Z. Wahrscheinlichkeitstheorie verw. Gebiete
60, 535–554 (1982)

Zeitschrift für
Wahrscheinlichkeitstheorie
und verwandte Gebiete

On a New Stopping Rule for Stochastic Approximation

Donna F. Stroup[1] and Henry I. Braun[2]

1 University of Texas at Austin, Austin, Tx, 78712, USA
[2] Educational Testing Service, Division of Measurement, Statistics and Data Analysis Research, Princeton, N.J. 08541, USA

Summary. A new stopping rule for the Robbins-Monro process, based on an F-statistic criterion is proposed and its asymptotic behavior established. On the basis of evidence obtained through experimental sampling, the procedure seems to work well over a wide variety of situations. A two-stage procedure, coupling the new rule with an earlier one proposed by Sielken [1973] is recommended for practical use.

1. Introduction

Since the seminal paper of Robbins and Monro [1951], a large number of papers have dealt with the purely probabilistic aspects of stochastic approximation. Schmetterer [1961] and Scheber [1973] provide good reviews of the relevant literature. Comparatively little attention has been devoted to the development of stopping rules which would permit the use of the procedure in practice. Two exceptions are a theoretical treatment by Farrell [1962] and a more practical approach by Sielken [1973].

In this paper we present a new stopping rule for the Robbins-Monro process, which is based on the first passage time of an F-statistic criterion below a fixed boundary. This new stopping time has finite moments of all orders, and it is shown that in an appropriate sequence of problems, a normalized version of the stopping time converges in distribution to a fixed random variable. Because the classical theorems of Billingsley [1968] cannot be applied, a somewhat different approach, which may be of independent interest, is required to show that the randomly stopped Robbins-Monro process is asymptotically normally distributed. Asymptotically valid confidence interval estimates for the parameter of interest may therefore be obtained.

By means of experimental sampling, the small-sample properties of the procedure are shown to be well approximated by the asymptotics. Comparisons with Sielken's method are carried out and some suggestions for the practical use of the procedure are put forward.

0044-3719/82/0060/0535/$04.00

2. Background and Preliminary Results

2.1 Notation

Stochastic approximation is concerned with the problem of determining what level of input is necessary to produce a given level of response. For any real number x, the level of input, the corresponding observed response $y(x)$ is a random variable with expectation $M(x)$; in general, M is assumed to be monotone and non-decreasing. For a given value α, assume $M(x)=\alpha$ has a unique root, denoted by θ. Let x_1 be any constant and define $x_2, x_3, \ldots$ recursively by

$$x_{n+1}=x_n-a_n(y_n-\alpha). \tag{2.1}$$

where y_n is a random variable whose conditional distribution given x_n coincides with the distribution of the random variable $y(x_n)$ with variance σ^2, and $\{a_n\}$ is a sequence of positive constants which converge to 0 as $n\to\infty$. Robbins and Monro [1951] proved the convergence in probability of x_n to θ under certain conditions, Blum [1954] strengthened this to almost sure convergence and Sacks [1958] demonstrated that $\sqrt{n}(x_n-\theta)$ is asymptotically normal with mean 0 and variance $\tau^2=A^2\sigma^2/(2AM'(\theta)-1)$.

R.L. Sielken [1973] applied this result to form a stopping rule of the sort proposed by Chow and Robbins [1965]. Heretofore, this is the only generalized treatment of the stopping problem. Sielken prescribed that sampling be stopped as soon as the length of the confidence interval based on the *asymptotic* distribution of x_n is as small as desired, or equivalently as soon as the estimated standard deviation is sufficiently small. In particular, let $\gamma(0<\gamma<\frac{1}{2})$ and $d>0$ be fixed and let K_γ be such that

$$1-2\gamma=\int_{-K_\gamma}^{K_\gamma}(2\pi)^{-\frac{1}{2}}\exp(-\tfrac{1}{2}x^2)\,dx;$$

that is, let K_γ be the upper $100(1-\gamma)$-percentile of the standard normal distribution. For an approximate $100(1-2\gamma)\%$ confidence interval of the form (x_n-d, x_n+d), sampling ceases as soon $K_\gamma\tau/\sqrt{n}\leqq d$ or, equivalently, as soon as $K_\gamma^2\tau^2/d^2\leqq n$. Now, let $v=K_\gamma^2\tau^2/d^2$. The apparent drawback to implementing this procedure is that v, which determines the stopping rule, contains τ^2, which is generally unknown and must be estimated. Recall that $\tau^2=A^2\sigma^2/(2AM'(\theta)-1)$, where $\sigma^2=\mathrm{var}(y_n|x_n)$ (or, in general, $\lim_{n\to\infty}\mathrm{var}(y_n|x_n)$) and $M'(\theta)$ are the two unknown quantities.

Burkholder [1956] has proposed estimators of σ^2 and $M'(\theta)$ and obtained sufficient conditions for these estimators to converge with probability one. The estimator of $M'(\theta)$ requires that at the n-th step in the successive approximation procedure, an observation be taken on $y(x_n)$ and $y(x_n+j_n)$, where $\{j_n\}$ is a sequence of positive constants such that $j_n n^\lambda$ converges to a positive limit as $n\to\infty$, for some λ, $0<\lambda<\frac{1}{2}$. Specifically, let $\{w_n\}$ be a sequence of random variables such that the conditional distribution of w_n for a given x_n coincides with the distribution of $y(x_n+j_n)$. Form

$$t_n=n^{-1}\sum_{i=1}^{n}\{(w_i-y_i)/j_i\} \tag{2.2}$$

and

$$\hat{\sigma}_n^2 = \tfrac{1}{2}\left\{ n^{-1} \sum_{i=1}^{n} (w_i - \alpha)^2 + n^{-1} \sum_{i=1}^{n} (y_i - \alpha)^2 \right\}. \tag{2.3}$$

Burkholder's results imply that under certain assumptions, $t_n \to M'(\theta)$ and $\hat{\sigma}_n^2 \to \sigma^2$, both with probability one as $n \to \infty$.

Now let $v_n = K_\gamma^2 \tau_n^2 / d^2$, where

$$\tau_n^2 = A^2 \hat{\sigma}_n^2 / (2 A t_n - 1) \tag{2.4}$$

and let v and τ be the corresponding quantities with $\hat{\sigma}_n^2$ and t_n replaced by σ^2 and $M'(\theta)$, respectively. Let the stopping rule implied by (2.4) be denoted by

$$N_s = N_s(d, \gamma) = \inf\{n\colon v_n \leqq n\}. \tag{2.5}$$

The resulting interval is then $(x_{N_s} - d, x_{N_s} + d)$.

Sielken [1973] showed that N_s has the following asymptotic properties:

$$\lim_{d \to 0} P\{|x_{N_s} - \theta| \leqq d\} = 1 - 2\gamma \tag{2.6}$$

$$\lim_{d \to 0} (N_s / v) = 1 \quad \text{a.s.} \tag{2.7}$$

Thus the level of the sequentially determined bounded length confidence interval converges to the prescribed level as the desired length converges to zero, and N_s is, in the sense of expression (2.7), asymptotically efficient.

Empirical Monte Carlo results of Sielken [1971] and Stroup [1979] indicate that the small sample properties of N_s are reasonably close to its asymptotic ones when the regression function M is linear or piecewise-linear, the distribution function of $y(x) - M(x)$ is short-tailed (normal, uniform or chi-squared), and the variance of the observations, σ^2, is not much larger than the deviation of the initial observation $|x_1 - \theta|$.

In some cases, quantal response for example, it is desirable to accommodate non-linear response functions. In practice, the effect of the initial observation and of the distributional form of $y(x) - M(x)$ should be minimized. For these reasons, we consider more closely certain aspects of Sielken's rule.

