The Annals of Probability
1982, Vol. 10, No. 1, 178–187

ALMOST SURE APPROXIMATIONS TO THE ROBBINS-MONRO AND KIEFER-WOLFOWITZ PROCESSES WITH DEPENDENT NOISE[1]

By David Ruppert

University of North Carolina at Chapel Hill

We study a recursive algorithm which includes the multidimensional Robbins-Monro and Kiefer-Wolfowitz processes. The assumptions on the disturbances are weaker than the usual assumption that they be a martingale difference sequence. It is shown that the algorithm can be represented as a weighted average of the disturbances. This representation can be used to prove asymptotic results for stochastic approximation procedures. As an example, we approximate the one-dimensional Kiefer-Wolfowitz process almost surely by Brownian motion and as a byproduct obtain a law of the iterated logarithm.

1. Introduction. The recursive algorithm

$$X_{n+1} = X_n - a_n(f(X_n) + e_n + \beta_n), \qquad n = 1, 2, \cdots, \tag{1.1}$$

where e_n and β_n are random vectors in R^k, β_n converges to 0 almost surely, f is a measurable function (possibly unknown) from R^k to R^k, and a_n is a positive random variable, has been studied by Kushner (1977) and Ljung (1978). They have shown that (1.1) includes the Robbins-Monro (RM) (1951) and Kiefer-Wolfowitz (KW) (1952) stochastic approximation processes, which are methods for locating roots of

$$f(x) = 0. \tag{1.2}$$

With the KW process, our goal is to locate a point in R^k where the unknown, real-valued function V attains a local maximum. It is assumed that for each x in R^k we can observe not $V(x)$ but rather $V(x)$ plus additive noise. To locate such a point, we look for a solution to (1.2) with f equal to the gradient of V. The KW algorithm is given by (1.1), with β_n equal to the error which results from approximating f by finite differences of values of V, and e_n equal to a function of the random errors added to the observations of V used to estimate these differences.

With the RM process, for any x in R^k we can observe $f(x)$ plus additive noise, and the goal is to find a solution to (1.2). The RM algorithm is of the form (1.1), with β_n equal to 0, and e_n equal to the noise added to the observation of $f(x)$.

Most of the classical results for the RM and KW processes require the assumption

$$E(e_n \mid X_1, e_1, \cdots, e_{n-1}) = 0. \tag{1.3}$$

Wasan (1969) discusses many of these results and has an extensive bibliography. We mention several of the major results. Blum (1954) gives sufficient conditions for X_n to converge almost surely to a solution θ of (1.2). Chung (1954) studies the asymptotic behavior of the moments of $(X_n - \theta)$ and uses the method of moments to prove that $(X_n - \theta)$, suitably normalized, converges weakly to a normal distribution. Sacks (1958) proves asymptotic normality by other methods and under weaker conditions. Fabian (1968) proves a theorem which subsumes much of the earlier work on asymptotic normality.

Received December 1978; revised February 1981.

[1] This research was supported by the National Science Foundation through Grant MCS78-01240.

AMS 1970 *subject classifications.* Primary 62L20, secondary 60F15.

Key words and phrases. Stochastic approximation, Robbins-Monro process, Kiefer-Wolfowitz process, dependent random variables, almost sure invariance principle.

More recently, the asymptotic behavior of X_n has been further elucidated. McLeish (1976), Nevel'son and Has'minskii (1976, page 153), and Walk (1977) prove weak invariance principles for RM processes; the latter treats RM processes taking values in a separable real Hilbert space. Also, laws of the iterated logarithm have appeared; e.g., see Gaposkin and Krasulina (1974) and Major (1973).

Kersting (1977) approximates the one-dimensional RM process almost surely by a weighted sum of i.i.d. random variables. This approximation allows results for i.i.d. random variables to be applied in a straightforward manner to the RM process, and many of the results mentioned above are simple corollaries of Kersting's representation.

It should be mentioned again that the above results all assume (1.3). Ljung (1978) states that in many applications, assumption (1.3) is violated because the disturbances e_n are correlated. He, Kushner (1977), and Kushner and Clark (1978) have weakened (1.3) considerably. Ljung establishes only almost sure convergence, not the speed of convergence. Kushner and Clark do prove rates for weak convergence. They define a continuous time process by piecewise constant interpolation of $n^{\gamma}(X_n - \theta)$, where γ depends upon a_n, β_n, and f. They show that with a suitable translation of the time variable, with the translation depending upon n, this process converges weakly in $(D[0, \infty))^k$ to the stationary solution of a stochastic differential equation.

In this paper, (1.3) is not assumed. Using techniques of Kersting (1977), we approximate X_n almost surely as a weighted average of $e_1, \cdots, e_n$. Results for X_n then follow immediately as corollaries of theorems for sums of dependent random variables. We work with a form of (1.1) which is sufficiently general to include the multidimensional RM and KW processes. Kersting's work is confined to the one-dimensional RM process.

The methods of this paper are different than those of Kushner and Clark. They deal with certain measures induced by X_n and prove weak convergence of these measures. We examine the sample paths of X_n and prove strong limit theorems.

