The Annals of Statistics
1981, Vol. 9, No. 5, 1020–1025

A STOCHASTIC APPROXIMATION BY OBSERVATIONS ON A DISCRETE LATTICE USING ISOTONIC REGRESSION[1]

BY H. G. MUKERJEE

The University of Iowa

A new non-parametric stochastic approximation procedure for estimating the roots of a non-decreasing regression function is described. The observations are taken on a discrete lattice and the estimation is based on the roots of the sample isotonic regression function fitted to the observed values. Asymptotic properties of the estimator are proven. When specialized to the bio-assay case it gives asymptotic results similar to those obtained by Derman for his up-and-down method but under weaker assumptions than Derman required.

1. Introduction. Let R denote the real line, N the positive integers, and I the set of all integers. For each $x \in R$ let $H(\cdot \mid x)$ be a distribution function and let $m(x) = \int y \, dH(y \mid x)$ define the corresponding regression function. m is assumed to be unknown; however, the experimenter is allowed to take unbiased observations from $H(\cdot \mid x)$ for any x. Suppose θ is a root of the equation $m(x) = \alpha$. The object is to estimate θ.

When θ is unique, Robbins and Monro (1951) suggest taking x_1 as an arbitrary initial estimate of θ and generating future estimates by $x_{n+1} = x_n - a_n(y_n - \alpha)$ for $n > 1$, where the conditional distribution of y_n given the past is $H(\cdot \mid x_n)$ and $\{a_n\}$ is a fixed positive sequence. Typically $a_n = a/n$ for some $a > 0$. This procedure and its many variations have been studied extensively. The convergence of x_n to θ in different modes has been studied as have the asymptotic normality of x_n and some optimal properties. The book by Wasan [6] contains an extensive bibliography. From a practical viewpoint there are two difficulties with this procedure. It may be that the stimulus (the variable x) can be changed only by integral multiples of some unit. Secondly, "sample preparation" at "odd" values of x may be difficult, impossible, or expensive. For these cases it may be desirable or necessary to take observations at points of some lattice $L = L(d_0, h) = \{d_i = d_0 + ih : i \in I\}$ for some $d_0 \in R$ and $h > 0$.

For many experimental situations it is reasonable to assume that the regression function m is non-decreasing. Under this assumption we fit an isotonic regression function [1, 2] to the observed values taken on some lattice L and base our estimate on the solution of $\hat{m}(x) = \alpha$, where $\hat{m}$ is the sample isotonic regression function. A similar procedure can be used with antitonic regression functions when m is assumed to be non-increasing. In the Robbins-Monro procedure the estimates $\{x_n\}$ of θ are Markovian in nature. After n steps the entire influence of the past is contained in the estimate $x_{n+1} = x_n - a_n(y_n - \alpha)$. This makes the estimate vulnerable to one or more "bad" observations (x_n going the wrong direction) near the end. Using isotonic regression it is reasonable to expect that the "weights" of positive observations on the right and negative observations on the left will "soften the blow" of occasional "bad" observations.

In the bio-assay problem of response-no response to various dosages of a treatment, the regression function is a distribution function, $m(x)$ being the probability of response (indicated by the value 1) and $1 - m(x)$ being the probability of no response (valued 0) at

[1] This research is based on portions of the author's Ph.D. dissertation (1977) at the State University of New York—Binghamton.

Received November, 1977. Revised December, 1980.
AMS 1970 *subject classification.* 62L20.
Key words and phrases. Stochastic approximation, isotonic regression.

dosage x, $x \in R$. Dixon and Mood [4] approximate θ in this case using the so-called up-and-down method by observations on a lattice. The results proven for this procedure and its many variations [7] depend on the parametric assumption that the regression function (or one obtained by a suitable transformation of the variate) is a normal distribution function. Intuitively, observations far away from θ contribute to the knowledge of θ via this parametric assumption. Thus the results are globally dependent on this model assumption which may not be very good. Derman [3] describes a non-parametric up-and-down method and derives a result concerning the asymptotic properties of his estimate. Our method specialized to the bio-assay case gives similar results under weaker assumptions. It is also argued that in a sense this procedure is asymptotically more efficient.

2. Model and procedure. For notational convenience we assume $\alpha = 0$ from now on.

(1) We assume that m is a real valued function, that γ and δ are real numbers with $\gamma \le \delta$, that $\sup_{x \le t} m(x) < 0$ for each $t < \gamma$ and $\inf_{x \ge t} m(x) > 0$ for each $t > \delta$, that $m(\gamma) \le 0$ and $m(\delta) \ge 0$, and that $m(x) = 0$ if $\gamma < x < \delta$; γ and δ may be equal.