It is clear from (2.3) that $\hat{\sigma}_n^2$ is a biased estimator of σ^2 and that the early observations, the ones responsible for determining N_s, contribute to the bias. Actually, the second observation at each input level (w_n, the response at $x_n + j_n$) is necessary only for the estimate of slope. It is used to estimate the variance only because it is available and increases the degrees of freedom in $\hat{\sigma}_n^2$. In a sense, the observations w_n distort the variance estimator. We will develop estimators s_n^2 of σ^2 and b_n of $M'(\theta)$ which improve on those of Burkholder.

In the following, take $\alpha = 0$ without loss of generality. As an alternative to Sielken's rule, we define

$$u_n(k) = \sum_{i=n-k+1}^{n} y_i^2 / k s_n^2 \qquad (n = k, k+1, \ldots;\ k = 1, 2, \ldots) \tag{2.8}$$

where s_n^2 is an estimator of the variance σ^2. As a stopping rule, take

$$N_k = \inf\{n\colon u_n(k) < c\} \tag{2.9}$$

for some positive constant c. The intuition underlying this rule is that as x_n converges to θ, the numerator of $u_n(k)$ should be numerically close to the denominator inasmuch as $M(x_n)$ near zero implies that y_n will be determined by $\varepsilon_n = y_n - M(x_n)$. To formulate the procedure, this implication is reversed so that values of $u_n(k)$ near one are taken as an indication that x_n is near θ.

2.2 Assumptions

The following conditions on $\{a_n\}$, $M(x)$, $y(x)$, and $\varepsilon(x) = y(x) - M(x)$ are summarized here for completeness.

A 1: *The sequence* $\{a_n\}$ *has the form* $\{A/n\}$ *where* A *is a constant such that* $2AM'(\theta) > 1$.

Y 1: *The distribution function of* $y(x)$, *denoted* $F(\cdot|x)$, *is such that* $F(y|\cdot)$ *is Borel-measureable for every* y.

M 1: *For each* ε *in* $(0, 1)$, $\inf_{\varepsilon < |x-\theta| < \varepsilon^{-1}} \{M(x) - \alpha\} \cdot (x - \theta) > 0$.

M 2: *For some constants* K_1 *and* K_2, $|M(x) - \alpha| \leqq K_1 + K_2 |x - \theta|$ *for all* x.

M 3: $M(x)$ *is continuously differentiable at* θ.

E 1: $y(x) - M(x) = \varepsilon(x)$ *have a common distribution with zero mean and* $\varepsilon(x)$ *is independent of the previous part of the* $\{x\}$ *process.*

E 2: $\sup_x E|\varepsilon(x)|^2 < \infty$.

E 3: *The fourth moments of the distributions of* $\varepsilon(x)$ *exist and have a common bound.*

E 4: *For all* x, $E\varepsilon^3(x)$ *exists and equals zero.*

2.3 Estimating Variance and Slope

Clearly, two independent observations of the response at level x_n (rather than a displaced second observation) will permit an unbiased estimate of the response variance. Let $y_{n,1}$ and $y_{n,2}$ be two independent and identically distributed observations of the response variable at level x_n. Thus $E(y_{n,1}|x_n) = E(y_{n,2}|x_n) = M(x_n)$. At step n, form

$$\bar{y}_n = \tfrac{1}{2}(y_{n,1} + y_{n,2});$$
$$e_n = (y_{n,1} - \bar{y}_n)^2 + (y_{n,2} - \bar{y}_n)^2;$$
$$s_n^2 = n^{-1} \sum_{i=1}^{n} e_i.$$

Since we have replicate observations at each level, we can modify the original process to be

$$x_{n+1} = x_n - a_n \bar{y}_n$$

with $\operatorname{var}(\bar{y}_n|x_n) = \frac{1}{2}\operatorname{var}(y_n|x_n)$. The following lemma is an immediate consequence of the strong law of large numbers (Breiman [1968]).

Lemma 1. *Under assumptions A 1, M 1, M 2, E 2, E 3,*

$$s_n^2 \to \sigma^2 \quad \text{(a.s.)} \quad as \quad n \to \infty.$$

As an estimator for the slope $M'(\theta)$ we consider the least squares regression estimate, since under assumption *M 3*, the regression is essentially linear in a neighbourhood of θ. The drawback to this estimator is that early observations may be far from the target value and the slope of M at these points may be very different from $M'(\theta)$. This problem is also present in Burkholder's estimate t_n, used by Sielken.

With pairs of observations $(x_i, \bar{y}_i)$, $i=1,2,\ldots,n$, a modification of the usual least squares estimate is

$$b_n=\frac{\sum_{i=1}^{n} \bar{y}_i(x_i-\bar{x}_n)}{\sum_{i=1}^{n}(x_i-\bar{x}_n)^2} \tag{2.10}$$

where $\bar{x}_n=n^{-1}\sum_{i=1}^{n} x_i$. The following result will be useful in investigating the properties of the proposed stopping rule.

Theorem 1 (Lai and Robbins). *Under assumptions A 1, Y1, M 1, M 2, M 3, E 1 and E 3,*

$$b_n \to M'(\theta) \quad \text{(a.s.)}.$$

One version of this result can be found in Lai and Robbins [1979]. Its proof as well as some generalizations are contained in an unpublished manuscript (Lai and Robbins, 1978). The critical step of the proof is to show that the condition

$$\lim_{n\to\infty} \sum_{1}^{n}(x_i-x_n)^2/\log n \to \infty$$

implies strong consistency of the least squares estimator. That condition certainly holds for the scheme defined by (2.1). However, Lai and Robbins were concerned with more general adaptive stochastic approximation schemes in which an updated estimate of $M'(\theta)$ is incorporated into the $\{a_n\}$ sequence.

3. Properties of the Prediction Stopping rule N_k

3.1 Properties for a Fixed Value for k

Using the averaged responses $\bar{y}_n$ and the variance estimator s_n^2 defined in Sect. 2, the stopping statistic becomes

$$u_n(k)=\sum_{i=n-k+1}^{n} \bar{y}_i^2/ks_n^2 \qquad (n=k,k+1,\ldots).$$

Let us henceforth write y_n for $\bar{y}_n$ and ε_n for $\bar{\varepsilon}_n=\bar{y}_n-M(x_n)$.

Considerations of the asymptotic properties (as $k\to\infty$) of the sequence of stopping times $N_k=\inf\{n: u_n(k)<c\}$ will be based on the weak convergence of a process derived from $u_n(k)$. The asymptotic normality of the randomly stopped process x_{N_k} does not follow from existing results (e.g. Billingsley [1968]), since

the sufficient condition that N_k converge in probability to a fixed random variable as $k\to\infty$ is not satisfied here.

Theorem 2. *Under assumptions A 1, Y1, M 1, M 2, M 3, E 1, E 2, $2ku_n(k)$ converges in distribution to a central chi-squared random variable with k degrees of freedom, as $n\to\infty$.*

Proof. With $u_n(k)$ defined as above,

$$s_n^2\, ku_n(k)=\sum_{i=0}^{k-1} M^2(x_{n-i})+2\sum_{i=0}^{k-1}\varepsilon_{n-i}M(x_{n-i})+\sum_{i=0}^{k-1}\varepsilon_{n-i}^2.$$

Thus,

$$D(k,n)=\left|2\,ku_n(k)-\sum_{i=0}^{k-1}(\sqrt{2}\varepsilon_{n-i}/s_n)^2\right|$$
$$\leqq 2s_n^{-2}\sum_{i=0}^{k-1}M^2(x_{n-i})+4s_n^{-2}\left|\sum_{i=0}^{k-1}\varepsilon_{n-i}M(x_{n-i})\right|.$$

Since $M(x_n)\to 0$ (a.s.) as $n\to\infty$, $M(x_{n-i})\to 0$ (a.s.) for any fixed i. Hence the first term on the $RHS\to 0$ as $n\to\infty$. An application of Kolmogorov's inequality shows that the second term must tend to 0 in probability. Consequently, $D(k,n)\xrightarrow{p}0$ as $n\to\infty$. Since $s_n^2\to\sigma^2=E(\varepsilon^2(x))$, $2ku_n(k)$ converges in distribution to the sum of k independent random variables, each having a (central) chi-squared distribution with one degree of freedom.

