Almost Sure Approximations to the Robbins-Monro and Kiefer-Wolfowitz Processes with Dependent Noise

David Ruppert[1]

Abbreviated Title: RM and KW Processes

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.

Key Words and Phrases: Stochastic approximation, Robbins-Monro process, Kiefer-Wolfowitz process, dependent random variables, almost sure invariance principle.

[1]Research supported by National Science Foundation, Grant NSF MCS78-01240.

AMS 1970 subject classification. Primary 62L20; Secondary 60F15.

1. Introduction. The recursive algorithm

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

where e_n and β_n are random vectors in R^k, β_n converges to 0 almost surely, f is a 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

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

Wasan (1969) discusses many of these results and has 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.

More recently, the asymptotic behavior of X_n has been further elucidated. McLeish (1976), Nevel'son and Has'minskii ((1973), 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., 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 $e_1,\ldots,e_n$. 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 multi-dimensional 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 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 flucuation behavior of Brownian motion hold also for the RM and KW processes. For example corollary 4.1 is law of the iterated logarithm for the KW process; it follows directly from the law of the iterated logarithm for Brownian motion.

2. <u>Notation and assumptions</u>. If x is a vector in R^k, let $x^{(i)}$ be its ith coordinate, $||x|| = (\sum_{i=1}^k (x^{(i)})^2)$, and $||x||_\infty = \max\{|x^{(j)}|: j = 1,\ldots,k\}$. If A is a matrix, let $A^{(ij)}$ be its i,jth entry and A^T be its transpose. Let I_k be the k×k identity 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 $0 \leq \tau \leq 1/4$, 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×k symmetric positive definite matrix and $\eta > 0$. Suppose

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

Let $\lambda_k \geq \lambda_{k-1} \geq \cdots \geq \lambda_1$ be the eigenvalues of D.

A4. Supose that for every $\delta > 0$,

$$\sum_{n=1}^{\infty} n^{-(\frac{1}{2}+\delta)} e_n \text{ converges.}$$

A5. Assume $X_n \to 0$.

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

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.

N1. 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.

Let P be an orthogonal matrix such that $P'DP = \text{diag}(\lambda_i)$.

N2. For any x in R^k let

$$\tilde{x} = P'x.$$

3. General Results.

Lemma 3.1. Assume A1 to A5. Suppose $\delta < \gamma < \lambda_1$. Then $n^{\delta} X_n \to 0$.

Proof. Define $Y_n = (n-1)^{\delta} X_n$. Since

$$(n(n-1)^{-1})^{\delta} = 1 + \delta\, n^{-1} + 0(n^{-2})$$

it follows from A1, A3, A5, and N2 that

$$\tilde{Y}_{n+1} = \tilde{Y}_n - n^{-1}(\mathrm{diag}(\lambda_i - \delta) + B_n)\tilde{Y}_n - n^{-1+\delta}\,(n^{\tau}\,\tilde{e}_n + n^{-2\tau}\,\tilde{\beta}_n)$$

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

$$\sum_{n=1}^{\infty} n^{-1+\delta-2\tau}\,\tilde{\beta}_n \text{ converges.}$$

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

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

Therefore, if we set $A_i = \lambda_i - \delta$,

$$\tilde{Y}_{n+1} = \tilde{Y}_n - n^{-1}\,(\mathrm{diag}(A_i) + B_n)\tilde{Y}_n + d_n \tag{3.1}$$

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. To complete the proof we need only show that

$$\limsup\, ||\tilde{Y}_n|| < \infty\,. \tag{3.2}$$

Choose $\epsilon > 0$, and then choose N such that

$$\sup_{n\geq 1} \left|\left|\sum_{k=N}^{N+n} d_k\right|\right| < \epsilon \quad \text{and} \tag{3.3}$$

$$\sup\{|B_{N+n}^{(ij)}|: n \geq 1 \text{ and } i,j = 1,\ldots,k\} < A_1/(2k) \ . \tag{3.4}$$

