A NEW MULTIVARIATE ROBBINS-MONRO PROCEDURE

Running Title: Multivariate Robbins-Monro

David Ruppert[1]

AMS Subject Classification: 62L20

KEY WORDS AND PHRASES: Root finding, stochastic approximation.

[1]David Ruppert is Associate Professor, Department of Statistics, University of North Carolina, Chapel Hill, NC 27514. This research was supported by National Science Foundation Grant MCS-8100748.

ABSTRACT

Suppose that f is a function from $\mathbb{R}^k$ to $\mathbb{R}^k$ and for some θ, $f(\theta) = 0$. Initially f is unknown, but for any γ in $\mathbb{R}^k$ we can observe a random vector $Y(x)$ with expectation $f(x)$. The unknown θ can be estimated recursively by Blum's (1954) multivariate version of the Robbins-Monro procedure. Blum's procedure requires the rather restrictive assumption that infimum of the inner product $(x-\theta)^t f(x)$ over any compact set not containing θ be positive. Thus at each x, $f(x)$ gives information about the direction towards θ. Blum's recursion is $X_{n+1} = X_n - a_n Y_n$ where the conditional expectation of Y_n given $X_1,\ldots,X_n$ is $f(X_n)$ and $a_n > 0$. Unlike Blum's method, the procedure introduced in this paper does not attempt to move *directly* towards θ when updating X_n to X_{n+1}. Rather, except for random fluctuations it moves in a direction which decreases $\|f\|^2$, and it may follow a very circuitous route to θ. Consequently, it does not require that $(x-\theta)^t f(x)$ have a constant signum. This new procedure is somewhat similar to the multivariate Kiefer-Wolfowitz procedure applied to $\|f\|^2$, but unlike the latter it converges to θ at rate $n^{-\frac{1}{2}}$.

1. INTRODUCTION

This paper is concerned with a multivariate version of a problem first studied by Robbins and Monro (1951). Suppose that f is an unknown function from $\mathbb{R}^k$ to $\mathbb{R}^k$, and that for any x in $\mathbb{R}^k$ we can observe a random vector Y(x) with expectation f(x). Let α in $\mathbb{R}^k$ be known, and suppose there is a unique θ such that $f(\theta) = \alpha$. The goal is to estimate θ.

For example, suppose that k = 2 and x gives the doses of two drugs that affect blood chemistry. The concentrations of two chemicals in the blood are measured after administration of the drugs, and f(x) gives the expected concentrations as a function of the doses. If α gives the ideal concentrations of the two blood components, then θ gives the correct doses.

By choosing the appropriate measurement scales, we can without loss of generality assume that $\alpha = 0$.

Blum's (1954) version of the Robbins-Monro (RM) process begins with an initial estimate X_1 of θ. Given $X_1,\ldots,X_n$, one observes Y_n, such that $E^{\mathcal{F}_n}Y_n = f(X_n)$ where $\mathcal{F}_n$ is the σ-algebra generated by $X_1,\ldots,X_n$. Then X_n is updated by the recursion

$$X_{n+1} = X_n - a_nY_n \tag{1.1}$$

where a_n is a suitably chosen positive sequence converging to 0. Convergence of X_n to θ is proved under the assumption that for each $\varepsilon > 0$

$$\inf\{(x-\theta)^t f(x) : \varepsilon < \|x\| < \varepsilon^{-1}\} > 0 \tag{1.2}$$

The importance of (1.2) can be easily seen as follows. From (1.1) and the fact that $E^{\mathcal{F}_n}Y_n = f(X_n)$, we have

$$E^{\mathcal{F}_n}\| X_{n+1}-\theta\|^2 = \| X_n-\theta\|^2 - 2a_n(X_n-\theta)^t f(X_n) + a_n^2 E^{\mathcal{F}_n}\| Y_n\|^2$$

If we could ignore the term of order a_n^2, then $\| X_n-\theta\|^2$ would be a positive supermartingale and would converge a.s. The term of order a_n^2 can be handled using a theorem on "almost" positive supermartingales (Robbins and Siegmund, 1971), which also can be used to show that

$$\sum_{n=1}^{\infty} a_n(X_n-\theta)^t f(X_n) \quad \text{converges a.s.} \tag{1.3}$$

