Almost sure convergence of stochastic gradient processes with matrix step sizes

Jean-Marie Monnez

Almost Sure Convergence of Stochastic Gradient Processes with Matrix Step Sizes

by

Jean-Marie Monnez

Université Henri Poincaré

IECN Laboratoire de Mathématiques

B.P. 239 - F 54506 Vandoeuvre lès Nancy Cedex

Abstract

We consider a stochastic gradient process, which is a special case of stochastic approximation process, where the positive real step size a_n is replaced by a random matrix A_n: $X_{n+1} = X_n - A_n \nabla g(X_n) - A_n V_n$. We give two theorems of almost sure convergence in the case where the equation $\nabla g = 0$ has a set of solutions.

Key Words and Phrases: Stochastic approximation, stochastic gradient

AMS 1991 Subject Classifications. Primary: 62**; Secondary:** L20.

1 Introduction

Let g be a function from $\mathbb{R}^p$ into $\mathbb{R}$. Note ∇g the gradient of g. Suppose that the equation $\nabla g = 0$ has a set of solutions. Suppose further that for any real x, one cannot observe $\nabla g(x)$ but only $\nabla g(x) + V_x$, V_x being a random error. In order to estimate a solution, one can use a stochastic gradient process (X_n), defined recursively by

$$X_{n+1} = X_n - a_n \nabla g(X_n) - a_n V_n. \tag{1}$$

The sequence of positive gains (a_n) verifies $\sum_{n=1}^{\infty} a_n = \infty, \sum_{n=1}^{\infty} a_n^2 < \infty$. (1) is a particular case of stochastic approximation process, whose convergence has been studied by different methods; see inter alia Ljung, Pflug, Walk (1992) for an overview.

Several applications can be found particularly in the field of statistics. In some of them, the gain a_n is replaced by a random matrix A_n, for instance to obtain an asymptotically efficient estimation of a parameter or to accelerate the convergence; the process is then defined by

$$X_{n+1} = X_n - A_n \nabla g(X_n) - A_n V_n. \tag{2}$$

It was supposed in most cases for this type of process that the equation $\nabla g = 0$ has a unique solution (see inter alia Nevel'son and Has'minskii, 1973, Spall, 2003). We consider here the case where it has a set of solutions and we give two theorems of almost sure convergence. The first one can be directly applied for instance to the

proof of the convergence of an extension of the k-means process of MacQueen (1967), which was the first motivation for this study.

All random variables are defined on a probability space $(\Omega, \mathcal{A}, P)$. Let (T_n) be a nondecreasing sequence of sub-σ-algebras of $\mathcal{A}$ such that $X_1, V_j, j \leq n-1$ and $A_j, j \leq n$, are T_n-measurable; then X_n is T_n-measurable.

In the following, $\langle ., . \rangle$ and $\|.\|$ are respectively the usual inner product and norm in $\mathbb{R}^p$; for a squared matrix A, $\|A\|$ is the spectral norm, $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ are respectively the greatest and the lowest eigenvalue of A; I is the identity matrix; the abbreviation $a.s.$ means almost surely.

2 First almost sure convergence theorem

2.1 Assumptions and theorem

Denote $F_n(X_n) = \frac{\|A_n^{\frac{1}{2}} \nabla g(X_n)\|^2}{\|A_n\|}$ and suppose:

(H1a) g is a non negative function.

(H1b) There exists $L > 0$ such that, for all $A, B \in \mathbb{R}^p$,

$g(B) - g(A) \leq \langle B - A, \nabla g(A) \rangle + L \parallel B - A \parallel^2 .$

H1b is verified if ∇g is Lipschitz continuous. H1b is used for $B = X_{n+1}$ and $A = X_n$; if the sequence (X_n) is bounded and g has second order partial derivatives which are continuous in the compact set containing it, then H1b can be omitted.

(H1c) There exists a set A such that $P(A) = 1$ and for all fixed $\omega \in A, \forall \epsilon > 0$,

$\exists \delta > 0, \exists N \in \mathbb{N}, \exists K > 0 : \forall p > n > N, (\parallel X_p - X_n \parallel < \delta) \Longrightarrow (F_p(X_p) \leq KF_n(X_n) + \epsilon)$.