If (3.2) does not hold then there exists j, M, n_1, and n_2 such that $N < n_1 < n_2$,

$$M > \epsilon \ ,$$

$$|\tilde{Y}_{n_1}^{(j)}| < M, \quad |\tilde{Y}_{n_2}^{(j)}| > 2M, \quad \text{and} \tag{3.5}$$

$$||\tilde{Y}_i||_\infty \leq 2M \text{ for } n_1 < i < n_2 \ . \tag{3.6}$$

We treat the case where $\tilde{Y}_{n_2}^{(j)} > 0$; the other case is similar. Let $m = \sup\{i < n_2: \tilde{Y}_i^{(j)} < M\}$. Then $m \geq n_1$. By (3.1), (3.3), (3.5), and (3.6),

$$\tilde{Y}_{n_2}^{(j)} = \tilde{Y}_{m+1}^{(j)} - \sum_{i=m+1}^{n_2-1} i^{-1}(\tilde{Y}_i^{(j)} A_j + \sum_{\ell=1}^{k} B_i^{(j\ell)} \tilde{Y}_i^{(\ell)}) + \sum_{i=m}^{n_2-1} d_i^{(j)}$$

$$\leq M - \sum_{i=m}^{n_2-1} i^{-1}(MA_j - 2M \sum_{\ell=1}^{k} |B_i^{(j\ell)}|) + \epsilon.$$

Therefore by (3.4)

$$\tilde{Y}_{n_2}^{(j)} \leq M + \epsilon - \sum_{i=n}^{n_2-1} i^{-1} M(A_j - A_1) \leq M + \epsilon < 2M$$

which contradicts (3.5) since $\tilde{Y}_{n_2}^{(j)} > 0$. □

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

<u>Theorem 3.1</u>. Assume A1 to A6 and $\lambda_1 > \eta$. Then there exists $\epsilon > 0$ such that for $i = 1,\ldots,k$

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

<u>Proof</u>. By A1, A3, and N2

$$\tilde{X}_{n+1} = \tilde{X}_n - n^{-1}(\mathrm{diag}(\lambda_i) + b_n) \tilde{X}_n + n^{-1}(n^\tau \tilde{e}_n + n^{-2\tau} \tilde{\beta}_n)$$

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

$$||b_n|| = O(n^{-\eta\delta}) \text{ for all } \delta < \gamma . \tag{3.7}$$

Fix $i \leq k$ and let $V_n = \tilde{X}_n^{(i)}$. Since $n^{\lambda_i} = (n-1)^{\lambda_i} + \lambda_i n^{\lambda_i - 1} + \nu_n$ where $\nu_n = O(n^{\lambda_i - 2})$,

$$n^{\lambda_i} V_{n+1} = (n-1)^{\lambda_i} V_n + \nu_n V_n - n^{-1+\lambda_i} b_n^{(i)} V_n + n^{-1+\lambda_i} (n^{\tau}\tilde{e}_n^{(i)} + n^{-2\tau} \beta_n^{(i)}). \tag{3.8}$$

Iterating (3.8) yields

$$n^{\lambda_i} V_{n+1} = V_2 + \sum_{j=2}^{n} (\nu_j - b_j^{(i)} j^{-1+\lambda_i}) V_n + \sum_{j=2}^{n} j^{-1+\lambda_i+\tau} \tilde{e}_j^{(i)} + \sum_{j=2}^{n} j^{-1+\lambda_i-2\tau} \tilde{\beta}_j^{(i)} . \tag{3.9}$$

Since $\lambda_i > \gamma$, it follows from (3.7), lemma 3.1, and $\nu_n = O(n^{\lambda_i - 2})$ that we can choose α satisfying $0 < \alpha < 2(\lambda_1 - \gamma)$ and

$$n^{-\lambda_i+\gamma} (\nu_n - n^{-1+\lambda_i} b_n^{(i)}) V_n = O(n^{-1-\alpha}) .$$

Therefore,

$$\sum_{n=1}^{\infty} |(\nu_n - n^{-1+\lambda_i} b_n^{(i)}) V_n n^{-\lambda_i+\gamma+\alpha/2}| < \infty .$$