Let $L = L(d_0, h)$ (as defined in Section 1) be an arbitrary lattice of observation points for some $d_0 \in R$ and $h > 0$. We wish to estimate $m^{-1}\{0\}$ by a point using a stochastic approximation procedure, the estimate after n steps being denoted by θ_n. We initially take "observations" $y_1, \cdots, y_{k_1}$, at "observation points" $x_1, \cdots, x_{k_1}$ (respectively), fixed or random, in L. At the n-th stage ($n > 1$) we take either one or two observations in a manner to be indicated later. It is assumed that the conditional distribution of an observation y given any value of the corresponding observation point x is given by $H(\cdot \mid x)$ and is otherwise independent of the past and any other observation in the present.

After n stages (through time n) let $(x_1, y_1), \cdots, (x_{k_n}, y_{k_n})$ be the ordered pairs of observation points x_i and observations y_i where k_n is the total number of observations taken through time n written in increasing order in time; the ordering among the observations taken at any given time (stage) having more than one observation will be given later.

Let $I_A(\cdot)$ be the indicator function of the set A.

For $r \le s$, both in L, let

$$n(r, s) = \textstyle\sum_{i=1}^{k_n} I_{[r,s]}(x_i),$$

let

$$A_n(r, s) = \textstyle\sum_{i=1}^{k_n} y_i I_{[r,s]}(x_i)/n(r, s) \qquad \text{if} \quad n(r, s) \ne 0,$$

and let $A_n(r, s) = 0$ if $n(r, s) = 0$. We define the estimate of m restricted to $\{x_1, x_2, \cdots, x_{k_n}\}$ to be the sample isotonic regression function m_n defined by

$$m_n(x) = \max_{r \le x} \min_{s \ge x} A_n(r, s), \qquad x \in \{x_1, x_2, \cdots, x_{k_n}\}.$$

It is the least squares fit of the observed values subject to the constraint that the fitted values define a non-decreasing function on $\{x_1, x_2, \cdots, x_{k_n}\}$ in the order of the reals. See the book by Barlow et al. [2] for an extensive discussion on the theory and applications of isotonic and antitonic regression functions. Although isotonic regression functions are usually used to estimate isotonic functions we do not need to assume that $m(\cdot)$ is monotone to get our results. Let x_{nm} and x_{nM} be the smallest and the largest values, respectively, of the observation points $x_1, \cdots, x_{k_n}$ through time n. Retaining the same symbol, we extend m_n to a function on R by connecting adjacent points on the graph of m_n by straight line segments, and by defining $m_n(x) = m_n(x_{nM})$ for $x > x_{nM}$ and $m_n(x) = m_n(x_{nm})$ for $x < x_{nm}$. The function m_n thus defined is a continuous polygonal non-decreasing function on R.

From the definition of m_n we note that only the following cases can occur:

a) $m_n(x) > 0$ for all x;

b) $m_n(x) < 0$ for all x;

c) $m_n^{-1}\{0\} \cap [x_{nm}, x_{nM}]$ is a single point or a non-empty finite closed interval: call it $[a, b]$.

Then in the corresponding cases above, θ_n is defined by:

a) $\theta_n = x_{nm} - h$;

b) $\theta_n = x_{nM} + h$;

c) $\theta_n = \frac{1}{2}(a + b)$.

The observation point(s) for $(n + 1)$st stage, given the preceding stages, are defined as follows:

i) if $\theta_n \in L$ we set $k_{n+1} = k_n + 1$ and $x_{k_n+1} = \theta_n$;

ii) if $\theta_n \notin L$ we set $k_{n+1} = k_n + 2$, $x_{k_n+1} = \max\{d \in L : d < \theta_n\}$, and $x_{k_n+2} = \min\{d \in L : \theta_n < d\}$.

3. Asymptotic results. We assume all events and random variables are defined on some appropriate probability space (Ω, Σ, P).

Notation. Equality (inequality) between random variables implies a.s. equality (inequality) only. All convergences are a.s. convergences. In definitions and other statements phrases such as "for each (or all) $n \in N$" and "as $n \to \infty$" will be frequently omitted when these implications are obvious. Expressions containing an empty sum as a multiplicand will be assumed to be zero.