Theorem 3. *Under assumptions A 1, Y1, M 1, M 2, M 3, E 1, and E 2, N_k has finite moments of all orders.*

Proof. For $n=k, k+1,\ldots$, define

$$u_n^*(k)=\frac{\varepsilon_n^2+\ldots+\varepsilon_{n-k+1}^2}{k\sigma^2} \tag{3.1}$$

and for some $0<\lambda<c$, let $B_i=\{\omega: u_i^*>c-\lambda\}$ and $C_i=\{\omega: u_i-u_i^*>\lambda\}$. Let I_n be the index set of a sequence of integers between $\sqrt{n}$ and n which are k units apart; that is, $I_n=\{n, n-k, n-2k,\ldots, n-d_nk\}$, where $d_n+1\geqq(n-\sqrt{n})/k\geqq d_n$.

A well known result (Chung, exercise 5, p. 44 [1974]) states that for any positive random variable N,

$$E(N^r)<\infty \quad \text{if and only if} \quad \sum_{n=1}^{\infty} n^{r-1}P(N>n)<\infty. \tag{3.2}$$

Now,

$$\{N_k>n\}\subset\bigcap_{i\in I_n}(B_i\cup C_i)=F_n; \tag{3.3}$$

and

$$F_n\subset\left(\bigcap_{i\in I_n}B_i\right)\cup\left[\bigcup_{i\in I_n}(C_i\setminus B_i)\right]=G_n \tag{3.4}$$

where $(C_i\setminus B_i)=C_i\cap B_i^c$. From (3.3) and (3.4),

$$P\{N_k>n\}\leqq P\left\{\bigcap_{i\in I_n}B_i\right\}+\sum_{i\in I_n}P\{C_i\setminus B_i\}. \tag{3.5}$$

We see, from Definition (3.1), that u_i^* is composed of independent random variables; and, for $i \in I_n$, the events B_i are independent, since they are determined by disjoint sets of independent variables.

Also from (3.1) for any i, $2ku_i^*(k)$ is distributed as a chi-squared random variable with k degrees of freedom, denoted χ_k^2. Thus,

$$P(B_i) = P(u_i^* > c - \lambda) = 1 - F\{2k(c-\lambda)\} = \rho < 1,$$

where $F(x) = P\{\chi_k^2 \leqq x\}$. Since the events $\{B_i : i \in I_n\}$ are independent,

$$P(\bigcap_{i \in I_n} B_i) = \prod_{i \in I_n} P(B_i) \leqq \rho^{(n-\sqrt{n})/k}$$

and

$$\sum_{n=1}^{\infty} n^{r-1} \rho^{(n-\sqrt{n})/k} < \infty.$$

Thus the theorem will follow from (3.2) and (3.5) by showing

$$\sum_{n=1}^{\infty} n^{r-1} [\sum_{i \in I_n} P(C_i \setminus B_i)] < \infty. \tag{3.6}$$

Let

$$u_i' = \frac{y_i^2 + \ldots + y_{i-k+1}^2}{k\sigma^2} = \frac{s_i^2}{\sigma^2} u_i;$$

then

$$C_i \subset \{|u_i - u_i'| > \lambda/2\} \cup \{|u_i' - u_i^*| > \lambda/2\} = D_i \cup E_i.$$

We consider first $E_i \setminus B_i$:

$$\begin{aligned} |u_i' - u_i^*| &= (k\sigma^2)^{-1} \left| \sum_{j=i-k+1}^{i} (y_j^2 - \varepsilon_j^2) \right| \\ &\leqq (k\sigma^2)^{-1} \sum_{j=i-k+1}^{i} M^2(x_j) + 2(k\sigma^2)^{-1} \sum_{j=i-k+1}^{i} |\varepsilon_j M(x_j)| \\ &\leqq \sigma^{-2} \max_{i-k+1 \leqq j \leqq i} M^2(x_j) + 2(k\sigma^2)^{-1} \max_{i-k+1 \leqq j \leqq i} |M(x_j)| \sum_{j=i-k+1}^{i} |\varepsilon_j|. \end{aligned}$$

Then

$$\begin{aligned} P\{E_i \setminus B_i\} &= P\{|u_i' - u_i^*| > \lambda/2 \quad \text{and} \quad u_i^* \leqq c - \lambda\} \\ &\leqq P\{\max_{i-k+1 \leqq j \leqq i} M^2(x_j) > \lambda\sigma^2/4\} \\ &\quad + P\left\{(k\sigma^2)^{-1} \max_{i-k+1 \leqq j \leqq i} |M(x_j)| \sum_{j=i-k+1}^{i} |\varepsilon_j| > \lambda/8 \quad \text{and} \quad u_i^* \leqq c - \lambda\right\}. \end{aligned} \tag{3.7}$$

Under assumption *A 1* and *M 2* (with $\theta = 0$), one obtains, $M^2(x_j) = 0(x_j^2)$ uniformly. Then by a result of Révész [1973],

$$P\{\max_{i-k+1 \leqq j \leqq i} M^2(x_j) > \lambda\sigma^2/4\} = O(i^{-t}), \quad \text{for each fixed} \quad t > 0.$$

We now pick t so large ($t>2(2r+1)/\varepsilon$) that this probability, the first term of (3.7), is $O(i^{-p})$, where $p>2r+1$.

Now, in the second probability of (3.7), the fact that $u_i^* \leqq c-\lambda$ gives an upper bound on $(k\sigma^2)^{-1} \sum_{j=i-k+1}^{i} |\varepsilon_j|$. Thus, by the argument above, for some $\lambda'>0$,

$$P\left\{(k\sigma^2)^{-1} \max_{i-k+1\leqq j\leqq i} |M(x_j)| \sum_{j=i-k+1}^{i} |\varepsilon_j| > \lambda/8 \quad \text{and} \quad u_i^* < c-\lambda\right\}$$
$$\leqq P\{\max_{i-k+1\leqq j\leqq i} |M(x_j)| > \lambda'\} = O(i^{-p}), p>2r+1.$$

Then, from (3.7), for i sufficiently large,

$$P\{E_i \setminus B_i\} = O(i^{-p}), p>2r+1. \tag{3.8}$$

Consider now the set D_i. Note that on $(B_i \cup E_i)^c$, $u_i' < c-\lambda/2$, so that on $(B_i \cup E_i)^c$, $|u_i - u_i'| > \lambda/2$ only if $|1-\sigma^2/s_i^2| > \lambda/2u_i' > \lambda/2(c-\lambda/2)$. Hence,

$$P\{D_i \setminus (B_i \cup E_i)\} \leqq P\{|1-\sigma^2/s_i^2| > \lambda/2(c-\lambda/2)\}.$$

Now, s_i^2 is an average of independent identically distributed random variables and, as $i\to\infty$, $s_i^2 \to \sigma^2$ a.s. Thus, we can apply a large deviation result of Cramér [1938] to conclude that

$$P\{D_i \setminus (B_i \cup E_i)\} = O(i^{-\frac{1}{2}} e^{-i}) \text{ as } i\to\infty. \tag{3.9}$$

Now, $P\{C_i \setminus B_i\} \leqq P\{D_i \setminus (B_i \cup E_i)\} + P\{E_i \setminus B_i\} = O(i^{-\frac{1}{2}} e^{-i}) + O(i^{-p})$ for $p>2r+1$, by (3.8) and (3.9). Finally,

$$\sum_{n=1}^{\infty} n^{r-1} \left(\sum_{i\in I_n} P\{C_i \setminus B_i\}\right) = \sum_{n=1}^{\infty} n^{r-1} \sum_{i\in I_n} O(i^{-p})$$

is a convergent series, since $r-1-(p-1)/2 < -1$ for $p>2r+1$. This proves (3.6) and hence the theorem.