In Section 2 we introduce the basic model. Section 3 presents general results. These results and the work of Philipp and Stout (1975) are used in Section 4 to show that the RW and KW processes can be redefined on richer probability spaces, without changing their distributions, so that they are approximated almost surely by Brownian motion. Because the approximation is almost sure and is sufficiently close, it is easy to show that results about the asymptotic fluctuation behavior of Brownian motion hold also for the RM and KW processes. For example, Corollary 4.1 is the law of the iterated logarithm for the KW process; it follows directly from the law of the iterated logarithm for Brownian motion.

2. Notation and assumptions. If x is a vector in R^k, let $x^{(i)}$ be its ith coordinate, and $\| x \| = (\sum_{i=1}^{k} (x^{(i)})^2)$. If A is a matrix, let $A^{(ij)}$ be its i,jth entry and A' be its transpose. Let I be the $k \times k$ identity matrix. For a square matrix M, let $\exp(M) = \sum_{i=0}^{\infty} M^i/i!$, and for $t > 0$, define $t^M = \exp((\log t)M)$. Let $\lambda_{\min}(M)$ be the minimum of the real parts of the eigenvalues of M. Define $\| M \| = (\sum_{i=1}^{p} \sum_{j=1}^{p} [M^{(ij)}]^2)^{1/2}$ whenever M is a $p \times p$ matrix. All relations between random variables are meant to hold almost surely.

The following assumptions define our basic model, which is a special case of algorithm (1.1).

A1. Suppose $0 \leq \tau \leq \frac{1}{4}$, X_1 is in R^k, and

$$X_{n+1} = X_n - n^{-1}(f(X_n) + n^{-2\tau}\beta_n + n^{\tau}e_n)$$

for $n \geq 1$ where e_n and β_n are random vectors in R^k.

A2. Suppose β is a vector in R^k and $\beta_n \to \beta$.

A3. Let D be a $k \times k$ matrix and $\eta > 0$. Suppose

$$f(x) = Dx + O(\| x \|^{1+\eta}) \qquad \text{as} \quad x \to 0.$$

Define $\gamma = \min(2\tau, \frac{1}{2} - \tau)$ or $\frac{1}{2} - \tau$ according as $P(\beta_n \neq 0 \text{ infinitely often}) > 0$ or not. Suppose $\lambda_{\min}(D) > \gamma$.

A4. Suppose that for each $k \times k$ matrix such that $\lambda_{\min}(M) > \frac{1}{2}$

$$\sum_{n=1}^{\infty} n^{-M} e_n \quad \text{converges.}$$

A5. Assume $X_n \to 0$.

A6. Let $\rho > 0$ and assume $\beta_n = \beta + O(n^{-\rho})$.

REMARKS. Assumption A4 holds, for example, when e_n is a martingale difference sequence with uniformly bounded second moments, and McLeish (1975) has theorems which imply A4 under a variety of mixing plus moment conditions. See, especially, his Theorem 2.10.

If (1.2) has a unique solution, then for convenience we take it to be 0, and under conditions given by Ljung (1978), Assumption A5 holds.

3. General results.

LEMMA 3.1. *Assume* A1 *to* A5. *Then* $n^{\delta} X_n \to 0$ *for all* $\delta < \gamma$.

PROOF. Select $\delta < \gamma$ and define $Y_n = (n-1)^{\delta} x_n$. Since

$$(n(n-1)^{-1})^{\delta} = 1 + \delta n^{-1} + O(n^{-2}),$$

it follows from A1, A3, and A5 that

$$Y_{n+1} = Y_n - n^{-1}(D - \delta I + B_n) Y_n - n^{-1+\delta}(n^{\tau} e_n + n^{-2\tau} \beta_n)$$

where the matrix B_n satisfies $B_n = O(n^{-1} + \| X_n \|^{\eta}) = o(1)$. Since $\beta_n = O(1)$ and $\delta < \gamma \leq 2\tau$ if $P(\beta_n \neq 0 \text{ } i.o.) > 0$,

$$\sum_{n=1}^{\infty} n^{-1+\delta-2\tau} \beta_n \quad \text{converges.}$$

By A4 and since $\delta < \gamma \leq \frac{1}{2} - \tau$,

$$\sum_{n=1}^{\infty} n^{-1+\delta+\tau} e_n \quad \text{converges.}$$

Therefore, if we set $A = D - \delta I$, then

$$Y_{n+1} = Y_n - n^{-1}(A + B_n) Y_n + d_n \tag{3.1}$$

where $\sum_{n=1}^{\infty} d_n$ converges. Let Ω^* be the subset of the probability space on which $B_n \to 0$ and $\sum_{n=1}^{\infty} d_n$ converges. Thus $P\Omega^* = 1$. Fix ω in Ω^*, and until the end of the proof write X instead of $X(\omega)$ for any random variable X.