The sequence $\{a_n\}$ is chosen so that

$$\sum_{n=1}^{\infty} a_n = \infty \ , \tag{1.4}$$

and (1.2)-(1.4) imply that $X_n \to 0$ a.s. Authors using condition (1.2) or something quite similar include Sacks (1958, assumption (A1*)), Schmetterer (1968, assumption (4.15)), and Walk (1977, assumption (2a)). (Walk's paper concerns the RM process in a general Hilbert spaces.) Blum's original work uses assumptions stronger than (1.2).

Unfortunately, (1.2) is a rather restrictive assumption, implying that at each x, f(x) "points away from θ". Clearly we can replace (1.2) by

$$\inf\{-(x-\theta)^t f(x): \varepsilon < \| x\| < \varepsilon^{-1}\} > 0; \tag{1.2'}$$

this requires only that we change (1.1) to $X_{n+1} = X_n + a_n Y_n$. An example of a function satisfying neither (1.2) nor (1.2') is $f(x_1,x_2) = (x_1^2,x_2^2)^t$.

An alternative to the multivariate RM procedure would be to apply the multivariate Kiefer-Wolfowitz (KW) procedure to minimize $\| f(x)\|^2$. (For the moment assume that the conditional variance of Y(x) is independent of x, so that $E^{\mathcal{F}_n}\| Y_n\|^2 = \| f(X_n)\|^2 +$ constant. Otherwise, the KW procedure will

find the minimizer of $E\| Y(x)\|^2$, not necessarily the solution to $f(x) = 0$.) The KW procedure will tend to follow the negative gradient of $\| f(x)\|^2$. Although it may not move from X_n *directly* towards θ, except for random fluctuations, it does move downhill. Therefore, under mild conditions X_n will converge to a local minimum of $\| f(x)\|^2$. If θ is the only local minimum, then $X_n \to \theta$.

Unfortunately, the rate of convergence to θ of the KW method is slower than the $n^{-\frac{1}{2}}$ rate of RM method, though modifications of the Kiefer-Wolfowitz method can produce rates arbitrarily close to $n^{-\frac{1}{2}}$ if f has derivatives of sufficiently high order (Fabian, 1971).

In this paper, we propose a new multivariate RM process which in some ways behaves as the Kiefer-Wolfowitz method applied to $\| f\|^2$, but which possesses the $n^{-\frac{1}{2}}$ rate of convergence even when f has only two derivatives.

Let $D(x)$ be the $k \times k$ derivative of $f(x)$. Then, the derivative of $\| f(x)\|^2$ is

$$2\, D^t(x)f(x).$$

The Hessian matrix of $\| f(x)\|^2$ is

$$H(x) = 2[D^t(x)D(x) + \sum_{i=1}^{k} H^{(i)}(x)f^i(x)] ,$$

where f^i is ith coordinate of f and $H^{(i)}$ is the Hessian of f^i. Thus $H(\theta) = 2\, D^t(\theta)D(\theta)$. The recursive procedure that is introduced here is

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

where $a > 0$, B_n is an estimate of $[D^t(\theta)D(\theta)]^{-1}$, and D_n is an estimate of $f(X_n)$. Blum's multivariate RM procedure uses one observation to construct f_n and thereby to update X_n to X_{n+1}. Our procedure uses $(2k)(m_n)$ observations to construct D_n and $[n^\gamma]$ observations to construct f_n, where $[\cdot]$ is the greatest

integer function, $\gamma > 0$, $m_n \to \infty$, and $m_n n^{-\gamma} \to 0$. We let $m_n \to \infty$ sufficiently fast (see below), so that the conditional variance of D_n, given $X_1,\ldots,X_n$, converges to 0. We require that $m_n n^{-\gamma} \to 0$ so that among the totality of observations used in constructing both D_n and f_n, the proportion used in estimating D converges to 0. These properties insure full asymptotic efficiency (see section 5).