Let $\sigma_0 = \sigma(x_1, \cdots, x_{k_1})$ and let $\sigma_k = \sigma(x_1, \cdots, x_{k+1}; y_1, \cdots, y_k)$, where $\sigma(\cdot)$ indicates the σ-fields generated by the random variables within the parentheses. Note that even though the ordering of the x_i's is arbitrary when multiple observations are taken at some stage, the conditional distribution of y_k given σ_{k-1} is still given by $H(\cdot \mid x_k)$ because of our assumption of independence of observations in the present conditioned on the past.

Let $z_k = y_k - m(x_k)$ and let $P_k(\cdot)$ and $E_k(\cdot)$ denote conditional probability and expectation, respectively, given σ_k. Note that

$$E_{k-1}(z_k) = 0. \tag{2}$$

If X is a random variable having distribution function $H(\cdot \mid x)$, let $F_x(t) = P\{|X - m(x)| \geq t\}$. Define $F(t) = \sup_x F_x(t)$. This implies $P_{k-1}(|z_k| \geq t) \leq F(t)$ for all $t \geq 0$. We assume

$$F(t) \to 0 \quad \text{as} \quad t \to \infty \qquad \text{and} \qquad \int_0^\infty t\,|dF(t)| < \infty. \tag{3}$$

Note that (3) implies

$$\int_0^\infty F(t)\,dt < \infty. \tag{3$'$}$$

Let $m = \max\{i \in I : m(d_0 + ih) < 0\}$ so that $d_m = \max\{d \in L : m(d) < 0\}$. Similarly let $M = \min\{i \in I : m(d_0 + ih) > 0\}$ so that $d_M = \min\{d \in L : m(d) > 0\}$. Note that d_m and d_M are well defined from our assumptions about m.

If A_n is a sequence of events (sets), the notation "A_n eventually" will mean that A_n happens for all n sufficiently large, i.e., "A_n eventually" is $\liminf A_n$.

THEOREM. *Under assumptions* (1) *and* (3)

$$P[d_m \leq x_k \leq d_M \text{ eventually}] = 1 \quad \text{and} \tag{A}$$

$$P[d_m \leq \theta_n \leq d_M \text{ eventually}] = 1 \tag{B}$$

for the procedure described in Section 2.

We first give two lemmas, the second of which contains the only result we need in the proof of the theorem from our distributional assumption (3).

Let B be any Borel subset of R. Let $n_B = \sum_{i=1}^n I_B(x_i)$ and let $\infty_B = \lim_{n\to\infty} n_B$, finite or infinite. Note that n_B and ∞_B are random variables (possibly extended). Note also that

$I_B(x_i)$ and n_B are both σ_{n-1}-measurable, facts we will use frequently without explicit reference.

(*) A sequence $\{z_k\}$, adapted to an increasing sequence $\sum_k$ of sigma-algebras, will be said to satisfy (*) if there exists a non-increasing function F on $[0, \infty)$ having $F(0) = 1$, satisfying $P[|z_{k+1}| \geq t \mid \sum_k](\omega) \leq F(t)$ a.s. for each k, and satisfying (3).

LEMMA 1. *If $\{z_k^*\}$ is adapted to $\sum_k$, satisfies (*) for F, and satisfies $E[z_{k+1}^* \mid \sum_k](\omega) = 0$ for each k, then*

$$\frac{1}{k} \sum_{i=1}^{k} z_i^* \rightarrow 0 \qquad \text{a.s.}$$

PROOF. When $\{z_k^*\}$ an independent sequence, the result is well known. The proof in this case is similar to the proof in the independent case.

LEMMA 2.

$$\frac{1}{n_B} \sum_{i=1}^{n} z_i I_B(x_i) \rightarrow 0 \qquad \text{a.s. } on\ [n_B \rightarrow \infty].$$

PROOF. If $\infty_B(\omega) \geq k$ let $z_k^*(\omega) = z_i(\omega)$ where i is the kth positive integer for which $x_i(\omega) \in B$. If $\infty_B(\omega) < k$ let $z_k^*(\omega) = 0$. The sequence $\{z_k^*\}$ adapted to the appropriate sequence of sigma-algebras satisfies (for $k \geq k_1$) the conditions of Lemma 1. The conclusion of Lemma 2 follows immediately from that of Lemma 1.