Abbreviate $\lambda_{\min}(A)$ and $\| A \|$ by λ and H. Since $\delta < \gamma < \lambda_{\min}(D)$, we have $\lambda > 0$. Choose ξ_0 in $(0, \lambda)$. By Hirsch and Smale (1974, page 146), there exists a norm $\| \ \|_*$ on $\mathbb{R}^p$ such that $\| [\exp(At)] x \|_* \leq \exp(\xi_0 t) \| x \|_*$ for all t in $\mathbb{R}$ and x in $\mathbb{R}^p$. Therefore, we can find $t_0 > 0$ and $\xi \in (0, \xi_0)$ such that

$$\| (I - At) x \|_* \leq (1 - \xi t) \| x \|_* \qquad \text{for} \quad t \in [0, t_0].$$

We will find ϵ, $\epsilon_0 > 0$, a non-negative sequence $\{\Delta(k)\}$, and a sequence of integers $\{n(k+1)\}$ such that $\Delta(k) \to 0$,

$$\| Y_{n(k+1)} \|_* \leq \max((1 - \varepsilon) \| Y_{n(k)} \|_*, \Delta_k (1 + \varepsilon^{-1})), \tag{3.2}$$

and

$$\| Y_{\ell} \|_* \leq (1 + \varepsilon_0) \| Y_{n(k)} \|_* \tag{3.3}$$

for $\ell \in \{n(k), \cdots, n(k+1) - 1\}$. Inequality (3.2) and Lemma 1 of Derman and Sacks (1959), applied with $a_k = \Delta(k)(1 + \varepsilon^{-1})$, $\delta_k = b_k = 0$, and $c_k = \varepsilon$, imply that $Y_{(n)k} \to 0$, and then $Y_n \to 0$ follows from (3.3). Choose $\varepsilon_0 > 0$ such that

$$J_1 = \xi - (1 + \varepsilon_0 + H)\varepsilon_0 > 0 \tag{3.4}$$

and define

$$J_2 = H + \varepsilon_0(1 + \varepsilon_0) + H\varepsilon_0. \tag{3.5}$$

Choose $\varepsilon > 0$ such that

$$J_1 \min\{t_0/2, (\varepsilon_0 - \varepsilon)J_2^{-1}\} = 2\varepsilon. \tag{3.6}$$

We now define $\{n(k)\}$ inductively. For each k we define

$$\Delta(k) = \sup\{\| \textstyle\sum_{i=n(k)}^{\ell} d_i \| : n(k) \le \ell\}.$$

First choose $n(1)$ such that $\Delta(n(1)) \le \varepsilon$, $n(1) \ge 2/t_0$, and $\sup\{\| B_k \| : k \ge n(1)\} < \varepsilon_0$. Suppose we have chosen $n(1), \cdots, n(k)$. Then let

$$m_k = \inf\{i > n(k) : \| Y_i - Y_{n(k)} \|_* > \varepsilon_0 \| Y_{n(k)} \|_*\},$$

$$m_k' = \inf\{i > n(k) : \textstyle\sum_{j=n(k)}^{i-1} j^{-1} \ge t_0/2\},$$

and $n(k+1) = \min\{m_k, m_k'\}$. Define $S_k = \{n(k), \cdots, n(k+1) - 1\}$ and $t_k = \sum_{i \in S(k)} i^{-1}$. Since $(n(k+1) - 1)^{-1} < t_0/2$, we have $t_k \le t_0$. By (3.1),

$$Y_{n(k+1)} = R_1 + R_2 + R_3 + R_4$$

where

$$R_1 = (I - \textstyle\sum_{i \in S_k} i^{-1} A) Y_{n(k)},$$

$$R_2 = -\textstyle\sum_{i \in S_k} i^{-1} B_i Y_i,$$

$$R_3 = -\textstyle\sum_{i \in S_k} i^{-1} A (Y_i - Y_{n(k)})$$

$$R_4 = \textstyle\sum_{i \in S_k} d_i.$$

Then

$$\| R_1 \|_* \le (1 - \xi t_k) \| Y_{n(k)} \|_*,$$

$$\| R_2 \|_* \le t_k \varepsilon_0 (1 + \varepsilon_0) \| Y_{n(k)} \|_*,$$

$$\| R_3 \|_* \le t_k H \varepsilon_0 \| Y_{n(k)} \|_*,$$

and

$$\| R_4 \|_* \le \Delta(k).$$

Therefore,

$$\| Y_{n(k+1)} \|_* \le (1 - J_1 t_k) \| Y_{n(k)} \|_* + \Delta(k). \tag{3.7}$$

If $n(k+1) = m_k'$, then $t_k \ge t_0/2$. On the other hand, if $n(k+1) = m_k$ and $\| Y_{n(k)} \|_* \ge \Delta(k)\varepsilon^{-1}$, then

$$\| Y_{n(k+1)} - Y_{n(k)} \|_* = \| \textstyle\sum_{i \in S_k} i^{-1} A Y_{n(k)} + R_2 + R_3 + R_4 \|_* \le (t_k J_2 + \varepsilon) \| Y_{n(k)} \|_*,$$

and therefore by the definition of m_k,

$$\varepsilon_0 \le t_k J_2 + \varepsilon \qquad \text{so that} \quad t_k \ge (\varepsilon_0 - \varepsilon) J_2^{-1}.$$

Therefore, if $\| Y_{n(k)} \|_* \ge \Delta(k)\varepsilon^{-1}$, then

$$t_k \ge \min\{t_0/2, (\varepsilon_0 - \varepsilon) J_2^{-1}\},$$

and by (3.6) and (3.7),

$$\| Y_{n(k+1)} \|_* \le [1 - (J_1 \min\{t_0/2, (\varepsilon_0 - \varepsilon) J_2^{-1}\} - \varepsilon)] \| Y_{n(k)} \|_* = (1 - \varepsilon) \| Y_{n(k)} \|_*.$$

If $\| Y_{n(k)} \|_* \le \Delta(k)\varepsilon^{-1}$, then by (3.7),

$$\| Y_{n(k+1)} \|_* \le \Delta(k)(1 + \varepsilon^{-1}).$$

Therefore, (3.2) holds. Finally, (3.3) holds by definition of m_k. □

Define $\delta(x, y)$ to equal 0 or 1 according to whether $x \neq y$ or $x = y$.