3.2 Weak Convergence of the Prediction Process

We consider now the weak convergence of a process derived from the statistic $u_n(k)$ in order to prove:

Theorem 4. *Under assumptions A 1, Y 1, M 1, M 2, M 3, E 1, E 2, E 3, N_k/k converges in distribution to a positive, non-degenerate random variable as $k\to\infty$.*

Before proving this theorem, we will consider several preliminary results. Since the errors ε_n are independent random variables with mean 0 and variance $\sigma^2/2$, $\{(2\varepsilon_n^2/\sigma^2)-1\}/\sqrt{2} = \varepsilon_n^2\sqrt{2}/\sigma^2 - 1/\sqrt{2}$ are independent, identically distributed random variables with mean zero. Without essential loss of generality, we also assume they have unit variance.

Let $\{W(t)\colon 0\leqq t\leqq 1\}$ be a Wiener Process. Define

$$T_n = \sum_{i=1}^{n} (\varepsilon_i^2\sqrt{2}/\sigma^2 - 1/\sqrt{2}),$$

and let $X_n(t)$ be the corresponding partial sum process:

$$X_n(t)=n^{-\frac{1}{2}}T_{[nt]}+n^{-\frac{1}{2}}(nt-[nt])(\varepsilon_{[nt]+1}^2\sqrt{2}/\sigma^2-1/\sqrt{2}). \tag{3.10}$$

Using these definitions, we now state and prove three lemmas, each under the assumptions of Theorem 4.

Lemma 2. *$X_n(t)$ converges weakly to $W(t)$ as $n\to\infty$.*

Proof. Since the random variables involved in the partial sum T_n are independent and identically distributed, the lemma follows directly from Donsker's Theorem.

Define

$$S_n=\sum_{i=1}^{n}(y_i^2\sqrt{2}/\sigma^2-1/\sqrt{2}),$$

and let $Y_n(t)$ be the corresponding process:

$$Y_n(t)=n^{-\frac{1}{2}}S_{[nt]}+n^{-\frac{1}{2}}(nt-[nt])(y_{[nt]+1}^2\sqrt{2}/\sigma^2-1/\sqrt{2}).$$

The following lemma relates the processes $X_n(t)$ and $Y_n(t)$.

Lemma 3. *With the preceding definitions,*

$$\sup_{0\leqq t\leqq 1}|X_n(t)-Y_n(t)|\to 0 \quad \textit{in probability, as } n\to\infty.$$

Proof. According to the definitions of $X_n(t)$ and $Y_n(t)$,

$$\begin{aligned}\sup_{0\leqq t\leqq 1}|X_n(t)-Y_n(t)| &= \max_{1\leqq j\leqq n}(\sqrt{2}/\sqrt{n}\,\sigma^2)\left|\sum_{i=1}^{j}(y_i^2-\varepsilon_i^2)\right| \\ &\leqq(\sqrt{2}/\sqrt{n}\,\sigma^2)\sum_{i=1}^{n}M(x_i)^2+\max_{i\leqq j\leqq n}(\sqrt{2}/\sqrt{n}\,\sigma^2)\left|2\sum_{i=1}^{j}M(x_i)\,\varepsilon_i\right|.\end{aligned} \tag{3.11}$$

From Lemma 1 and assumption *A 1*, the first term on the *RHS* of Eq. (3.11) tends to zero a.s. Further by *E 1* and Kolmogorov's inequality for the martingale difference sequence $\{M(x_n)\,\varepsilon_n\}$, one obtains for each $\varepsilon>0$,

$$P\left\{\max_{k\leqq j\leqq n}n^{-\frac{1}{2}}\left|\sum_{i=1}^{j}M(x_i)\,\varepsilon_i\right|>\varepsilon\right\}\leqq\varepsilon^{-2}n^{-1}\sum_{i=1}^{n}EM(x_i)^2\,\sigma^2/2=o(1).$$

Thus, the assertion follows.

Lemma 4. *$Y_n(t)\to W(t)$ weakly as $n\to\infty$.*

Proof. This result is implied by Lemmas 2 and 3 (see, e.g. Billingsley [1968], Theorem 4.1).

In order to prove Theorem 4, it suffices to show that for any fixed m, $P\{N_k>mk\}$ converges to a non-zero number as $k\to\infty$. Now,

$$P\{N_k>mk\}=1-P\{\min_{k\leqq i\leqq mk}u_i(k)\leqq c(k)\} \tag{3.12}$$

where $c(k)$ is the stopping boundary of the stopping rule N_k. Hence Theorem 4 will be proved by showing that the probability of expression (3.12) converges to some number between 0 and 1, as $k\to\infty$. We will relate the statistic $u_n(k)$ to a process involving $Y_n(t)$.

Proof of Theorem 4. Define, for some fixed m,

$$V_n(t)=Y_n(t)-Y_n(t-m^{-1}).$$

From Lemma 4 and Proposition 13.17 of Breiman [1968],

$$V_n(t)\to W(t)-W(t-m^{-1}) \tag{3.13}$$

weakly as $n\to\infty$.

Since

$$\begin{aligned}V_{mk}(t)&=(mk)^{-\frac{1}{2}}(S_{[mkt]}-S_{[mkt]-k})\\&\quad+(mk)^{-\frac{1}{2}}(mkt-[mkt])\,(\sigma^{-2}\sqrt{2})\,(y^2_{[mkt]+1}-y^2_{[mkt]-k+1}),\end{aligned}$$

we have that for $i=1,2,\ldots,mk$,

$$\begin{aligned}V_{mk}(i/mk)&=(mk)^{-\frac{1}{2}}(S_i-S_{i-k})\\&=(2k/m)^{\frac{1}{2}}\,(k\sigma^2)^{-1}\sum_{j=i-k+1}^{i}y_j^2-(k/2m)^{\frac{1}{2}}.\end{aligned}$$

Now, we define

$$U_{mk}(i/mk)=(2k/m)^{\frac{1}{2}}\,u_i(k)-(k/2m)^{\frac{1}{2}}\qquad(1\leqq i\leqq mk) \tag{3.14}$$

and an element of $C[0,1]$:

$$\begin{aligned}U_{mk}(t)&=(2k/m)^{\frac{1}{2}}\,u_{[mkt]}(k)-(k/2m)^{\frac{1}{2}}\\&\quad+(mkt-[mkt])\,(2k/m)^{\frac{1}{2}}\,[u_{[mkt]+1}(k)-u_{[mkt]}(k)].\end{aligned}$$

With this definition,

$$|U_{mk}(i/mk)-(\sigma^2/s_i^2)\,V_{mk}(i/mk)|=(k/2m)^{\frac{1}{2}}\,|1-\sigma^2/s_i^2|, \tag{3.15}$$

which converges to zero a.s. as $i\to\infty$.

Now, since $(\sigma^2/s_i^2)\to 1$ a.s. as $i\to\infty$, (3.13) and (3.15) imply that

$$U_n(t)\xrightarrow{w} W(t)-W(t-m^{-1})\qquad\text{as } n\to\infty \tag{3.16}$$

where $n=mk$ and $t=i/mk$.

Now suppose that the boundary $c(k)$ is of the form

$$c(k)=c(m/2k)^{\frac{1}{2}}+\tfrac{1}{2} \tag{3.17}$$

where c is some constant independent of k.
Then $u_i(k)\leqq c(k)$ iff $(2k/m)^{\frac{1}{2}}\,u_i(k)-(k/2m)^{\frac{1}{2}}\leqq c$.
From (3.14) and (3.17), it follows that

$$\min_{k\leqq i\leqq mk}u_i(k)\leqq c(k)\quad\text{iff}\quad\inf_{m^{-1}\leqq t\leqq 1}U_{mk}(t)\leqq c.$$

Thus, from (3.11) and (3.16), as $k\to\infty$,

$$P\{N_k > mk\}\to 1-P\{\inf_{m^{-1}\leqq t\leqq 1}[W(t)-W(t-m^{-1})]\leqq c\}\neq 0,1.$$

Hence N_k/k converges in distribution to a non-degenerate random variable as $k\to\infty$.

