (2.15) follows from (2.14) by taking

the Taylor expansion. (2.16) and (2.17) for $p = 2$ are due to Lai and Robbins (1979). Finally, (2.18) and (2.19) are direct consequences of (2.13). □

3. Asymptotic efficiencies. In the previous section we introduced stochastic approximation algorithms for estimating θ_p and established certain asymptotic properties of the procedures. Since we have been using transformations of y_n, it is natural to ask whether there exist other schemes that may yield better rates of convergence. Since we do not wish to make any parametric assumption on F, we shall study the efficiency in the spirit of parametric–nonparametric (semiparametric) formulation and show that (2.6) with a_n satisfying (2.9) or (2.10) are asymptotically efficient.

Our idea is to reformulate the regression model (1.1) so that the information bounds can be derived. Specifically, consider the model

$$y_n = M(x_n) + \varepsilon_n, \tag{3.1}$$

which is the same as (1.1) except we only assume ε_n to be i.i.d. with a common distribution F which, besides certain regularity conditions that will be imposed later, is arbitrary. This means that we do not require F to have zero mean or zero median. However, the family of the original stochastic approximation model (1.1) is not enlarged since $M(\cdot)$ is arbitrary and can absorb the location shift of ε_n. Moreover, the problem is to estimate $\theta_p = M^{-1}(m_p(F))$.

Our next step is to consider a smaller family by assuming M to be known. By deriving the lower information bound for the problem of estimating θ_p within the smaller class, we will see from (2.12) that the x_n defined by (2.6) are asymptotically efficient.

Listed below are some conditions which will be used later.

(3.2a) $M^{-1}(\cdot)$ exists in a neighborhood of $m_p(F)$ and $M(\cdot)$ is continuously differentiable in a neighborhood of $\theta_p = M^{-1}(m_p)$,

(3.2b) ε_n has a density function f and $\int |t|^{2(p-1)} f(t)\,dt < \infty$,

(3.2c) ε_n has a density function f and $f(m_p) > 0$.

Let $\{x_n\}$ be any sequence of adaptive designs, in the sense that $x_n \in \sigma\{y_k,\ k \le n-1\}$. Then the likelihood function becomes

$$L_n = \prod_{i=1}^{n} f(y_i - M(x_i)) = \prod_{i=1}^{n} f(\varepsilon_i).$$

Clearly, $\{\varepsilon_i,\ i = 1, \ldots, n\}$ is a sufficient statistic since we have assumed $M(\cdot)$ to be known. Thus, we may confine ourselves to the class of statistics involving only $\{\varepsilon_i\}$.

Let $p \ge 1$ be fixed and let f be the true density satisfying (3.2b) or (3.2c) depending on $p > 1$ or $p = 1$. Following Ibragimov and Has'minskii [(1981), Chapter IV], we consider one-dimensional parametric subfamilies $\{f_\theta\}$ passing through f, i.e., $f_{\theta_p} = f$, which are locally asymptotically normal (LAN) at

$\theta = \theta_p$. We call an estimator T_n of θ_p regular at f if, for any one-dimensional LAN subfamily passing through f, T_n is a regular estimator. The concept of regular estimator can be found in Ibragimov and Has'minskii [(1981), II.9].

THEOREM 3. *Let $p \geq 1$ be fixed. Assume conditions* (3.2a) *and* (3.2b) *or* (3.2c), *depending on whether $p > 1$ or $p = 1$ holds. Then, for any regular estimator T_n of θ_p, there exists a distribution function G such that*

$$\sqrt{n}\,(T_n - \theta_p) \to_{\mathscr{D}} N\big(0, I_p(\theta_p)^{-1}\big) * G, \tag{3.3}$$

where $*$ denotes the convolution operator and where

$$I_p(\theta_p) = \begin{cases} \big[2M'(\theta_p) f(m_p)\big]^2, & \text{if } p = 1, \\ \dfrac{\big[(p-1)M'(\theta_p)\int_{-\infty}^{\infty} |t - m_p|^{p-2}\, dF(t)\big]^2}{\int_{-\infty}^{\infty} |t - m_p|^{2(p-1)}\, dF(t)}, & \text{if } p > 1. \end{cases} \tag{3.4}$$