THEOREM 3.1 *Assume A1 to A6. Then there exists* $\varepsilon > 0$ *such that* $n^{\gamma}X_{n+1} = -(D - 2\tau I)^{-1}\beta\delta(\gamma, 2\tau) - n^{-1/2}(\sum_{k=1}^{n}(k/n)^{D+(\tau-1)I}e_k)\delta(\gamma, \tfrac{1}{2} - \tau) + O(n^{-\varepsilon})$.

PROOF. By A1 and A3,

$$X_{n+1} = X_n - n^{-1}(D + b_n)X_n - n^{-1}(n^{\tau}e_n + n^{-2\tau}\beta_n) \tag{3.8}$$

where $b_n = O(\|X_n\|^{\eta})$. By Lemma 3.1,

$$b_n = O(n^{-\eta\delta}) \qquad \text{for all} \quad \delta < \gamma. \tag{3.9}$$

We now assume that D has only one eigenvalue, λ. This involves no loss of generality since we can, by a change of basis, put D in real canonical form (Hirsch and Smale, 1974, page 130) so that

$$D = \operatorname{diag}(D_1, \cdots, D_q), \qquad q \leq p,$$

where each of $D_1, \cdots, D_q$ is a square matrix with exactly one eigenvalue. Then we can partition X_n in an analogous fashion, and our proof can be applied separately to each element of the partition. Since $\log(n-1) = \log n - n^{-1} + O(n^{-2})$, $(n-1)^D = \exp((\log n - n^{-1} + O(n^{-2}))D) = n^D - Dn^{D-I} - \nu_n$ where $\nu_n = O(n^{-2}\|n^D\|)$. By Hirsch and Smale (1974, page 146), $\|n^D\| = O(n^{\lambda+\varepsilon})$ for all $\varepsilon > 0$, so

$$\nu_n = O(n^{-2+\lambda+\varepsilon}) \qquad \text{for all} \quad \varepsilon > 0. \tag{3.10}$$

Then from (3.8)

$$n^D X_{n+1} = (n-1)^D X_n + (\nu_n - n^{D-I}b_n)X_n - n^{D-I}(n^{\tau}e_n + n^{-2\tau}\beta_n),$$

and upon iteration we obtain

$$n^D X_{n+1} = X_2 + \sum_{j=2}^{n}(\nu_j - j^{D-I}b_j)X_j - \sum_{j=2}^{n} j^D j^{-1+\tau}e_j - \sum_{j=2}^{n} j^D j^{-1-2\tau}\beta_j. \tag{3.11}$$

Since $\lambda > \gamma$, it follows from (3.9), (3.10), and Lemma 3.1 that

$$n^{\gamma-\lambda}\|(\nu_n - n^{-1+\lambda}b_n)X_n\| = O(n^{-1-\alpha})$$

for some $\alpha > 0$. We can and will assume that $\alpha < 2(\lambda - \gamma)$. Therefore,

$$\sum_{n=1}^{\infty} n^{-\lambda+\gamma+\alpha/2}\|(\nu_n - n^{-1+\lambda}b_n)X_n\| < \infty,$$

and it follows from Kronecker's lemma that

$$n^{-\lambda+\gamma+\alpha/2}\sum_{j=2}^{n}\|(\nu_j - j^{-1+\lambda_i}b_j)X_j\| = o(1).$$

Since $\|n^{-D}x\| \leq K(n^{-\lambda+\alpha/4}\|x\|)$ for some K and all $x \in \mathbb{R}^p$,

$$\|n^{-D+\gamma I}\sum_{j=2}^{n}(\nu_j - j^{-1+\lambda}b_j)X_j\| = O(n^{-\alpha/4}). \tag{3.12}$$

Using A6 and an argument similar to the one which established (3.12), we can prove that

$$\|n^{-D+\gamma I}\sum_{j=2}^{n} j^{D-(\gamma+1)I}(\beta_j - \beta)\| = O(n^{-\rho/2}).$$

For all $\varepsilon > 0$,

$$n^{-D+\gamma I}\sum_{j=2}^{n} j^{-I+D-\gamma I} = n^{-D+\gamma I}\left(\int_1^n x^{-I+D-\gamma I}dx + O\left(\int_1^n x^{\lambda-\gamma-2+\varepsilon}dx\right)\right)$$

$$= (D - \gamma I)^{-1} + O(n^{-1+\varepsilon}).$$

Thus, for some $\varepsilon > 0$,

$$n^{-D+\gamma I}\sum_{j=1}^{n} j^D j^{-1-2\tau}\beta_j = (D - \gamma I)^{-1}\beta\delta(2\tau, \gamma) + O(n^{-\varepsilon}). \tag{3.13}$$