Note to Theorem 4. If we consider a sequence of stochastic approximation problems indexed by the increasing sample size $n=mk$, then N_k/k converges in distribution as $n\to\infty$ provided that $Y_n(t)-Y_n(t-m^{-1})$ is not degenerate; i.e. if $k/n=m^{-1}$ does not converge to zero.

3.3 The Asymptotic Normality of the Randomly Stopped Robbins-Monro Process

The weak convergence of the process $U_n(t)$, derived from $u_n(k)$, provides us with further insight into the nature of the stopping rule N_k. We now consider the problem of finding a confidence interval using x_{N_k}. The stochastic approximation process $\{x_n\}$ is essentially a partial sum of dependent random variables. Using this representation, McLeish [1976] has shown (under certain assumptions) the weak convergence of the Robbins-Monro process. Walk [1977] has proved a more general version of this theorem under slightly weaker conditions, but McLeish's representation is more convenient.

Define $x(t)=x_{[t]+1}$ and $W_n(t)=n^{\frac{1}{2}-\beta}[nt]^{\beta}\,(x(nt)-\theta)$ where $\beta=AM'(\theta)$. Then if $\beta>\frac{1}{2}$, W_n converges weakly to $W_{(\beta)}$, a Gaussian process of independent increments, with mean 0 and variance $A^2\sigma^2 t^{2\beta-1}/(2\beta-1)$.

Note that the choice $A=(M'(\theta))^{-1}$ minimizes the asymptotic variance of $W_{(\beta)}$, and with this value the limiting process is Brownian motion. Using McLeish's process $W_n(t)$ and the process $U_n(t)$ defined in the proof of Theorem 4, we can determine the asymptotic distribution of x_{N_k}; namely, that

$$N_k^{\frac{1}{2}}(x_{N_k}-\theta)/[A^2 s_{N_k}^2/2(2Ab_{N_k}-1)]^{\frac{1}{2}}$$

(with s_n and b_n defined in Sect. 2) converges in distribution as $k\to\infty$, to a Gaussian random variable with mean zero and variance one. We first show how X_{N_k} can be expressed in terms of $W_n(\cdot)$ evaluated at a random time.

Let T be the limit in distribution of N_k/k as $k\to\infty$, which exists and is a proper random variable by Theorem 4. With $\theta=0$, let $W_n(t)$ be defined as for McLeish's result (with $\beta=1$). Now, for any real number m, we define truncated versions $N_k(m)$ and $T(m)$ of N_k and T respectively:

$$N_k(m)=\begin{cases}N_k & \text{if } N_k\leqq km,\\ km & \text{otherwise}\end{cases}$$

and

$$T(m)=\begin{cases}T & \text{if } T\leqq m\\ m & \text{otherwise.}\end{cases}$$

Then $N(m)/km \to T(m)/m$ in distribution as $k \to \infty$. Define two processes on $[0,1]$:

$$\tau^2(t) = A^2\sigma^2 t^{2\beta-1}/2(2\beta-1) = A^2\sigma^2 t/2$$

and

$$\hat{\tau}_n^2(t) = A^2 s_{nt}^2 t^{2Ab} nt^{-1}/2(2Ab_{nt}-1).$$

For m fixed, let $m_k = km$, and define $\Phi_k(t,\omega) = tN_k(m)/m_k$. Then

$$\frac{W_{m_k}(\Phi_k(t))}{\hat{\tau}_{m_k}(\Phi_k(t))} = \frac{m_k^{-\frac{1}{2}}[tN_k(m)]\, x_{tN_k(m)}(tN_k(m)/m_k)^{-(2Ab_{m_k\Phi_k(t)}-1)/2}}{(A^2 s^2_{m_k\Phi_k(t)}/2(2Ab_{m_k\Phi_k(t)}-1)^{\frac{1}{2}}}.$$

Because of the almost sure convergence of the variance estimator s_n^2 to σ^2, of the slope estimator b_n to β, and of Ab_n to 1, we see that as $k \to \infty$,

$$\left| \frac{W_{m_k}(\Phi_k(1))}{\hat{\tau}_{m_k}(\Phi_k(1))} - \frac{(N_k(m))^{\frac{1}{2}} x_{N_k(m)}}{(A^2 s^2_{N_k(m)}/2(2Ab_{N_k(m)}-1))^{\frac{1}{2}}} \right|$$

converges to zero a.s. Thus, if we show that for any real number $m = mk/k$, as $k \to \infty$, $W_{m_k}(\Phi_k(1))/\hat{\tau}_{m_k}(\Phi_k(1))$ converges in distribution to a standard normal random variable, it will follow that

$$(N_k(m))^{\frac{1}{2}} x_{N_k(m)}/[A^2 s^2_{N_k(m)}/2\{2Ab_{N_k(m)}-1\}]^{\frac{1}{2}}$$

converges in distribution to the standard normal as $k \to \infty$.

The result will then follow from Theorem 4.2 of Billingsley [1968] since

$$\lim_{m\to\infty} \lim_{k\to\infty} P\{N_k(m) \neq N_k\} = \lim_{m\to\infty} \lim_{k\to\infty} P\{N_k > km\} = \lim_{m\to\infty} P\{T > m\} = 0,$$

because T is a proper random variable.

The asymptotic normality of $W_{m_k}(\Phi_k(1))/\hat{\tau}_{m_k}(\Phi_k(1))$ will follow from the stronger result (see Theorem 5 below) that for any real number m, the process $W_{m_k}(\Phi_k)/\hat{\tau}_{m_k}(\Phi_k)$ converges weakly to a Gaussian process. The proof follows the lines of Chap. 17 in Billingsley [1968]. Thus we show that the sequence (W_n, U_n) converges weakly to an appropriate limit and apply the continuous mapping theorem. We first present two lemmas.

Lemma 5. *Let* $\phi_i = 1 + O(i^{-1})$ *and define*

$$B_n(t) = (2/\sigma^2 n)^{\frac{1}{2}} \sum_{j=1}^{[nt]} \phi_j \varepsilon_j.$$

Let $X_n(t)$ *be defined as in Eq.* (3.10). *Then* X_n *and* B_n *are uncorrelated for any* n.

Proof. Without loss of generality, take $t < s$. Then

$$\begin{aligned} &\mathrm{Cov}[X_n(t), B_n(s)] = \mathrm{Cov}[X_n(t), B_n(t) + B_n(s) - B_n(t))] \\ &= \mathrm{Cov}[X_n(t), B_n(t)] + \mathrm{Cov}[X_n(t), (B_n(s) - B_n(t))]. \end{aligned}$$

Now, $X_n(t)$ involves centered ε_i^2 terms and $B_n(t)$ is a sum of ε_i terms. Thus, in view of *E 1* and *E 4*

$$\mathrm{Cov}[X_n(t), B_n(t)] = 0.$$

Also, since $X_n(t)$ and $[B_n(s)-B_n(t)]$ involve disjoint sets of independent ε's,

$$\mathrm{Cov}[X_n(t), B_n(s)-B_n(t)]=0.$$

Thus, $\mathrm{Cov}[X_n(t), B_n(s)]=0$, i.e. $X_n(t)$ and $B_n(s)$ are uncorrelated.

Let U denote the weak limit of $\{X_n\}$ and W denote the weak limit of $\{B_n\}$. By the proof of Theorem 4 and McLeish's proof, both U and W are Gaussian processes. Select versions $\tilde{U}$ and $\tilde{W}$ that are independent.

Lemma 6. $(X_n, B_n) \xrightarrow{w} (\tilde{U}, \tilde{W})$

Proof. Since the sequences $\{X_n\}$ and $\{B_n\}$ are individually tight, so is the sequence $\{(X_n, B_n)\}$. (Billingsley, 1968, p. 41). It remains to prove convergence of the finite dimensional distributions.