The estimator f_n is simply the mean of $[n^\gamma]$ observations with conditional expectation, given $X_1,\ldots,X_n$, equal to $f(X_n)$. The ith column of D_n is constructed as follows. Let e(i) be the ith column of the $k \times k$ identity matrix. Let Y(n,i,w) and Y(n,i,1) each be the mean of m_n observations with conditional expectation equal to $f(X_n + c_n e(i))$ and $f(X_n - c_n e(i))$, respectively. Then, the ith column of D_n is

$$D_n^{(i)} = [Y(n,i,2) - Y(n,i,1)]/(2c_n) .$$

We choose m_n and c_n so that $c_n \to 0$ and $m_n^{-1} c_n^{-2} \to 0$. With this choice of c_n and m_n, the bias of D_n as an estimate of $D(X_n)$ converges to 0, and as mentioned above the conditional variance of D_n also converges to 0.

B_n is constructed as follows. Let $\underline{\eta}_n$ and $\overline{\eta}_n$ be positive sequences such that $\underline{\eta}_n \downarrow 0$, $\overline{\eta}_n \uparrow \infty$, and certain other conditions (see section 2) are met.

Let

$$C_n = (n-1)^{-1} \sum_{i=1}^{n-1} D_i^t D_i .$$

Let $B_n = C_n^{-1}$ if all eigenvalues of C_n^{-1} lie between $\underline{\eta}_n$ and $\overline{\eta}_n$. Otherwise let B_n be some symmetric matrix whose eigenvalues are all between $\underline{\eta}_n$ and $\overline{\eta}_n$.

The procedure of this paper bears some resemblance to the one dimensional Venter (1967) procedure. It was shown by Chung (1954) that the univariate Robbins-Monro procedure is asymptotically optimal when $a_n = 1/(nf'(\theta))$. Of course, $f'(\theta)$ will typically be unknown. Venter introduced a consistent estimate b_n of $f'(\theta)$ and showed that symptotic optimality could be achieved

with an $a_n = 1/(n\, b_n)$.

Our procedure also estimates f' but at each X_n, not simply at θ. Moreover, the Venter process and the original RM process are consistent under roughly the same circumstances. Our procedure is an attempt to improve the consistency properties of Blum's multivariate RM process. The matrix sequence B_n does, however, play a role for our process which is analogous to that of b_n in the Venter process.

It should be mentioned that Blum's version of the RM process may be preferable to the one introduced here under certain circumstances, namely when (1.2) is known to hold or (1.2') is known to hold and $D^t(x)f(x) = 0$ for some $x \neq \theta$. However, when neither (1.2) nor (1.2') hold, our procedure, but not necessarily Blum's, at least tends to move in a direction decreasing $\| f(x)\|^2$.

2. NOTATION AND ASSUMPTIONS

Let $\mathbb{R}^k$ be k dimensional Euclidean space. If A is a $k\times\ell$ matrix, let A^{ij} be the i, jth entry of A, and if $\ell=1$ let $A^i = A^{i1}$. Also, let A^t be the transpose of A, and let $\| A\|^2 = \sum_{i=1}^{k} \sum_{j=1}^{\ell} (A^{ij})^2$. $A^{-t} = (A^{-1})^t$.

All random variables are defined on the same probability space, and all relations between random variables are meant to hold with probability one. If $z_1,\ldots z_n$ are random variables, then $\sigma(z_1,\ldots,z_n)$ is the σ-algebra that they generate. If $\mathcal{F}$ is a σ-algebra, the $E^{\mathcal{F}}X$ and $\mathrm{Var}^{\mathcal{F}}X$ are, respectively, the conditional mean and variance matrix of the random vector X. If X is a random matrix, then $\mathrm{Var}^{\mathcal{F}_n}X$ is conditional variance matrix of X arranged as a column vector. $\xrightarrow{L}$ denotes convergence in law.

We say that $a_n \sim b_n$ if $a_n/b_n \to 1$. "O" and "o" notation has its usual meaning.

The following assumptions on f will be needed.