Since $-\lambda_i + \gamma + \gamma/2 < 0$, Kronecker's lemma implies that

$$n^{-\lambda_i+\gamma} \sum_{j=1}^{n} |(\nu_j - j^{-1+\lambda_i} b_j^{(i)}) V_j| = O(n^{-\alpha/2}) . \tag{3.10}$$

Using A6 and an argument similar to the one which established (3.10), we obtain

$$n^{-\lambda_i+\gamma} \sum_{j=2}^{n} j^{-1+\lambda_i-\gamma} (\tilde{\beta}_j^{(i)} - \tilde{\beta}^{(i)}) = O(n^{-\rho/2}) .$$

Furthermore, it is easy to show that

$$n^{-\lambda_i+\gamma} \sum_{j=2}^{n} j^{-1+\lambda_i-\gamma} = (\lambda_i-\gamma)^{-1} + O(n^{-\lambda_i+\gamma}) .$$

By N1, if $\gamma > 2\tau$, then with probability one $\tilde{\beta}_j^{(i)} = 0$ for all j sufficiently large. Thus,

$$(3.11) \qquad n^{-\lambda_i+\gamma} \sum_{j=1}^{n} j^{-1+\lambda_i-2\tau} \tilde{\beta}_j^{(i)} = \tilde{\beta}^{(i)} \delta(2\tau,\gamma) + O(n^{-\lambda_i+\gamma}) .$$

By (3.9) to (3.11), for some $\epsilon > 0$,

$$(3.12) \qquad n^{\gamma} V_{n+1} = n^{\gamma-\lambda_i} \sum_{j=1}^{n} j^{-1+\lambda_i+\tau} \tilde{e}_j^{(i)} + \tilde{\beta}^{(i)} \delta(2\tau,\gamma) + O(n^{-\epsilon}).$$

If $\gamma \neq \frac{1}{2}-\tau$, then by N1, $\gamma = \frac{1}{2} - \tau - \Delta$ for $\Delta > 0$. Thus by A4

$$\sum_{j=1}^{\infty} (j^{-1+\lambda_i+\tau})(j^{-\lambda_i+\gamma+\Delta/2}) \tilde{e}_j^{(i)} \text{ converges.}$$

Therefore, by Kronecker's lemma,

$$n^{-\lambda_i+\gamma+\Delta/2} \sum_{j=1}^{n} j^{-1+\lambda_i+\tau} \tilde{e}_j^{(i)} = o(1).$$

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

$$(3.13) \qquad n^{\gamma-\lambda_i} \sum_{j=1}^{n} j^{-1+\lambda_i+\tau} \tilde{e}_j^{(i)} =$$

$$n^{\frac{1}{2}-\tau-\lambda_i} \sum_{j=1}^{n} j^{-1+\lambda_i+\tau} \tilde{e}_j^{(i)} \delta(\gamma, \tfrac{1}{2}-\tau) + O(n^{-\epsilon}).$$

Substituting (3.13) into (3.12) completes the proof. □

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 Phillip 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 $\epsilon > 0$,

$$\sum_{k\leq t} e_k = B(t) + O(t^{\frac{1}{2}-\epsilon}) . \tag{4.1}$$

For example, suppose e_k is a strictly stationary ϕ-mixing process such that $\sum_{n=1}^{\infty}(\phi(n))^{\frac{1}{2}} < \infty$. (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, (4.1) holds. We will be interested in the asymptotic behavior of $\sum_{k\leq t} k^{\alpha}e_k$ where e_k satisfies (4.1), so the following lemma is useful.