Let r and s be any two positive integers, $t_1, \ldots, t_r$, $u_1, \ldots, u_s$ be arbitrarily chosen points in $[0,1]$, and $a_1, \ldots, a_r$, $b_1, \ldots, b_s$ any real constants. Define

$$L_n = \sum_{i=1}^{r} a_i X_n(t_i) + \sum_{i=1}^{s} b_i B_n(u_i)$$

and

$$L = \sum_{i=1}^{r} a_i \tilde{U}(t_i) + \sum_{i=1}^{s} b_i \tilde{W}(u_i).$$

It suffices to prove than $L_n \to L$ in distribution.

By virtue of Lemma 5 and the weak convergence of $\{X_n\}$ and $\{B_n\}$,

$$E(L_n)=0=E(L) \quad \text{and} \quad \mathrm{var}(L_n)\to\mathrm{var}(L).$$

Now consider the following representations, ignoring terms of linear interpolation:

$$\sum_{i=1}^{r} a_i X_n(t_i) = \sum_{i=1}^{r} \left\{ a_i n^{-\frac{1}{2}} \sum_{j=1}^{[nt_i]} (\sqrt{2}\sigma^{-2}\varepsilon_j^2 - 1/\sqrt{2}) \right\}$$

$$= \sum_{i=1}^{r} \left\{ \sum_{j=i}^{r} a_j n^{-\frac{1}{2}} \right\} \left\{ \sum_{j=[nt_{i-1}]+1}^{[nt_i]} (\sqrt{2}\sigma^{-2}\varepsilon_j^2 - 1/\sqrt{2}) \right\}$$

and

$$\sum_{i=1}^{s} b_i B_n(t_i) = \sum_{i=1}^{s} \left\{ \sum_{j=i}^{s} b_j (2/\sigma^2 n)^{\frac{1}{2}} \right\} \left\{ \sum_{j=[nt_{i-1}]+1}^{[nt_i]} \phi_j \varepsilon_j \right\}.$$

An application of the univariate central limit theorem shows that L_n is asymptotically normal. Hence, $L_n \to L$ in distribution.

Theorem 5. *Assume A 1, Y1, M 1, M 2, M 3, E 1, E 2, E 3, and E 4 hold. Suppose further that* $A=[M'(\theta)]^{-1}$ *and that* $M(x)$ *has the form*

$$M(x)=(x-\theta)M'(\theta)+0(x-\theta)^2).$$

Then $W_{m_k}(\Phi_k)/\hat{\tau}_{m_k}(\Phi_k)$ *converges weakly to a Gaussian process.*

Proof. From the proof of Theorem 4, we have

$$\sup_{0 \leqq t \leqq 1} |U_n(t) - X_n(t)| \xrightarrow{P} 0 \quad \text{as } n \to \infty.$$

McLeish [1976] shows that

$$\sup_{0 \leqq t \leqq 1} |W_n(t) - B_n(t)| \xrightarrow{P} 0 \quad \text{as } n \to \infty.$$

Hence, by Lemma 6,

$$(U_n, W_n) \xrightarrow{w} (\tilde{U}, \tilde{W}).$$

Since N_k is a continuous function of $\tilde{U}$ (except for a set of paths of measure 0 under $\tilde{U}$) and the composition function is also continuous, we have $W_{m_k}(\Phi_k(t)) \to W(tT(m)/m)$ weakly as $k \to \infty$. Furthermore,

$$W_{m_k}(\Phi_k(t))/\hat{\tau}_{m_k}(\Phi_k(t)) = \{W_{m_k}(\Phi_k(t))/\tau(tT(m)/m)\}\ \{\tau(tT(m)/m)/\hat{\tau}_{m_k}(\Phi_k)\}.$$

Given any $\varepsilon > 0$, there exists $K = K(\varepsilon)$ such that for all $k > K$ and for all t,

$$P\{|1 - \tau(tT(m))/\hat{\tau}_{m_k}(tT(m))| > \varepsilon\} < \varepsilon.$$

Thus, the theorem will follow from Slutsky's Theorem (see, for example, Billingsley [1968], Theorem 5.1, Corollary 2) if we establish that $W_{m_k}(\Phi_k(t))/\tau(tT(m)/m)$ converges in distribution to the standard normal as $k \to \infty$.

For any real number a, as $k \to \infty$,

$$\begin{aligned} P\{W_{m_k}(\Phi_k)/\tau(tT(m)/m) \leqq a\} &\to P\{W(tT(m)/m)/\tau(tT(m)/m) \leqq a\} \\ &= \int P\{W(tT(m)/m)/\tau(tT(m)/m) \leqq a \,|\, T(m) = t'\}\, dP\{T(m) \leqq t'\} \\ &= \int F(a)\, dP\{T(m) \leqq t'\} = F(a), \end{aligned}$$

where F is the standard normal distribution function, (McLeish [1976]). That is, conditional on $T(m)$, $W(tT(m)/m)/\tau(tT(m)/m)$ has a standard Gaussian distribution. Evaluation at $t = 1$ yields the conclusion of the theorem.

3.4 On the Choice of the Boundary c

To fully specify the prediction stopping rule and the associated confidence interval of θ, it remains to consider the stopping boundary c. We have seen in Theorem 2 that for large n the statistic $2ku_n(k)$ behaves like a chi-square random variable with k degrees of freedom. Since y_i has mean $M(x_i)$ and variance $\sigma^2/2$, the noncentrality of $2ku_n(k)$ is approximately

$$\lambda_n = 2\sigma^{-2} \sum_{i=n-k+1}^{n} M^2(x_i).$$

Suppose u is a χ_r^2 random variable with noncentrality parameter λ. N. Marakathavalli [1955] gives the uniformly most powerful unbiased size p test of the

hypothesis $H_0 \cdot \lambda = 0$ against $H_1 \cdot \lambda > 0$ as: Reject H_0 when $u \geqq c_p$, where

$$\int_0^{c_p} f_k(u|\lambda=0)\, du = 1-p,$$

for $f_k(\cdot|\lambda)$ the density of u.

In the case of the prediction statistics, the y_i are asymptotically normal and independent, so we consider the noncentrality λ_n. Acceptance of H_0 is in this case equivalent to termination of sampling at step n. So to completely define the prediction stopping time

$$N_k = \inf\{n\colon u_n(k) \leqq c\},$$

we might consider constants c of the form $c = \chi_k^2(p)/2k$ for varying values of p, where $\chi_k^2(p)$ is the upper p-percentile point of a central χ_k^2 distribution.

4. Small Sample Behavior of the Prediction Rule

4.1 Properties of x_{N_k}

Extensive Monte Carlo Studies (Stroup, [1979]) have shown that even for $k=2$ the properties of x_{N_k} are fairly close to the asymptotic results proved in the previous section. A suitable boundary was obtained by chosing $c(2)$ to be the upper 10 percent point of the χ_2^2 distribution divided by 4. Interestingly, the empirical coverage probabilities were not very different for the three error distributions considered: Gaussian (0, 1), uniform, and chi-squared with three degrees of freedom.

To illustrate the results, we confine ourselves to presenting some properties of $x_N = x_{N_2}$ where the regression function is $M(x) = x$, the errors are standard normal, and $x_1 = 5$. Based on 200 replications of the Robbins-Monro process, we find the sample central moments of $x_N/(\hat{\tau}^2/N)^{\frac{1}{2}}$ to be: $\mu_1 = -0.086$, $\mu_2' = 1.02$, $\mu_3' = -0.20$, $\mu_4' = 3.62$. The sample coefficients of skewness and kurtosis are $g_1 = -0.19$ and $g_2 = 0.47$. Neither is significantly different from zero under the null hypothesis that the normalized, randomly stopped Robbins-Monro estimator has a standard Gaussian distribution. Finally, Fig. 1 displays the p percent pseudo-variances (see Tukey [1977]) of $x_N/(\hat{\tau}^2/N)^{\frac{1}{2}}$ for $p = 0.75$, 0.80,

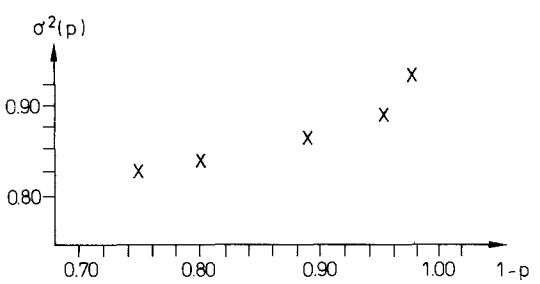

Fig. 1. Pseudovariances of the Sampling Distribution of $x_N/(\hat{\tau}/\sqrt{N})$

0.90, 0.95, 0.975. The pattern suggests that the distribution is perhaps sightly more long-tailed than standard Gaussian variates. Nonetheless, these results support the validity of the Gaussian approximation even for small values of k.