F1. (i) $f = (f_1,\ldots,f_k)^t$ is a twice differentiable function from $\mathbb{R}^k$

to $\mathbb{R}^k$. θ is in $\mathbb{R}^k$. $f(\theta) = 0$.

(ii) $D(X)$ is the derivative of f, i.e., $D^{ij}(x) = (\partial/\partial x_j)f^i(x)$.

(iii) $D(\theta)$ is nonsingular.

(iv) For all $\varepsilon > 0$

$$\inf\{\| D^t(x)f(x)\| : \varepsilon \le \| f(x)\| \le \varepsilon^{-1}\} > 0.$$

(v) $\sup\{\| D(x)\| \} < \infty$.

(vi) Let $H(x)$ be the Hessian of $\| f(x)\|^2$, i.e., $H^{ij}(x) = (\partial^2/\partial x_i \partial x_j)\| f(x)\|^2$.

Then,

F2. For all $\varepsilon > 0$

$$\inf\{\| f(x)\| : \varepsilon \le \|x-\theta\| \le \varepsilon^{-1}\} > 0.$$

The modified Robbins-Monro algorithm will be described formally by the following assumptions.

A1. (i) $X_{n+1} = X_n - an^{-1}B_nD_nf_n$ where $a > 0$, X_n is in $\mathbb{R}^k$ and B_n and D_n are $k \times k$ random matrices.

(ii) $c_n > 0$, $c_n \downarrow 0$, m_n is an integer, $m_nc_n^2 \uparrow \infty$, $\gamma > 0$, $m_nn^{-\gamma} \to 0$, $\underline{\eta}_n \downarrow 0$, $\bar{\eta}_n^2n^{-2} = o(\underline{\eta}_nn^{-1})$, $\bar{\eta}_n \uparrow \infty$,

$$\sum n^{-1}\underline{\eta}_n = \infty \quad , \tag{2.1}$$

and

$$\sum n^{-1}\bar{\eta}_n[c_n^2 + n^{-1}\bar{\eta}_n(c_n^2 + c_n^{-2}m_n^{-1} + n^{-\gamma})] < \infty \tag{2.2}$$

(iii) Let $\mathcal{F}_n = \sigma(X_1,\ldots,X_n)$. Then B_n is $\mathcal{F}_{n-1}$ measurable, $E^{\mathcal{F}_n}f_n = f(X_n)$, $E^{\mathcal{F}_n}D_n = D(X_n) + O(c_n^2)$, $\| \mathrm{Var}^{\mathcal{F}_n}f_n\| = O(n^{-\gamma})$, and $\| \mathrm{Var}^{\mathcal{F}_n}D_n\| = O(m_n^{-1}c_n^{-2})$. Given $\mathcal{F}_n$, f_n and D_n are conditionally independent.

(iv) B_n is symmetric and all eigenvalues of B_n are between $\underline{\eta}_n$ and $\bar{\eta}_n$.

A2. (i) $a > (1+\gamma)/2$

(ii) If $X_n \to \theta$, then $B_n \to (D^t(\theta)D(\theta))^{-1}$ and $n^\gamma \operatorname{Var}^{F_n} f_n \to S$ for some matrix S, and for all $r > 0$

$$\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} E\, I\{\| V_i \|^2 \geq r\, i\} \| V_i \|^2 = 0,$$

where $V_n = n^\gamma (\bar{D}_n f(X_n) - D_n^t f_n)$, and $\bar{D}_n = E^{F_n} D_n$.

Remarks on the assumptions. F1 (vi) corresponds to the assumption of a bounded Hessian which has been used in the study of the Kiefer-Wolfowitz algorithm, e.g., Fabian (1971, assumption (2.2)). Any substantially weaker condition would probably require that the step sizes, $a_n B_n D_n f_n$, in A1 (i) be modified to prevent increasingly large oscillations. Otherwise, X_n might have no finite limit points.

Assumptions F1 (iv) and F2 are also similar to conditions that would be needed if the KW process were applied to $\| f(x) \|^2$. See Fabian (1971, assumption (2.2) equation (1)). They imply that θ is the only local minimum of $\| f(x) \|^2$.