<u>Lemma 4.1</u>. Let e_n be a sequence of random variables. For any number α, define S_α on $[0,\infty)$ by

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

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

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

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

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

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

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

Proof. If $\alpha > -\frac{1}{2}$, define $N(k,\alpha) = \sum_{j=1}^{k} j^{2\alpha}$ and then define B_α by $B_\alpha(0) = 0$ and for $N(k-1,\alpha) < t \leq N(k,\alpha)$, $k = 1,2,\ldots$, let

$$(4.4) \qquad B_\alpha(t) = B_\alpha(N(k,\alpha)) + k^\alpha \left(B_0\left(\left(\frac{t-N(k-1,\alpha)}{k^{2\alpha}}\right) - (k-1)\right) - B_0(k-1)\right) .$$

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

Let $\alpha < 0$ and let n be a positive integer. Define $b_k^n = k^\alpha - (k+1)^\alpha$ for $k = 1,\ldots,n-1$ and $b_n^n = n^\alpha$. Then by (4.3)

$$(4.5) \qquad S_\alpha(n) = \sum_{k=1}^{n} \left(\sum_{j=k}^{n} b_j^n\right) e_k = \sum_{j=1}^{n} b_j^n \left(\sum_{k=1}^{j} e_k\right)$$

$$= \sum_{j=1}^{n} b_j^n B_0(j) + \sum_{j=1}^{n} b_j^n j^{\frac{1}{2}-\epsilon}, \text{ and}$$

$$(4.6) \qquad \sum_{j=1}^{n} b_j^n B_0(j) = \sum_{k=1}^{n} \left(\sum_{j=k}^{n} b_j^n\right)(B_0(k) - B_0(k-1))$$

$$= \sum_{k=1}^{n} k^\alpha (B_0(k) - B_0(k-1)).$$

Since $k^\alpha - (k+1)^\alpha = O(k^{\alpha-1})$,

$$(4.7) \qquad \lim_{n\to\infty} \sum_{j=1}^{n} b_j^n j^{\frac{1}{2}-\epsilon} \text{ exists and is finite if } \alpha < -\tfrac{1}{2} .$$

Since $(B_0(k) - B_0(k-1))$ is an independent sequence of standard normal random variables, (4.6) implies that

$$(4.8) \qquad \lim_{n\to\infty} \sum_{j=1}^{n} b_j^n B_0(j) \text{ exists and is finite, if } \alpha < -\tfrac{1}{2} .$$

By (4.5), (4.7), and (4.8), if $\alpha < -\frac{1}{2}$ then (4.2) holds. If $\alpha > -\frac{1}{2}$, then

$$(4.9) \qquad \sum_{j=1}^{n} j^{\frac{1}{2}-\epsilon} b_j^n = O\left(\sum_{j=1}^{n-1} j^{-\frac{1}{2}-\alpha-\epsilon}\right) + n^{\frac{1}{2}-\epsilon-\alpha} = O(n^{\alpha+\frac{1}{2}-\zeta}) \text{ for some } \zeta > 0 .$$

For $-\frac{1}{2} < \alpha < 0$, it follows from (4.5), (4.6), and (4.9) that

$$(4.10) \qquad S_\alpha(n) = B_\alpha(N(n,\alpha)) + O(n^{\alpha+\frac{1}{2}-\zeta}), \quad n = 1,2,\ldots .$$

Now it will be shown that (4.10) holds for $\alpha \geq 0$. Define $c_1 = 1$ and $c_k = k^\alpha - (k-1)^\alpha$ for $k \geq 2$. Then for $\alpha \geq 0$

$$S_\alpha(n) = \sum_{k=1}^{n} (\sum_{j=1}^{k} c_j)e_k = \sum_{j=1}^{n} c_j(\sum_{k=j}^{n} e_k)$$

$$= \sum_{j=1}^{n} c_j(B_0(n) - B_0(j-1) + O(n^{\frac{1}{2}-\epsilon})),$$

$$\sum_{j=1}^{n} c_j(B_0(n) - B_0(j-1)) = \sum_{j=1}^{n} c_j (\sum_{k=j}^{n} B_0(k) - B_0(k-1))$$

$$= \sum_{k=1}^{n} k^\alpha(B_0(k) - B_0(k-1)), \text{ and}$$

$$\sum_{j=1}^{n} c_j\, n^{\frac{1}{2}-\epsilon} = O(n^{\frac{1}{2}+\alpha-\epsilon}) .$$

Therefore (4.10) holds for $\alpha > 0$. For $\alpha = 0$, (4.10) holds by assumption.