By (3.11) to (3.13), for some $\varepsilon > 0$,

$$n^{\gamma}X_{n+1} = -n^{-D}n^{\gamma}\sum_{j=1}^{n} j^{D}j^{-1+\tau}e_j - (D - \gamma I)^{-1}\beta\delta(2\tau, \gamma) + O(n^{-\varepsilon}). \tag{3.14}$$

If $\gamma \neq \frac{1}{2} - \tau$, then $\gamma = \frac{1}{2} - \tau - \Delta$ for some $\Delta > 0$. Thus by A4

$$\sum_{j=1}^{\infty} (j^{D}j^{-1+\tau})e_j j^{-\lambda+\gamma+\Delta/2} \quad \text{converges,}$$

and therefore by Kronecker's lemma

$$n^{-\lambda+\gamma+\Delta/2}\sum_{j=1}^{n} j^{D-I+\tau I}e_j = o(1).$$

Since $n^{-D}n^{\gamma} = O(n^{-\lambda+\gamma+\varepsilon})$ for all $\varepsilon > 0$,

$$n^{-D}n^{\gamma}\sum_{j=1}^{n} j^{D-I+\tau I}e_j = o(n^{-\Delta/3}).$$

Thus we have shown that for some $\varepsilon > 0$,

$$n^{-D+\gamma I}\sum_{j=1}^{n} j^{D-I+\tau I}e_j = n^{(1/2-\tau)I-D}\sum_{j=1}^{n} j^{D-I+\tau I}e_j\delta\left(\gamma, \frac{1}{2} - \tau\right) + O(n^{-\varepsilon}). \tag{3.15}$$

Substituting (3.15) into (3.14) completes the proof. □

Since in some applications D will be symmetric, and therefore upon a change of basis diagonal, we state the following special case of Theorem 3.1.

COROLLARY 3.1 *Suppose A1 to A6 hold and* $D = \operatorname{diag}(\lambda_1, \cdots, \lambda_j)$. *Then there exists* $\varepsilon > 0$ *such that, for* $i = 1, \cdots, k$,

$$n^{\gamma}X_{n+1}^{(i)} = -(\lambda_i - 2\tau)^{-1}\beta^{(i)}\delta(\gamma, 2\tau) - n^{-1/2}\sum_{k=1}^{n} (k/n)^{\lambda_i+\tau-1}e^{(i)}\delta\left(\gamma, \frac{1}{2} - \tau\right) + O(n^{-\varepsilon}).$$

4. The one-dimensional RM and KW processes. Theorem 3.1 enables us to use theorems for sums of dependent random variables to prove theorems for stochastic approximation processes with dependent noise. In this section we apply the work of Philipp and Stout (1975) to the RM and KW processes in R^1. Their monograph gives sufficient conditions so that a sequence of random variables e_n can be redefined on a richer probability space without changing its distribution, together with a Brownian motion $B(t)$ on $[0, \infty)$ such that, for some $\varepsilon > 0$,

$$\sum_{k \le t} e_k = B(t) + O(t^{1/2-\varepsilon}). \tag{4.1}$$

For example, suppose e_k is a strictly stationary ϕ-mixing process such that $\sum_{n=1}^{\infty} (\phi(n))^{1/2} < \infty$. (See Philipp and Stout, page 26, for the definition of ϕ-mixing.) If $Ee_1 = 0$ and $E|e_1|^{2+\delta} < \infty$ for some $\delta > 0$, then $\lim_{n\to\infty} n^{-1}E(\sum_{k=1}^{n} e_k)^2$ exists (Philipp and Stout, page 26). Call this limit σ^2. Suppose that $\sigma^2 > 0$. Then without loss of generality $\sigma^2 = 1$ can be assumed. Then by their Theorem 4.1, Equation (4.1) holds. We will be interested in the asymptotic behavior of $\sum_{k \le t} k^{\alpha}e_k$ where e_k satisfies (4.1), so the following lemma is useful.

LEMMA 4.1. *Let* e_n *be a sequence of random variables. For any number* α, *define* S_α *on* $[0, \infty)$ *by*

$$S_\alpha(t) = \sum_{k \le t} k^{\alpha}e_k.$$

Suppose there exists a standard Brownian motion $B_0(t)$ *on* $[0, \infty)$ *and a positive number* ε *such that*

$$S_0(t) = B_0(t) + O(t^{1/2-\varepsilon}).$$

Then for $\alpha < -\frac{1}{2}$,

(4.2) $\lim_{t\to\infty} S_\alpha(t)$ *exists and is finite,*

and for $\alpha > -\frac{1}{2}$, *there exists a standard Brownian motion* B_α *and a positive number* ε' *such that*

$$S_\alpha(t) = B_\alpha(t^{2\alpha+1}(2\alpha+1)^{-1}) + O(t^{\alpha+1/2-\varepsilon'}). \tag{4.3}$$

PROOF. Define $N(0, \alpha) = 0$ and $N(k, \alpha) = \sum_{j=1}^{k} j^{2\alpha}$. Then define for $N(k-1, \alpha) < t \le N(k, \alpha)$

$$B_\alpha(t) = \sum_{j=1}^{k-1} j^\alpha(B_0(j) - B_0(j-1)) + k^\alpha(B_0(k^{-2\alpha}(t - N(k-1, \alpha)) + k - 1) - B_0(k-1)).$$