Given the method described in the introduction of constructing B_n, D_n, and f_n, assumptions A1 (iii), A1 (iv) and A2 (ii) are natural. To have $E^{F_n} D_n = D(X_n) + O(c_n^2)$, it is sufficient that the second derivative of f be uniformly bounded. Simple conditions sufficient for A2 (ii) can be found using standard martingale techniques.

3. THEOREMS

Theorem 3.1. Assume F1 and A1. Then $f(X_n) \to 0$. If F2 also holds, then $X_n \to \theta$.

Theorem 3.2. Assume F1, F2, A1, and A2. Then

$$n^{(1+\gamma)/2}(X_n - \theta) \xrightarrow{L} N(0, [a^2/(2a-1-\gamma)] D^{-1}(\theta) S D^{-t}(\theta))$$

Corollary 3.3. Under F1, F2, A1, and A2, the choice $a = (1+\gamma)$ is asymptotically optimal. With this choice,

$$N_n^{\frac{1}{2}}(X_n-\theta) \xrightarrow{L} N(0,D^{-1}(\theta)S\,D^{-t}(\theta))$$

where N_n is the total number of observations needed to construct X_n.

4. PROOFS

Proof of theorem 3.1. Define $\bar{D}_n = E^{F_n}D_n$, $d_n = D_n-\bar{D}_n$, and $\varepsilon_n = f_n-f(X_n)$. By F1 (ii), A1 (i), and A1 (iii) there is a ζ in (0,1) such that

$$E^{F_n}\| f(X_{n+1})\|^2 \le \| f(X_n)\|^2 - 2an^{-1}f^t(X_n)D(X_n)B_nD^t(X_n)f(X_n)$$

$$-2an^{-1}f^t(X_n)[\bar{D}_n-D(X_n)]B_nD^t(X_n)f(X_n) \tag{4.1}$$

$$+ a^2n^{-2}E^{F_n}[(B_nD_n^tf_n)^tH(X_n-\zeta an^{-1}B_nD_n^tf_n)(B_nD_n^tf_n)] .$$

By F1 (v) and A1 (iii)

$$|f^t(X_n)[\bar{D}_n-D(X_n)]B_nD^t(X_n)f(X_n)| \tag{4.2}$$

$$= O(\| f(X_n)\|^2\lambda_nc_n^2),$$

where λ_n is the largest eigenvalue of B_n. By F1 (vi)

$$E^{F_n}(B_nD_n^tf_n)^tH(X_n-\theta an^{-1}B_nD_n^tf_n)(B_nD_n^tf_n) \tag{4.3}$$

$$= O(\lambda_n^2E^{F_n}\| D_n^tf_n\|^2).$$

By A1 (iii) and F1 (v)

$$E^{F_n}\| D_n^tf_n\|^2 = \| D^t(X_n)f(X_n)\|^2 +$$

$$f^t(X_n)[\bar{D}_n\bar{D}_n^t - D(X_n)D^t(X_n)]f(X_n) + f^t(X_n)(E^{F_n}d_nd_n^t)f(X_n)$$

$$+ E^{F_n}[\varepsilon_n^t(\bar{D}_n\bar{D}_n^t + d_nd_n^t)\,\varepsilon_n] \tag{4.4}$$

$$= \| D^t(X_n)f(X_n)\|^2 + O[(c_n^2 + c_n^{-2}m_n^{-1})\| f(X_n)\|^2 + n^{-\gamma}] .$$

By (2.2), (4.1) to (4.4), and A1 (iv)

$$E^{F_n}\| f(X_{n+1})\|^2 \le \| f(X_n)\|^2 [1 + 2an^{-1}\underline{\eta}_n c_n^2 + n^{-2}\bar{\eta}_n^2(c_n^2 + c_n^{-2}m_n^{-1})]$$

$$- (2an^{-1}\underline{\eta}_n - \bar{\eta}_n^2 n^{-2})\| D^t(X_n)f(X_n)\|^2 + \bar{\eta}_n^2 n^{-\gamma-2}$$

$$= \| f(X_n)\|^2 (1+\mu_n)$$

$$- (2an^{-1}\underline{\eta}_n)(1 + o(1))\| D^t(X_n)f(X_n)\|^2 + \nu_n$$

where $\sum\mu_n < \infty$ and $\sum\nu_n < \infty$.