One can easily show that for $\alpha > -\frac{1}{2}$, there exists $\Delta > 0$ such that

$$\sum_{j \leq t} j^{2\alpha} = (2\alpha + 1)^{-1}\, t^{2\alpha+1} + O(t^{2\alpha+1-\Delta}) . \tag{4.11}$$

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 $\mu > 0$

$$B(t + O(t^{1-\delta})) = B(t) + O(t^{\frac{1}{2}-\delta/2+\mu}) . \tag{4.12}$$

By (4.10) to (4.12), (4.3) holds for $\alpha > -\frac{1}{2}$. □

<u>Lemma 4.2</u>. Suppose a_k are real numbers and $\sum_{k=1}^{\infty} a_k$ converges. Let b_k^n, $k = 1,\ldots,n$ and $n \geq 1$, be positive numbers. Suppose $\sup_n b_n^n \leq M < \infty$, and for each n suppose $b_k^n \leq b_{k+1}^n$. Assume that for each k, b_k^n decreases to 0 as $n \to \infty$. Then $\lim_{n\to\infty} \sum_{k=1}^{n} a_k b_k^n = 0$.

<u>Proof</u>. Fix $\epsilon > 0$. Choose N such that if $n \geq m \geq N$, then $|\sum_{k=m}^{n} a_k| \leq \epsilon$. Then choose $N' \geq N$ such that $b_k^n \leq \epsilon$ for $k = 1,\ldots,N$ and $n \geq N'$. For $k \leq n$, define $c_k^n = b_k^n - b_{k-1}^n$ if $k \geq 2$ and $c_1^n = b_1^n$. Then if $n \geq N$,

$$|\sum_{k=N}^{n} a_k\, b_k^n| = |\sum_{k=N}^{n} a_k(\sum_{j=1}^{k} c_j^n)|$$

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

Therefore if $n \geq N'$,

$$|\sum_{k=1}^{n} a_k\, b_k^n| \leq |\sum_{k=1}^{N} a_k\, b_k^n| + |\sum_{k=N+1}^{n} a_k\, b_k^n| \leq (\sum_{k=1}^{N} |a_k|)\epsilon + M\epsilon \ . \qquad \square$$

Now we treat the KW process, and for this we need several assumptions.

A7. M is a function from R^1 to R^1. The number θ is the unique solution of

$$\inf_x M(x) = M(\theta) \quad \text{and} \quad M'(\theta) = 0 \ .$$

Hencefore we assume also that $\theta = 0$.

A8. M has two continuous derivatives, M''' exists in a neighborhood of 0, $d > 0$, $M'''(x) = M'''(0) + O(|X|^d)$ as $x \to 0$, $\sup_x |M''(x)| < \infty$, $\{x\colon |M'(x)| < \delta\}$ is compact for some $\delta > 0$, and $\{x\colon M(x) \leq C\}$ is compact for all $C > M(0)$.

A9. Let a and c be positive. Suppose $aM''(0) > 1/3$.

A10. Let u_n be a random variable. Define $S(t) = \sum_{k \leq t} u_k$. Suppose $\sigma^2 > 0$, $B(t)$ is a standard Brownian motion on $[0,\infty)$, and for some $\delta > 0$

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

A11. Let X_1 be a random variable and define X_n recursively by

$$X_{n+1} = X_n - a\, n^{-1}\, Y_n \quad \text{where}$$

$$Y_n = (M(X_n + cn^{-1/6}) - M(X_n - cn^{-1/6}) + u_n)/(2\, cn^{-1/6}) \ .$$

Theorem 4.1. Suppose A7 to A11 hold. Let $A = aM''(0) - (5/6)$ and $B = a\,c^2\,M'''(0)/(6a\,M''(0) - 2)$. Then for some $\epsilon > 0$,

$$n^{1/3}\,X_n = B + n^{-\frac{1}{2}} \sum_{k=1}^{n} (k/n)^A\, e_k + O(n^{-\epsilon}) \;. \tag{4.13}$$

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

$$X(t) = B + (a\sigma)(2c)^{-1}\,(2A+1)^{-\frac{1}{2}}\,t^{-A-\frac{1}{2}}\,Z(t^{2A+1}) + O(t^{-\epsilon}) \;. \tag{4.14}$$