Since B_0 is a standard Brownian motion, so also is B_α. Now

$$\begin{aligned} S_\alpha(n) &= n^\alpha e_n + \sum_{k=1}^{n-1} \left(\sum_{j=k}^{n-1} (j^\alpha - (j+1)^\alpha) + n^\alpha\right) e_k \\ &= n^\alpha \sum_{k=1}^{n} e_k + \sum_{j=1}^{n-1} \sum_{k=1}^{j} (j^\alpha - (j+1)^\alpha) e_k \\ &= n^\alpha S_0(n) + \sum_{j=1}^{n-1} (j^\alpha - (j+1)^\alpha) S_0(j) \\ &= n^\alpha B_0(n) + \sum_{j=1}^{n-1} (j^\alpha - (j+1)^\alpha) B_0(j) + O(n^{\alpha+1/2-\varepsilon} + \sum_{j=1}^{n-1} j^{\alpha-1/2-\varepsilon}) \\ &= \sum_{j=1}^{n} j^\alpha(B_0(j) - B_0(j-1)) + O(n^{\alpha+1/2-\varepsilon} + \sum_{j=1}^{n} j^{\alpha-1/2-\varepsilon}). \end{aligned}$$

Thus,

$$S_\alpha(n) = B_\alpha(N(n, \alpha)) + O(n^{\alpha+1/2-\varepsilon}). \tag{4.4}$$

(If $\alpha - \varepsilon = -\frac{1}{2}$, then the remainder is $O(\log n)$, not $O(n^{\alpha+1/2-\varepsilon})$). Therefore, if $\alpha < -\frac{1}{2}$, then (4.2) must hold.

One can easily show that for $\alpha > -\frac{1}{2}$

$$N(k, \alpha) = (2\alpha+1)^{-1} t^{2\alpha+1} + O(t^{2\alpha}). \tag{4.5}$$

In the proof of their Lemma 3.5.3, Philipp and Stout (1975) show that if $1 > \delta > 0$ and B is a Brownian motion on $[0, \infty)$, then for each $\mu > 0$

$$B(t + O(t^{1-\delta})) = B(t) + O(t^{1/2-\delta/2+\mu}). \tag{4.6}$$

By (4.4) to (4.6), (4.3) holds whenever $\alpha > -\frac{1}{2}$. □

LEMMA 4.2. *Suppose* a_k *are real numbers and* $\sum_{k=1}^{\infty} a_k$ *converges. Let* b_k^n, $k = 1, \cdots, n$ *and* $n \ge 1$, *be positive numbers. Suppose* $\sup_n b_n^n \le M < \infty$, *and for each* n *suppose* $b_k^n \le b_{k+1}^n$. *Assume that, for each* k, b_k^n *decreases to* 0 *as* $n \to \infty$. *Then* $\lim_{n\to\infty} \sum_{k=1}^{n} a_k b_k^n = 0$.

PROOF. Fix $\varepsilon > 0$. Choose N such that if $n \ge m \ge N$, then $|\sum_{k=m}^{n} a_k| \le \varepsilon$. Then choose $N' \ge N$ such that $b_k^n (\sum_{j=1}^{N} |a_j|) \le \varepsilon$ for $k = 1, \cdots, N$ and $n \ge N'$. For $k \le n$, define $c_k^n = b_k^n - b_{k-1}^n$ if $k \ge 2$ and $c_1^n = b_1^n$. Then if $n \ge N$,

$$|\sum_{k=N}^{n} a_k b_k^n| = |\sum_{k=N}^{n} a_k (\sum_{j=1}^{k} c_j^n)| = |\sum_{j=1}^{n} c_j^n (\sum_{k=\max(j,N)}^{n} a_k)| \le b_n^n \varepsilon \le M\varepsilon.$$

Therefore if $n \ge N'$,

$$|\sum_{k=1}^{n} a_k b_k^n| \le |\sum_{k=1}^{N} a_k b_k^n| + |\sum_{k=N+1}^{n} a_k b_k^n| \le \varepsilon(1 + M).$$ □

Now we apply our results to the one-dimensional KW process. Let M be a differentiable function from R' to R'. Suppose that θ is the unique solution to

$$\inf_x M(x) = M(\theta)$$

and that $M'(x) = 0$ implies that $x = \theta$. We will assume that M''' exists and is continuous in a neighborhood of θ. For convenience, we also assume that $\theta = 0$. The KW process is

defined by the recursion

$$X_{n+1} = X_n - an^{-1}Y_n,$$

where Y_n is an estimate of $M'(X_n)$. Since we assume only that, for each x, an estimate of $M(x)$, but not of $M'(x)$, can be directly observed, Y_n is defined as follows. Let c and τ be positive constants and, for $i = 1, 2$, let $M(X_n + (-1)^i cn^{-\tau})$ be estimated by $Y_{n,i} = M(X_n + (-1)^i cn^{-\tau}) + \mu_{n,i}$. Then set

$$Y_n = (Y_{n,2} - Y_{n,1})/(2cn^{-\tau}).$$

If we define β_n by

$$a^{-1}n^{-2\tau}\beta_n = (M(X_n + cn^{-\tau}) - M(X_n - cn^{-\tau}))/(2cn^{-\tau}) - M'(X_n),$$

then

$$(ac^2/12)(M'''(\eta_n) + M'''(\rho_n))$$

where $$|X_n - \rho_n| \le cn^{-\tau} \quad \text{and} \quad |X_n - \eta_n| \le cn^{-\tau}.$$