Therefore, by theorem 1 of Robbins and Siegmund (1971), $\lim_{n\to\infty} \| f(X_n)\|$ exists and is finite and

$$\sum_{n=1}^{\infty} n^{-1}\underline{\eta}_n\| D^t(X_n)f(X_n)\|^2 < \infty .$$

Then by F1 (iv) $\| f(X_n)\|^2 \to 0$, and therefore $X_n \to \theta$ if F2 holds.

Proof of theorem 3.2. By theorem 3.1 and A2 (ii), $B_n \to (D^t(\theta)D(\theta))^{-1}$, $\lambda_n \to 1$, and $D(X_n) \to D(\theta)$. For each $\eta > 0$

$$f^t(X_n)D(X_n)B_nD^t(X_n)f(X_n)$$

$$\ge (1-\eta)\| f(X_n)\|^2$$

for all sufficiently large n. Therefore, by (4.1) to (4.4), for each $\eta > 0$

$$E^{F_n}\| f(X_{n+1})\|^2 \le \| f(X_n)\|^2 [1-2a(1-\eta)n^{-1} +$$

$$0(c_n^2 n^{-1} + n^{-2})] + 0(n^{-2-\gamma})$$

$$\le \| f(X_n)\|^2 (1-2a(1-2\eta)n^{-1}) + 0(n^{-2-\gamma})$$

for all n sufficiently large. Note that $(n+1)^{1+\varepsilon} = n^{1+\varepsilon} + n^{\varepsilon}(1+\varepsilon) + 0(n^{\varepsilon-1})$. For each $0 < \varepsilon < \gamma$ and for each $\eta > 0$ and for all large n,

$$(n+1)^{1+\varepsilon}E^{F_n}\| f(X_{n+1})\|^2 \le (1-[2a(1-\eta)-(1+\varepsilon)]n^{-1})n^{1+\varepsilon}\| f(X_n)\|^2$$

$$+ 0(n^{-1-(\gamma-\varepsilon)}) .$$

Therefore, by A2 (i) and another application of theorem 1 of Robbins and Siegmund (1971), $\lim_{n\to\infty} n^{1+\varepsilon}\| f(X_n)\|^2$ exists and is finite for all $\varepsilon < \gamma$, whence

$$n^{1+\varepsilon}\| f(X_n)\|^2 \to 0 \tag{4.5}$$

for all $\varepsilon < \gamma$.

We can now apply theorem 2.2 of Fabian (1968) with $\Gamma_n = aB\bar{D}_n^t f(X_n)$, $F = aI$, $\alpha = 1$, $\beta = (1+\varepsilon)$, $U_n = X_n$, $\Phi_n = aB_n$, $\Phi = aD^{-1}(\theta)D^{-t}(\theta)$, $V_n = n^{\gamma}(\bar{D}_n f(X_n) - D_n^t f_n)$, $T_n = T = 0$, $P = I$, $\Lambda = aI$, and $\Sigma = \lim_{n\to\infty} E^{\mathcal{F}_n}(V_n V_n^t)$. Note that $(\Lambda^{(ii)} + \Lambda^{(jj)} - \beta_+) = (2a-1-\gamma)$ for all i and j. We need to calculate Σ more explicitly.