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}, $-u_n$, $M'(X_n) - (M(X_n+cn^{-1/6}) - M(X_n-cn^{-1/6}))$, $a/(2cn^{5/6})$ 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). We now show that A1, A2, and A3 hold with $f(x) = a\,M'(x)$, $D = \lambda_1 = a\,M''(0)$, $\beta = (a\,c^2/6)\,M'''(0)$, $e_n = a\,u_n/(2c)$ and $\tau = 1/6$ (so that $\gamma = 1/3 = 2\tau = 1/2 - \tau$). By Taylor expansions,

$$M(X_n + cn^{-1/6}) - M(X_n - c_n^{-1/6})/2\,cn^{-1/6} = M'(X_n) + 1/12\,c^2\,n^{-1/3}\,(M'''(\eta_n) + M'''(\rho_n))$$

where $|\eta_n| < |X_n|$ and $|\rho_n| < |X_n|$. Therefore if we set $\beta_n = (ac^2/12)(M'''(\eta_n) + M'''(\rho_n))$,

$$a\,Y_n = f(X_n) + n^{-2\tau}\,\beta_n + n^{\tau}\,e_n$$

and A1 holds. A2 follows from A8. A3 holds because A8 implies that $f(x) = Dx + O(x^2)$ as $x \to 0$. Thus we have shown that A1 to A5 hold. By lemma 3.1, $X_n = O(n^{-\epsilon})$ for an $\epsilon > 0$; this and A8 imply that $\beta_n = \beta + O(|X_n|^d) = \beta + O(n^{-\epsilon d})$ and so A6 holds. We now invoke theorem 3.1 to prove that (4.13)

holds. By A.10

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

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

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

This and (4.13) imply (4.14) . □

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

Corollary 4.1. Suppose A7 to A11 hold. Then

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

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.

A13. Assume f is a function from R^1 to R^1 and the 0 is the unique solution of

$$f(x) = 0.$$

A14. Let u_n be a sequence of random variables. Suppose $B(t)$ is a standard Brownian on $[0,\infty)$, $\epsilon > 0$, $\sigma > 0$, and

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

A15. $X_{n+1} = X_n - a\ n^{-1}\ (f(X_n) + u_n)$.

A16. f has a continuous derivative and $f(x) = f'(0)x + 0(|x|^{1+\epsilon})$ as $x \to 0$ for $\epsilon > 0$.

A17. Define $V(x) = \int_0^x f(y)dy$. Suppose $\{x: V(x) \leq C\}$ is compact for all $C < \sup V(x)$.

A18. Suppose $\sup|f'(x)| < \infty$ and $\{x: |f(x)| \leq \delta\}$ is compact for some $\delta > 0$.

Theorem 4.2. Suppose A13 to A19 hold and $a f'(0) > \frac{1}{2}$. Define $D = ah'(0)-1$. Then for some $\epsilon > 0$.

$$n^{\frac{1}{2}} X_n = n^{-\frac{1}{2}} a \sum_{k=1}^{n} (k/n)^D u_k + 0(n^{-\epsilon}) .$$

Define $X(t)$, $t \geq 0$ by $X(t) = n^{\frac{1}{2}} X_n$ on $n \leq t < n + 1$. Then there exists a standard Brownian $Z(t)$ and an $\epsilon > 0$ such that

$$X(t) = \frac{a\sigma\, t^{-D-\frac{1}{2}}}{(2D + 1)^{\frac{1}{2}}} Z(t^{2D+1}) + 0(t^{-\epsilon}) .$$

Proof. Similar to the proof of theorem 4.1. □

References

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

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

Fabian, V. (1968). On asymptotic normality in stochastic approximation. *Ann. Math. Statist. 39* 1327-1332.

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.

Kersting, Gotz (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 Applic.* *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, Rhode Island.

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, Rhode Island.

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. Wahrscheinlichkeitstheorie und Verw. Gebiete* *39* 135-150.

Wasan, M.T. (1969). Stochastic Approximation. Cambridge Univ. Press, New York.