PROOF. First consider the case $p > 1$. Let

$$h(t) = \begin{cases} -|t - M(\theta_p)|^{p-1}, & \text{if } t < M(\theta_p), \\ 0, & \text{if } t = 0, \\ +|t - M(\theta_p)|^{p-1}, & \text{if } t > M(\theta_p). \end{cases} \tag{3.5}$$

In analogy with the approach of Ibragimov and Has'minskii (1981), define one-dimensional parametric families

$$f_\lambda(t) = f(t)\big[1 + \big(\lambda - M(\theta_p)\big)h_N(t)\big], \tag{3.6}$$

where

$$h_N(t) = \begin{cases} h(t), & \text{if } -k(N) \leq t \leq N, \\ 0, & \text{otherwise,} \end{cases}$$

and $k(N)$ are determined by $\int_{-k(N)}^{N} h(t) f(t)\, dt = 0$. Thus, (3.6) defines a parametric family for each N with λ in a close neighborhood of $M(\theta_p)$. It is certainly LAN at $\lambda = M(\theta_p)$ from II.2 of Ibragimov and Has'minskii (1981), with its Fisher information

$$I_{p,N}^{\lambda} = \int_{-\infty}^{\infty} h_N^2(t)\, f(t)\, dt. \tag{3.7}$$

Now let $\theta(\lambda) = M^{-1}(m_p(f_\lambda))$, the parameter which we want to estimate. Since $m_p(f_\lambda)$ is the unique solution of the equation

$$-\int_{-\infty}^{m} |t - m|^{p-1} f_\lambda(t)\, dt + \int_{m}^{\infty} |t - m|^{p-1} f_\lambda(t)\, dt = 0$$

and since conditions (3.2a) and (3.2b) are satisfied, we have

$$\frac{dm_p(f_\lambda)}{d\lambda}\bigg|_{\lambda=m_p} = \frac{(p-1)\int_{-\infty}^{\infty}|t-m_p|^{p-2}f(t)\,dt}{\int_{-\infty}^{\infty}h(t)h_N(t)f(t)\,dt}. \tag{3.8}$$

Since $\theta(\lambda) = M^{-1}(m_p)$, combining (3.7) and (3.8) yields the Fisher information quantity with respect to θ at θ_p,

$$I_{p,N}(\theta_p) = \left[\frac{M'(\theta_p)(p-1)\int_{-\infty}^{\infty}|t-m_p|^{p-2}f(t)\,dt}{\int_{-\infty}^{\infty}h(t)h_N(t)f(t)\,dt}\right]^2 \int_{-\infty}^{\infty} h_N^2 f(t)\,dt.$$

We know that T_n is a regular estimator for this LAN family. Thus by Hájek–Le Cam's convolution theorem [cf. Ibragimov and Has'minskii (1981), II.9] we get, for some distribution function G_N, $\sqrt{n}(T_n - \theta_p) \to_{\mathscr{D}} N(0, [I_{p,N}(\theta_p)]^{-1} * G_N$. Now letting $N \to \infty$, we have $I_{p,N}(\theta_p) \to I_p(\theta_p)$, implying $N(0, I_{p,N}(\theta_p)^{-1}) \to N(0, [I_p(\theta_p)]^{-1})$. Thus, $G_N \to G$ for some distribution G. Hence, $\sqrt{n}(T_n - \theta_p) \to_{\mathscr{D}} N(0, [I_p(\theta_p)]^{-1}) * G$. For $p = 1$, we introduce similar parametric families (3.6) with

$$h(t) = \begin{cases} -1, & \text{if } t < M(\theta_p), \\ 0, & \text{if } t = 0, \\ +1, & \text{if } t > M(\theta_p) \end{cases}$$

and h_N, the corresponding truncations of h. The above argument can then be employed to show that (3.3) still holds. □

Acknowledgments. The author wishes to thank Professors Tze Leung Lai and Adam Martinsek and Mr. Bertrand Clarke for helpful discussions. He also thanks an Associate Editor and referees for many valuable comments and suggestions leading toward numerous improvements.

REFERENCES

ANBAR, D. (1973). On the optimal estimation methods using stochastic approximation procedures. *Ann. Statist.* **1** 1175–1184.

ANDERSON, T. W. and TAYLOR, J. (1976). Some experimental results in the statistical properties of least squares estimates in a control problem. *Econometrica* **44** 1289–1302.