If V_1, V_2, W_1, and W_2 are random variables possessing finite second moments such that (V_1,V_2) is independent of (W_1,W_2), then

$$\mathrm{Cov}(V_1W_1, V_2W_2) = \mathrm{Cov}(V_1,V_2)\ \mathrm{Cov}(W_1,W_2)$$

$$+ \mathrm{Cov}(V_1,V_2)(EW_1)(EW_2) + \mathrm{Cov}(W_1,W_2)(EV_1)(EV_2).$$

Applying this fact coordinatewise to $D_n^t f_n$ and using A1 (ii) and A1 (iii) one can show that

$$\mathrm{Var}^{\mathcal{F}_n}(D_n^t f_n) = \bar{D}_n^t(\mathrm{Var}^{\mathcal{F}_n} f_n)\bar{D}_n$$

$$+ O(\| f(X_n)\|^2\ c_n^{-2}m_n + n^{-\gamma}c_n^{-2}m_n) \quad .$$

Then by A2 (ii) and (4.5)

$$\Sigma = a^2 \lim_{n\to\infty} n^{\gamma}\mathrm{Var}^{\mathcal{F}_n}(D_n^t f_n)$$

$$= a^2\ D^t(\theta)S\ D(\theta)$$

Proof of corollary 3.3. It is trivial to show that $a^2/(2a-1-\gamma)$ is minimized in a, subject to the constraint that $a > (1+\gamma)/2$, by $a = (1+\gamma)$. The corollary then follows because $N_n = \sum_{i=1}^{n}\{[i^{\gamma}]+m_i\} \sim \sum_{i=1}^{n}[i^{\gamma}] \sim n^{1+\gamma}/(1+\gamma)$.

5. ASYMPTOTIC EFFICIENCY

We will not treat the subject of efficiency in great detail, but we will study a simple example. Suppose D is nonsingular and known, $f(X) = D(X-\theta)$, and Var $Y(x) = S$ for all x. If Y(x) is normally distributed, then $x-D^{-1}Y(x) \sim N(\theta, D^{-1}SD^{-t})$. Thus, if $Z_1,\ldots,Z_{N_n}$ is any sequence of random variables, then the maximum likelihood estimate of θ based on $Y(Z_1),\ldots,Y(Z_{N_n})$ is

$$\hat{\theta} = N_n^{-1} \sum_{i=1}^{N_n} Z_i - D^{-1}Y(Z_i)$$

Also $\hat{\theta} \sim N(\theta, N_n^{-1}D^{-1}\ S\ D^{-t})$. Therefore, $\hat{\theta}$ and our estimator, X_n, have the same asymptotic distributions, even though our estimator has been devised for a much more general problem.

When deriving the asymptotic distribution of X_n, it is crucial that $\text{Var}^{F_n} D_n \to 0$ so that $\text{Var}^{F_n} D_n^t f_n \doteq D^t(\theta)(\text{Var}^{F_n} f_n)D(\theta)$. Then because $\text{Var}^{F_n}(D_n^t f_n)$ is determined by $\text{Var}^{F_n} f_n$, full efficiency is obtained by having $m_n n^{-\gamma} \to 0$, so that the ratio of the number of observations used to construct D_n to the number used to construct f_n converges to 0.

REFERENCES

1. Blum, J.R. (1954). Multidimensional stochastic approximation methods. Ann. Math. Statist. 25, 737-744.

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

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

4. Fabian, V. (1971). Stochastic approximation. In Optimizing Methods in Statistics (J.S. Rustagi, ed.) 439-470. Academic Press: New York.

5. Robbins, H. and Monro, S. (1951). A stochastic approximation method. Ann. Math. Statist. 22, 400-407.

6. Robbins, H. and Siegmund, D. (1971). A convergence theorem for non negative almost supermartingales and some applications. In Optimizing Methods in Statistics. (J.S. Rustagi, ed.) 233-257. Academic Press: New York.

7. Sacks, J. (1958). Asymptotic distributions of stochastic approximation procedures. Ann. Math. Statist. 29, 373-405.

8. Schmetterer, L. (1969). Multidimensional stochastic approximation. (In Multivariate Analysis, II, P.R. Krishnaiah, ed.) Academic Press: New York.

8. Venter, J. (1967). An extension of the Robbins-Monro procedure. Ann. Math. Statist. 38, 181-190.

9. Walk, H. (1977). An invariance principle for the Robbins-Monro Process in a Hilbert Space. Z. Wahrscheinlickkeitstheorie verw. Gebiete, 31, 135-150.