Therefore, if we assume that $0 \le \tau \le \frac{1}{4}$ and define $\mu_n = \mu_{n,2} - \mu_{n,1}$, then A1 holds with $f(x) = aM'(x)$, and $e_n = (a\mu_n)/(2c)$. Since $P(\beta_n \ne 0 \text{ infinitely often}) > 0$ unless very restrictive assumptions are made, $\gamma = \min(2\tau, \frac{1}{2} - \tau)$, and therefore γ is equal to its maximum, $\frac{1}{3}$, when $\tau = \frac{1}{6}$. By Theorem 2.5 of Fabian (1971), the rate $n^{-1/3}$ cannot be improved for the KW process. Therefore, we can set $\tau = \frac{1}{6}$. (If M has a continuous derivative of order $s + 1$ in a neighborhood of 0 for s an even integer, then the rate $n^{-s/2(s+1)}$ is obtainable by the procedure given in Fabian's (1971) Theorem 2.6.

REMARK. A referee has pointed out that Theorem 3.1 implies that if $\tau > \frac{1}{6}$, then $n^{2\tau}X_n \to -(6aM''(0) - 12\tau)^{-1}ac^2M'''(0)$. I do not know if this result has been published.

We will also make several additional assumptions.

A7. M has two continuous derivatives, $\sup_x |M''(x)| < \infty$, $\{x: |M'(x)| < \delta\}$ is compact for some $\delta > 0$, and $\{x: M(x) \le c\}$ is compact for all $c > M(0)$.

A8. $M'''(x) = M'''(0) + O(|x|^d)$ as $x \to 0$ for some $d > 0$.

A9. $aM''(0) > \frac{1}{3}$.

A10. Define $S(t) = \sum_{k \le t} \mu_k$. Suppose $\sigma > 0$, $B(t)$ is a standard Brownian motion on $[0, \infty)$, and for some $\delta > 0$

$$S(t) = \sigma B(t) + O(t^{1/2-\delta}).$$

THEOREM 4.1. *Suppose A7 to A10 hold. Let* $A = aM''(0) - (\frac{5}{6})$ *and* $B = -ac^2M'''(0)/(6aM''(0) - 2)$. *Then for some* $\varepsilon > 0$,

$$n^{1/3}X_n = B - n^{-1/2}\sum_{k=1}^{n} (k/n)^A e_k + O(n^{-\varepsilon}). \tag{4.7}$$

Define $X(t)$, $t \ge 0$, *by* $X(t) = n^{1/3}X_n$ *if* $n \le t < n + 1$. *Then there exists a standard Brownian motion* $Z(t)$ *on* $[0, \infty)$ *and* $\varepsilon > 0$ *such that*

$$X(t) = B + (a\sigma)(2c)^{-1}(2A + 1)^{-1/2}t^{-A-1/2}Z(t^{2A+1}) + O(t^{-\varepsilon}). \tag{4.8}$$

PROOF. We first note that A5 holds by Lemma 1 of Ljung (1978). To apply this lemma, his $X(n)$, $e(n)$, $\beta(n)$, $\gamma(n)$, and $f(x)$ are set equal to our X_{n+1}, $-(\mu_n n^{1/6}/(2c))$, $-\beta_n n^{-1/3} = M'(X_n) - (M(X_n + cn^{-1/6}) - M(X_n - cn^{-1/6})/(2cn^{-1/6}))$, an^{-1}, and $-M'(x)$, respectively. Ljung's Condition B1 is verified by using (4.2) of Lemma 4.1, Lemma 4.2, and his Equation (15). After using his Lemma 3 to verify his Condition B2, it is clear that all conditions of his Lemma 1 are satisfied.

A4 holds because of A10 and (4.2). A3 holds with $D = \lambda_1 = aM''(0)$ because A8 implies that $f(x) = Dx + O(x^2)$ as $x \to 0$. Let $\beta = (ac^2/6)M'''(0)$. By Lemma 3.1, $X_n = O(n^{-\varepsilon})$ for

an $\varepsilon > 0$; this and A8 imply that $\beta_n = \beta + O(|X_n|^d + n^{-d/6}) = \beta + O(n^{-\varepsilon})$ for some $\varepsilon > 0$, and so A6 holds. We now invoke Theorem 3.1 to prove that (4.7) holds. By A10

$$\sum_{k \le t} e_k = a\sigma/(2c)B(t) + O(t^{1/2-\delta}).$$

Therefore, by Lemma 4.1, there exists a Brownian motion $Z(t)$ and an $\varepsilon > 0$ such that

$$\sum_{k \le t} k^A e_k = a\sigma(2c)^{-1}(2A+1)^{-1/2}Z(t^{2A+1}) + O(t^{A+1/2-\varepsilon}).$$

This and (4.7) imply (4.8). □

Theorem 4.1 yields results on the asymptotic fluctuation behavior of X_n. Here is a simple example.

COROLLARY 4.1. *Suppose* A7 *to* A10 *hold. Then*

$$\limsup_{n\to\infty} \frac{n^{1/3}X_n}{\sqrt{2\log(\log n)}} = \frac{a\sigma}{(2c)(2A+1)^{1/2}}.$$

PROOF. Straightforward. Use the law of the iterated logarithm for Brownian motion. □

Now we state, without proof, an analogue of Theorem 4.1 for the RM process.