PROOF OF THE THEOREM. Let $\epsilon > 0$ be arbitrary. From the way m_n was defined, $m_n(\infty) = m_n(x_{nM}) = \max_{r \leq x_{nM}} A_n(r, x_{nM})$. Thus $m_n(\infty) \geq A_n(d_M, x_{nM}) = A_n(d_M, \infty)$ if there is any observation in $[d_M, \infty)$ after n stages. We note that in case (b) we have $x_{n+1,M} = x_{k_{n+1}} = x_{nM} + h$ and that $x_{n+1,M} = x_{nM}$ otherwise. Thus $m_n(\infty) < 0$ i.o. if and only if $x_{nM} \rightarrow \infty$, which in turn implies $n(d_M, \infty) \rightarrow \infty$.

From Lemma 2, on $[n(d_M, \infty) \rightarrow \infty]$,

$$\frac{1}{n(d_M, \infty)} \sum_{i=1}^{k_n} z_i I_{[d_M, \infty)}(x_i) = A_n(d_M, \infty) - \frac{1}{n(d_M, \infty)} \sum_{i=1}^{k_n} m(x_i) I_{[d_M, \infty)}(x_i) \rightarrow 0 \qquad \text{a.s.}$$

and

$$\frac{1}{n(d_M, \infty)} \sum_{i=1}^{k_n} m(x_i) I_{[d_M, \infty)}(x_i) \geq m(d_M) > 0$$

for all n large enough so that $A_n(d_M, \infty) > 0$ for all n large enough. Thus

$$\begin{aligned} P[m_n(\infty) < 0 \text{ i.o.}] &= P[m_n(\infty) < 0 \text{ i.o.}, x_{nM} \rightarrow \infty] \\ &\leq P[m_n(\infty) < 0 \text{ i.o.}, n(d_M, \infty) \rightarrow \infty] \\ &\leq P[m_n(\infty) \leq 0 \text{ i.o.}, n(d_M, \infty) \rightarrow \infty] \\ &\leq P[A_n(d_M, \infty) \leq 0 \text{ i.o.}, n(d_M, \infty) \rightarrow \infty] = 0. \end{aligned} \tag{4}$$

Similarly, one can show that $P[m_n(-\infty) > 0 \text{ i.o.}] = 0$.

Since $x_{nm} \leq x_i \leq x_{nM}$ for $k_n < i \leq k_{n+1}$ if $m_n(x_{nm}) \leq 0$ and $m_n(x_{nM}) \geq 0$, there exists $n_0 \in N$ such that if

$$F = [d_{m-n_0} \leq x_n \leq d_{M+n_0} \qquad \text{for all } n \in N]$$

then $P(F) \geq 1 - \epsilon$.

Let $F_k = [x_n = d_{M+k} \text{ i.o.}]$ for $k = 0, 1, \cdots, n_0$. We will prove that

$$P[F \cap F_k \cap \{m_n(d_{M+k}) \leq 0 \text{ i.o.}\}] = 0 \qquad \text{for} \quad k = 0, 1, \cdots, n_0. \tag{5}$$

Now $P[F \cap F_k \cap \{m_n(d_{M+k}) \leq 0 \text{ i.o.}\}]$

$$\leq P[F \cap F_k \cap \{\min_{i \geq M+k} A_n(d_{M+k}, d_i) \leq \text{i.o.}\}]$$

$$\leq \textstyle\sum_{i=M+k}^{M+n_0} P[F \cap F_k \cap \{A_n(d_{M+k}, d_i) \leq 0 \text{ i.o.}\}]$$

and, using the same argument used in proving (4), we see that the last expression is zero, proving (5).

We now show that if k' is the largest integer k such that $x_n = d_{M+k'}$ i.o. then $P[k' \geq 1] = 0$. Let

$$\tilde{F} = F - \cap_{0 \leq k \leq n_0}[F_k \cap \{m_n(d_{M+k}) \leq 0 \text{ i.o.}\}].$$

By (5), $P(\tilde{F}) = P(F)$. We will show that $\tilde{F} \cap \{k' \geq 1\}$ is empty. Suppose not. Since $x_n = d_{M+k'}$ i.o., $m_n(d_{M+k'}) \leq 0$ only finitely often on $\tilde{F}$. But according to our procedure if $m_n(d_{M+k'}) > 0$ and we take an observation at $d_{M+k'}$ at the $(n+1)$st stage, then $m_n(d_{M+k'-1}) < 0$, $x_{k_n+1} = d_{M+k'-1}$, and $x_{k_n+2} = d_{M+k'}$. Thus $x_n = d_{M+k'-1}$ i.o. and $m_n(d_{M+k'-1}) \leq 0$ i.o. on $\tilde{F} \cap [k' \geq 1]$. This is a contradiction. Hence, $P[k' \leq 0] \geq P[\tilde{F} \cap \{k' \leq 0\}] = P(\tilde{F}) = P(F) \geq 1 - \epsilon$. Since $\epsilon > 0$ was arbitrary we have shown that $P[\limsup_n x_n \leq d_M] = 1$. Similarly, one can show that $P[\liminf_n x_n \geq d_m] = 1$, thus proving (A). According to our procedure if θ_n is outside of $[d_m, d_M]$ then x_{k_n+1} or x_{k_n+2} or both are outside of $[d_m, d_M]$ and thus (A) implies (B). This completes the proof of the theorem.