4.2 *Performance Relative to Sielken's Rule*

Sielken's rule and the prediction rule differ in that the former aims at a specific interval length while the latter does not. However, a comparison between the two can be effected as follows. For a particular choice of the parameters x_1, $M'(\theta)$, k, and c, find the average width of the confidence interval with the prediction rule. Using this value as the predetermined width d for Sielken's rule and using the same values for x_1 and $M'(\theta)$, carry out Sielken's rule. Table 1 gives the results for three situations. (Mean square error is given for x_N, s_N^2, and b_N as estimates for θ, σ^2, and $M'(\theta)$, respectively.) Here, we see that the empirical coverage frequency L of Sielken's rule is far from the target value $\gamma = 0.90$, and the improvement over Sielken's method is suggested by the smaller mean square error and bias for the estimate of $M'(\theta)$.

Judged by the criterion of sample size, Sielken's rule performs as well as, or perhaps better than, the prediction rule in this situation (note smaller mean sample sizes and smaller variance of sample size). But the situation investigated here is rather specialized: small error variance (1) and initial estimates $(0, 2, 5)$ very near the target value ($\theta = 0$). So, let us compare the two rules under less ideal conditions. The first two columns of Table 2 give results when the errors

Table 1. A comparison of the prediction rule and Sielken's rule for selected widths of the 90% confidence interval - Gaussian distribution, mean = 0, variance = 1. Based on 50 independent runs. (Same sequence of Gaussian Errors for both S and P.)

		S	P	S	P	S	P
	Width	0.79	0.79	0.47	0.47	0.42	0.42
	VAR[a]		5.17e 0		1.11e−1		4.72e−2
	x_1	0	0	2	2	5	5
	$M'(\theta)$	1	1	1.25	1.25	1.25	1.25
	k		1		2		2
	c		$c_{0.10}$		$c_{0.10}$		$c_{0.10}$
N[b]	MEAN	7.72e 0	9.18e 0	2.17e 1	1.30e 1	1.96e 1	1.21e 1
	VAR	1.78e 2	4.61e 1	2.79e 2	8.75e 2	2.57e 2	5.66e 2
x_N	MEAN	−1.10e−2	2.01e−2	1.09e−2	−4.92e−2	−1.54e−3	7.92e−2
	MSE	1.78e−1	8.97e−2	3.27e−2	4.47e−2	2.84e−2	3.88e−2
	BIAS	1.22e−4	4.10e−4	1.19e−4	2.41e−3	2.39e−6	6.27e−3
s_N^2	MEAN	1.13e 0	1.13e 0	1.80e 0	1.27e 0	1.82e 0	1.20e 0
	MSE	4.75e−1	3.56e−1	9.88e−1	6.33e−1	1.12e 0	5.98e−1
	BIAS	1.84e−2	1.78e−1	6.46e−1	7.54e−2	6.85e 0	3.97e−2
b_N	MEAN	2.07e 0	1.64e 0	1.98e 0	1.26e 0	1.92e 0	1.21e 0
	MSE	2.34e 0	7.97e−1	1.72e 0	7.81e−2	1.48e−1	1.84e−2
	BIAS	1.14e 0	4.05e−1	5.33e−1	6.10e−3	4.45e−1	1.40e−3
L		9.20e−1	8.80e−1	9.80e−1	9.00e−1	9.80e−1	9.40e−1

[a] Sample variance of the width for the prediction rule

[b] N-th stage in RM procedure, at which time $2N$ values have been observed

Table 2. Behavior of the two stage procedure when the variance is large and the initial estimate is poor - Gaussian distribution. Based on 50 independent runs with:

		σ^2 5	x_1 10	$M'(\theta)$ 1.25	k 2	c $c_{0.05}$	d 0.4
		Sielken alone		Prediction as 1st stage		Sielken as 2nd stage[a]	
N	MEAN	1.00e 2		2.99e 1		5.08e 1	
	VAR	0.00e 0[b]		5.64e 2		1.06e 2	
x_N	MEAN	2.41e 2		1.76e 0		2.03e−2	
	MSE	5.93e 2		3.18e 1		1.66e−2	
	BIAS	5.79e 2		3.10e 0		4.12e−4	
	VAR	1.45e 1		2.87e 1		1.61e−2	
s_N^2	MEAN	9.21e 0		5.50e 0		7.25e 0	
	MSE	1.94e 1		4.13e 0		6.03e 0	
	BIAS	1.77e 1		2.50e−1		5.07e 0	
	VAR	1.70e 0		3.88e 0		9.60e−1	
b_N	MEAN	6.17e−1		1.79e 0		1.76e 0	
	MSE	9.54e 0		2.62e 0		1.63e 0	
	BIAS	4.01e−1		2.87e−1		2.68e−1	
	VAR	9.14e 0		2.33e 0		1.36e 0	
W	MEAN			3.94e−1			
	VAR			2.88e 0			

[a] This column refers only to the second stage of the RM process where Sielken's rule is being used
[b] N was assigned the value 100 if sampling had not terminated by that iteration. The zero variance reflects the fact that Sielken's rule failed to terminate in any of the 50 independent runs

are Gaussian but with large variance (5), and a poor initial observation ($x_1 = 10$). Even in this relatively mild deviation from ideal conditions, Sielken's rule failed to converge (in 100 iterations) for any of the 50 independent runs.

The investigation of Sielken's estimates of variance and slope indicates that large values of $|x-\theta|$ or $|M(x)-\alpha|$ introduce a bias term. The discussion of the stopping boundary of the prediction rule indicates that this rule terminates at a point when the noncentrality is small. Therefore, a two-stage procedure seems to be suggested as another possible use for the prediction rule: use the prediction rule initially to find a point x_{N_k} which is near θ (in the sense of small noncentrality); then use x_{N_k} as an improved initial estimate for Sielken's procedure. This procedure appears to be useful, for example, in cases where the convergence of Sielken's procedure is poor. The third column of Table 2 adds the results for Sielken's rule used in this way. Note the improvement of the two-stage procedure over Sielken's procedure in the parameter estimates as well as the empirical frequency of coverage.

One of the possible uses for the RM stochastic approximation process is in estimating percentage points of a quantal response curve. In this case, the observation $y(x)$ takes two values only, 0 and 1, with $P(y(x)=1)=M(x)$. Two forms commonly used for $P(y(x)=1)$ are the probit form

$$M(x)=\int_{-\infty}^{b(x-a)}(2\pi)^{-\frac{1}{2}}\exp(-\tfrac{1}{2}t^2)\,dt$$

and the logit form

$$M(x)=[1+\exp\{-b(x-a)\}]^{-1}$$

for positive constants a and b. In this work, the logit form will be discussed, partly because this allows more direct programming and analysis.