BEGUN, J. M., HALL, W. J., HUANG, W.-M. and WELLNER, J. A. (1983). Information and asymptotic efficiency in parametric–nonparametric models. *Ann. Statist.* **11** 432–452.

BERGER, E. (1986). Asymptotic behaviour of a class of stochastic approximation procedures. *Probab. Theory Related Fields* **71** 517–552.

BILLINGSLEY, P. (1968). *Weak Convergence of Probability Measures*. Wiley, New York.

BLUM, J. R. (1954). Approximation methods which converge with probability one. *Ann. Math. Statist.* **25** 382–386.

CHUNG, K. L. (1954). On a stochastic approximation method. *Ann. Math. Statist.* **25** 463–483.

FABIAN, V. (1973). Asymptotically efficient stochastic approximation: The RM case. *Ann. Statist.* **1** 486–495.

FABIAN, V. (1983). A local asymptotic minimax optimality of an adaptive Robbins Monro stochastic approximation procedure. In *Mathematical Learning Models—Theory and Algorithms*.

Lecture Notes in Statist. **20** (Herkenrath, Kalin and Vogel, eds.) 43–49. Springer, New York.

GOODWIN, G. C., RAMADGE, P. J. and CAINES, P. E. (1981). Discrete time stochastic adaptive control. *SIAM J. Control Optim.* **19** 829–853.

IBRAGIMOV, I. A. and HAS'MINSKII, R. Z. (1981). *Statistical Estimation: Asymptotic Theory.* Springer, New York.

KIEFER, J. and WOLFOWITZ, J. (1952). Stochastic estimation of the maximum of a regression function. *Ann. Math. Statist.* **23** 462–466.

LAI, T. L. and ROBBINS, H. (1978). Limit theorems for weighted sums and stochastic approximation processes. *Proc. Nat. Acad. Sci. U.S.A.* **75** 1068–1070.

LAI, T. L. and ROBBINS, H. (1979). Adaptive design and stochastic approximation. *Ann. Statist.* **7** 1196–1221.

LAI, T. L. and ROBBINS, H. (1981). Consistency and asymptotic efficiency of slope estimates in stochastic approximation schemes. *Z. Wahrsch. Verw. Gebiete* **56** 329–360.

LAI, T. L. and ROBBINS, H. (1982). Iterated least squares in multiperiod control. *Adv. in Appl. Math.* **3** 50–73.

LAI, T. L. and WEI, C. Z. (1987). Asymptotically efficient self-tuning regulators. *SIAM J. Control Optim.* **25** 466–481.

MARTINSEK, A. T. (1988). Comparison of slope estimates in adaptive stochastic approximation. Technical Report, Univ. Illinois.

PHILIPP, W. and STOUT, W. F. (1986). Invariance principles for martingales and sums of independent random variables. *Math. Z.* **192** 253–264.

ROBBINS, H. and MONRO, S. (1951). A stochastic approximation method. *Ann. Math. Statist.* **22** 400–407.

ROBBINS, H. and SIEGMUND, D. (1971). A convergence theorem for non-negative almost supermartingales and some applications. In *Optimizing Methods in Statistics* (J. Rustagi, ed.) 233–257. Academic, New York.

RUPPERT, D. (1981). Stochastic approximation of an implicitly defined function. *Ann. Statist.* **9** 555–566.

SACKS, J. (1958). Asymptotic distribution of stochastic approximation method. *Ann. Math. Statist.* **29** 373–405.

SEN, P. K. (1981). *Sequential Nonparametrics: Invariance Principles and Statistical Inference.* Wiley, New York.

STEIN, C. (1956). Efficient nonparametric testing and estimation. *Proc. Third Berkeley Symp. Math. Statist. Prob.* **1** 292–301.

WALK, H. (1977). An invariance principle for the Robbins–Monro process in a Hilbert space. *Z. Wahrsch. Verw. Gebiete* **39** 135–150.

WALK, H. (1988). Limit behavior of stochastic approximation processes. *Statist. Decisions* **6** 109–128.

WEI, C. Z. (1987). Multivariate adaptive stochastic approximation. *Ann. Statist.* **15** 1115–1130.

ZELLNER, A. (1971). *An Introduction to Bayesian Inference in Econometrics.* Wiley, New York.

DEPARTMENT OF STATISTICS
101 ILLINI HALL
UNIVERSITY OF ILLINOIS
CHAMPAIGN, ILLINOIS 61820