4. Remarks. A) It is clear from the proof of the theorem that a variety of strategies can be employed in the first few stages of the experiment without affecting the asymptotic results. In applications it might be advisable to start with a coarser grid $L' \subset L$ and have enough observation points sufficiently spread out to cover $m^{-1}\{0\}$ with a reasonable degree of certainty. It is also possible in cases a) and b) to choose the next observation point more than one unit (of h) farther than the appropriate extreme observation point. Moreover, this could be subjectively based on the shape of m_n.

B) Suppose $m^{-1}\{0\}$ is some unique point θ, $\theta \notin L$, and m is linear between d_m and d_M. Since $d_m \leq x_n \leq d_M$ eventually with probability 1 we expect $m_n^{-1}\{0\} = \theta_n \notin L$ frequently. In this case it is possible to improve on the estimate of θ by taking approximately $|m_n(x_{k_n+1})|/[m_n(x_{k_n+2}) - m_n(x_{k_n+1})]$ proportion of the observations at x_{k_n+2} and $|m_n(x_{k_n+2})|/[m_n(x_{k_n+2}) - m_n(x_{k_n+1})]$ proportion of the observations at x_{k_n+1} $(n+1)$st at the stage whenever $m_n^{-1}\{0\} = \theta_n \notin L$. As a matter of fact any procedure for choosing one or more observations and observation points when $\theta_n \notin L$ will give the conclusion of the theorem provided only that $x_n = d_i$ i.o. and $x_n = d_{i+1}$ i.o. whenever $\theta_n \in (d_i, d_{i+1})$ i.o. For instance one could take a single observation by an appropriate randomization procedure like tossing a coin.

C) For the bio-assay case Derman [3] obtains

$$P[\limsup_n \theta_n \leq \theta + h, \liminf_n \theta_n \geq \theta - h] = 1$$

for his procedure under the assumptions that $m^{-1}\{0\} = \theta$ and that m is strictly increasing in the interval $[\theta - h, \theta + h]$. Our procedure applies to a fairly general class of regression functions and gives results similar to those obtained by Derman for the bio-assay case, but our model assumptions are weaker than Derman's.

D) In all up-and-down methods $\{x_n\}$ is an irreducible Markov chain where all states are recurrent and non-null [3]. Thus with probability 1 some fraction of the observations is taken far away from the root as $n \to \infty$. In the non-parametric case this amounts to a (possibly large) loss of efficiency not incurred in our method.

Acknowledgments. The author wishes to express his gratitude to his thesis advisor, Professor David L. Hanson, for doing a great job of advising. The author would also like to thank the referees whose helpful suggestions have substantially simplified and improved the presentation.

REFERENCES

[1] AYER, M., BRUNK, H. D., EWING, G. M., REID, W. T., and SILVERMAN, E. (1955). An empirical distribution function for sampling with incomplete information. *Ann. Math. Statist.* **26** 641–647.

[2] BARLOW, R. E., BARTHOLOMEW, D. J., BREMNER, J. M., and BRUNK, H. D. (1972). *Statistical Inference Under Order Restrictions.* Wiley, London.

[3] DERMAN, C. (1957). Non-parametric up-and-down experimentation, *Ann. Math. Statist.* **28** 795–797.

[4] DIXON, W. J. and MOOD, A. M. (1948). A method for obtaining and analyzing sensitivity data. *J. Amer. Statist. Assoc.* **43** 109–126.

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

[6] WASAN, M. T. (1969). *Stochastic Approximation.* Cambridge University Press.

[7] WETHERILL, G. B. (1963). Sequential estimation of quantal response curves, *J. Roy. Statist. Soc. Ser. B* **25** 1–48.

DEPARTMENT OF STATISTICS
UNIVERSITY OF IOWA
IOWA CITY, IOWA 52242