This situation differs from those discussed previously in that $\mathrm{var}\{y(x)\}$ is not constant at each step but varies with x. However, since $\mathrm{var}\{y(x)\}<\infty$ for each x and $\lim_{x\to\infty}\mathrm{var}\{y(x)\}=\sigma^2<\infty$, the theorems on almost sure convergence of x_n to θ and asymptotic normality of $\{x_n\}$ continue to apply. Wetherill [1963] considered the performance of the RM procedure for a fixed sample size. In order to assess the performance of the prediction rule here, we first consider the asymptotic distribution of the statistic $u_n(k)$. Now, since y_n has a Bernoulli distribution with parameter $M(x_n)$, y_n^2 is distributed as y_n. Thus $y_n^2+y_{n-1}^2+\ldots+y_{n-k+1}^2$ is asymptotically (for large n) distributed as the sum of k Bernoulli variables with mean $\alpha=M(\theta)$, i.e., as a binominal (k,α) variable. Thus for large n, $u_n(k)$ is asymptotically distributed as $(k\sigma^2)^{-1}$ times a binomial (k,α) variable. In this case, we know, in fact, that

$$\sigma^2=\lim_{x\to\theta}\mathrm{var}\{y(x)\}=\lim_{x\to\theta}M(x)\{1-M(x)\}=\alpha(1-\alpha).$$

So the boundary c_p can be taken to be

$$\mathrm{Bin}_p(k,\alpha)/\{k\alpha(1-\alpha)\},$$

where $\mathrm{Bin}_p(k,\alpha)$ is the p-percentile point of a binomial (k,α) distribution.

Table 3. A comparison of the prediction rule and Sielken's rule for the quantal response problem

		S	P	S	P	S	P
	Width	0.47	0.47	2.25	2.25	0.10	0.10
	VAR[a]		1.06		1.16		1.97
	x_1	0	0	0.5	0.5	0.5	0.5
	A	4	4	4	4	4.5	4.5
	k		2		2		2
	c		1.00		1.00		1.00
N	MEAN	2.07e 1	2.46e 1	3.16e 1	6.59e 1	1.00e 2	7.75e 1
	VAR	4.31e 2	1.55e 1	4.09e 2	2.07e 1	0.0[b]	1.03e 2
x_N	MEAN	5.32e−1	−4.59e−1	−2.06e−1	−6.26e−1	1.50e 1	1.92e−1
	MSE	3.01e−1	2.47e−1	4.97e−1	4.15e−1	2.27e 2	3.75e 0
	BIAS	2.83e−1	2.11e−1	4.25e−2	3.92e−1	2.26e 2	3.68e−2
s_N^2	MEAN	1.23e 0	2.12e−1	1.50e−1	2.05e−1	4.99e 0	2.07e−1
	MSE	2.40e 0	6.48e−2	3.26e−1	1.06e−1	3.59e 1	7.03e−2
	BIAS	9.60e−1	1.44e−3	1.01e−1	2.02e−3	2.24e 1	1.85e−3
b_N	MEAN	1.80e 0	2.25e−1	8.92e−1	6.47e−1	6.22e−1	3.01e−1
	MSE	3.67e 0	3.31e−1	8.48e−1	5.69e−1	4.57e−1	5.61e−1
	BIAS	2.40e 0	6.25e−4	4.13e−1	1.58e−1	1.38e−1	2.60e−3
L		6.40e−1	9.00e−1	7.80e−1	9.00e−1	0	9.20e−1

[a] Sample variance of the width for the prediction rule

[b] Sielken's rule failed to terminate within 100 iterations in any of the 50 independent runs

Experimental sampling was done for the case $M(x)=\{1+\exp(-x)\}^{-1}$, with $\alpha=0.5$ (i.e., the median). The value $k=2$ was used for the prediction rule, and the boundary value was taken to be 1.0. As in Table 1, Sielken's rule was used with the preassigned width taken to be the value of W obtained in the runs of the prediction rule. The results are given in Table 3. The prediction rule performs markedly better in terms of the slope estimate and the empirical coverage frequency L.

4.3 Remarks

It should be noted that we have not investigated here the results of applying Sielken's stopping rule N_S but with the improved estimators s_n^2 and b_n rather than Burkholder's $\hat{\sigma}_n^2$ and t_n. It is assumed that the reduction of bias would be beneficial in this case as well. In practice, it is expected that two stage procedures of Sect. 4.3 would be preferable, since Burkholder's estimates perform well in a neighbourhood of θ.

More generally, further research needs to be done to determine efficient stopping boundaries and to extend these results to the more practical adaptive stochastic approximation schemes studied by Lai and Robbins.

Acknowledgements. The authors wish to thank S. Livingston for suggesting the need for practical stopping rules. The authors benefitted greatly from many discussions with P. Bloomfield and J. McBride as well as the detailed comments on an earlier draft by a referee.

References

Billingsley, P.: Convergence of Probability Measures. New York: John Wiley 1968

Blum, J.: Approximation Methods which Converge with Probability One. Ann. Math. Statist. **25**, 382–386 (1954)

Breiman, L.: Probability. Reading: Addison-Wesley 1968

Burkholder, D.: On a Class of Stochastic Approximation Processes. Ann. Math. Statist. **27**, 1044–1059 (1956)

Chow, Y., Robbins, H.: On the Asymptotic Theory of Fixed-width Sequential Confidence Intervals for the Mean. Ann. Math. Statist. **36**, 457–462 (1965)

Chung, K.L.: On a Stochastic Approximation Method. Ann. Math. Statist. **25**, 463–483 (1954)

Chung, K.L.: A Course in Probability Theory (2nd ed.). New York: Academic Press 1974

Cramer, H.: On a New Limit Theorem in Probability Theory. Actualities Sci. Ind. **36**, (1938)

Farrell, R.H.: Bounded Length Confidence Intervals for the Zero of a Regression Function. Ann. Math. Statist. **33**, 237–247 (1962)

Lai, T.L., Robbins, H.: Consistency and Asymptotic Efficiency of Slope Estimates in Stochastic Approximation Schemes. Unpublished Manuscript 1978

Lai, T.L., Robbins, H.: Adaptive Design and Stochastic Approximation. Ann. Statist. **7**, 1196–1221 (1979)

McLeish, D.L.: Functional and Random Central Limit Theorems for the Robbins-Monro Process. J. Appl. Probab. **13**, 148–154 (1976)

Marakathavalli, N.: Unbiased Test for a Specified Value of the Parameter in the Non-central F Distribution. Sankhya **15**, 321–330 (1955)

Neveu, J.: Mathematical Foundations of the Calculus of Probability. San Francisco: Holden Day 1965

Révész, P.: Robbins-Monro Procedure in a Hilbert Space and Its Application in the Theory of Learning Processes II. Studia Sci. Math. Hungar. **8**, 469–472 (1973)

Robbins, H., Monro, S.: A Stochastic Approximation Method. Ann. Math. Statist. **22**, 400–407 (1951)

Sacks, J.: Asymptotic Distribution of Stochastic Approximation Procedures. Ann. Math. Statist. **29**, 373–405 (1958)

Scheber, T.K.: Stochastic Approximation: A Survey. Master's Thesis, Thesis, Naval Postgraduate School, Monterey, California (1973)

Schmetterer, L.: Stochastic Approximation. Proceedings of the Fourth Berkeley Sympos. on Math. Statist. and Probab. **1**, 587–609. Univ. Calif. 1961

Sielken, R.L.: Some Stopping Times for Stochastic Approximation Procedures. Florida State University, Statistics Report M127 (1971)

Sielken, R.L.: Stopping Times for Stochastic Approximation Procedures. Z. Wahrscheinlichkeitstheorie verw. Gebiete **27**, 79–86 (1973)

Stroup, D.F.: Stopping Rules for Stochastic Approximation Procedures. Ph.D. dissertation, Princeton University (1979)

Tukey, J.W.: Exploratory Data Analysis. Reading, Mass: Addison-Wesley 1977

Walk, H.: An Invariance Principle for the Robbins-Monro Process in Hilbert Space. Z. Wahrscheinlichkeitstheorie verw. Gebiete **39**, 135–150 (1977)

Wetherill, G.: Sequential Estimation of Quantal Response Curves. J. Roy. Statist. Soc. B**25**, 1–48 (1963)

Wetherill, G., Chen, H., Vasudeva, R.B.: Sequential Estimation of Quantal Response Curves: A New Method of Estimation. Biometrika **53**, 439–454 (1966)

Received July 15, 1980; in revised form February 25, 1982