A11. Assume f is a function from R^1 to R^1 and the 0 is the unique solution of

$$f(x) = 0.$$

A12. Let u_n be a sequence of random variables. Suppose $B(t)$ is a standard Brownian motion on $[0, \infty)$, $\varepsilon > 0$, $\sigma < 0$, and

$$\sum_{k \le t} u_k = \sigma B(t) + O(t^{1/2-\varepsilon}).$$

A13. $X_{n+1} = X_n - an^{-1}(f(X_n) + u_n)$.

A14. f has a continuous derivative and $f(x) = f'(0)x + O(|x|^{1+\varepsilon})$ as $x \to 0$ for $\varepsilon > 0$.

A15. Define $V(x) = \int_0^x f(y)\,dy$. Suppose $\{x: V(x) \le C\}$ is compact for all $C < \sup V(x)$.

A16. Suppose $\sup |f'(x)| < \infty$ and $\{x: |f(x)| \le \delta\}$ is compact for some $\delta > 0$.

THEOREM 4.2 *Suppose* A11 *to* A16 *hold and* $af'(0) > \frac{1}{2}$. *Define* $D = af'(0) - 1$. Then for some $\varepsilon > 0$

$$n^{1/2}X_n = -n^{-1/2}a\sum_{k=1}^n (k/n)^D u_k + O(n^{-\varepsilon}).$$

Define $X(t)$, $t \ge 0$, *by* $X(t) = n^{1/2}X_n$ *on* $n \le t < n+1$. *Then there exists a standard Brownian motion* $Z(t)$ *and an* $\varepsilon > 0$ *such that*

$$X(t) = \frac{a\sigma t^{-D-1/2}}{(2D+1)^{1/2}} Z(t^{2D+1}) + O(t^{-\varepsilon}).$$

PROOF. Similar to the proof of Theorem 4.1. □

Acknowledgement. I wish to thank the referees for their careful reading of the original manuscript, for detecting several errors, and for suggestions which improved the clarity of the paper.

REFERENCES

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

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

DERMAN, C. and SACKS, J. (1959). On Dvoretzky's stochastic approximation theorem. *Ann. Math. Statist.* **30** 601–605.
FABIAN, V. (1968). On asymptotic normality in stochastic approximation. *Ann. Math. Statist.* **39** 1327–1332.
FABIAN, V. (1971). Stochastic approximation. In *Optimizing Methods in Statistics* (J. S. Rustagi, ed.). Academic, New York.
GAPOSKIN, V. F. and KRASULINA, T. P. (1974). On the law of the iterated logarithm in stochastic approximation processes. *Theor. Probability Appl.* **19** 844–850.
HIRSCH, M. and SMALE, S. (1974). *Differential Equations, Dynamical Systems, and Linear Algebra.* Academic, New York.
KERSTING, G. (1977). Almost sure approximation of the Robbins-Monro process by sums of independent random variables. *Ann. Probability* **5** 954–965.
KIEFER, J. and WOLFOWITZ, J. (1952). Stochastic estimation of the maximum of a regression function. *Ann. Math. Statist.* **23** 462–466.
KUSHNER, H. J. (1977). General convergence results for stochastic approximations via weak convergence theory. *J. Math. Anal. and Appl.* **61** 490–503.
KUSHNER, H. J. and CLARK, D. S. (1978). *Stochastic Approximation Methods for Constrained and Unconstrained Systems.* Springer-Verlag, New York.
LJUNG, L. (1978). Strong convergence of a stochastic approximation algorithm. *Ann. Statist.* **6** 680–696.
MAJOR, P. (1973). A law of iterated logarithm for the Robbins-Monro method. *Studia Sci. Math. Hungar.* **8** 92–102.
MCLEISH, D. L. (1975). A maximal inequality and dependent strong laws. *Ann. Probability* **3** 829–839.
MCLEISH, D. L. (1976). Functional and random central limit theorems for the Robbins-Monro process. *J. Appl. Probability* **13** 148–154.
NEVEL'SON, M. B. and HAS'MINSKII, R. Z. (1976). Stochastic approximation and recursive estimation. *Trans. of Math. Monographs Vol.* **47**, Amer. Math. Soc., Providence, R.I.
PHILIPP, W. and STOUT, W. (1975). Almost sure invariance principles for partial sums of weakly dependent random variables. *Amer. Math. Soc. Mem. No. 161*, Amer. Math. Soc., Providence, R.I.
ROBBINS, H. and MONRO, S. (1951). A stochastic approximation method. *Ann. Math. Statist.* **22** 400–407.
SACKS, J. (1958). Asymptotic distribution of stochastic approximation procedures. *Ann. Math. Statist.* **29** 373–405.
WALK, H. (1977). An invariance principle for the Robbins-Monro process in a Hilbert space. *Z. Wahrsch. verw. Gebiete* **39** 135–150.
WASAN, M. T. (1969). *Stochastic Approximation.* Cambridge Univ. Press, New York.

DEPARTMENT OF STATISTICS
THE UNIVERSITY OF NORTH CAROLINA
AT CHAPEL HILL
321 PHILIPPS HALL 039 A
CHAPEL HILL, NORTH CAROLINA 27514