This assumption is verified if ∇g is uniformly continuous, particularly if the sequence (X_n) is bounded and ∇g is continuous in the compact set containing it, and if the following assumption H4d holds. The sequence (X_n) is bounded if $g(X_n)$ converges, which is proved without H1c, and if $\lim_{\|x\| \to \infty} g(x) = \infty$ (if g is C^1, then g is a Liapunov function).

(H2) There exist four sequences of random variables $(B_n), (C_n), (D_n)$ and (E_n) in $\mathbb{R}^+$ adapted to the sequence of σ-algebras (T_n) such that $a.s.$

(H2a) $\parallel A_n^{\frac{1}{2}} E[V_n \mid T_n] \parallel^2 \leq B_n g(X_n) + C_n, \sum_1^\infty (B_n + C_n) < \infty$;

(H2b) $E[\parallel A_n V_n \parallel^2 \mid T_n] \leq D_n g(X_n) + E_n, \sum_1^\infty (D_n + E_n) < \infty$.

H2a and H2b are extensions of classical assumptions in the field of stochastic approximation. They are satisfied particularly in the case where $V_n = V_n^1 + V_n^2$, $E[V_n^1 \mid T_n] = 0$ $a.s.$, V_n^2 T_n-measurable and, if the following assumptions H4a, b hold, $\parallel A_n^{\frac{1}{2}} V_n^2 \parallel^2 \leq B_n g(X_n) + C_n$.

(H4a) For $n \geq 1, A_n$ is $a.s.$ a positive definite symmetrical matrix.

(H4b) $\sup_n \lambda_{\max}(A_n) < \min(\frac{1}{2}, \frac{1}{4L})$ $a.s.$

(H4c) $\sum_1^\infty \lambda_{\max}(A_n) = \infty$ $a.s.$

(H4d) $\sup_n \frac{\lambda_{\max}(A_n)}{\lambda_{\min}(A_n)} < \infty$ $a.s.$

In the case where $A_n = a_n I, a_n \in \mathbb{R}$, H4 gives $0 < a_n < \min(\frac{1}{2}, \frac{1}{4L}), \sum_1^\infty a_n = \infty$.

H4 holds in the case where A_n is a diagonal matrix with diagonal elements $a_n^1, ..., a_n^p$ such that $\frac{a}{n} \leq a_n^i < \min(\frac{1}{2}, \frac{1}{4L})$, $i = 1, ..., p$, $a > 0$ and $\sup_n \frac{\max a_n^i}{\min a_n^i} < \infty$ *a.s.*; this can be applied to the study of the convergence of the k-means process of Mac Queen (1967). H4 holds too for $A_n = \frac{1}{n} A(X_n)$ with suitable conditions on the matrices $A(X_n)$.

Theorem 1 *(a) Assume H1a, b, H2a, b and H4a, b hold; then* $g(X_n)$ *and* $\sum_1^\infty \parallel A_n^{\frac{1}{2}} \nabla g(X_n) \parallel^2$ *converge a.s.*

(b) Under the assumptions of (a), if moreover H1c, H4c, d hold, then $F_n(X_n)$ *and* $\nabla g(X_n)$ *converge a.s. to* 0*; if the sequence* (X_n) *is contained in a compact set* C *and* ∇g *is continuous in* C*, the distance of* X_n *to the set of stationary points of* g *converges a.s. to* 0*.*

In the case where $A_n = a_n I$, this theorem can be compared to theorem 1 of Dippon (1998), who supposes yet that ∇g is Lipschitz continuous and $E[V_n \mid T_n] = 0$.

2.2 Lemma

We use the following lemma in the proof of the theorem. Define the recursive process (W_n) in $\mathbb{R}^p$ and make assumption H3a:

$$W_{n+1} = (I - A_n)W_n + A_n V_n, W_1 = 0. \tag{3}$$

If $A_n = \frac{1}{n} I$, then $W_{n+1} = \frac{1}{n} \sum_1^n V_i$.

(H3a) $W_n \longrightarrow 0$ *a.s.*

Lemma 2 *Assume H1c, H3a and H4a, b, c hold; then, if* $\sum_1^\infty \parallel A_n^{\frac{1}{2}} \nabla g(X_n) \parallel^2$ *converges a.s.,* $F_n(X_n)$ *converges a.s.to* 0*; if furthermore H4d holds,* $\nabla g(X_n)$ *converges a.s. to* 0*; if moreover the sequence* (X_n) *is contained in a compact set* C *and* ∇g *is continuous in* C*, the distance of* X_n *to the set of stationary points of* g *converges a.s. to* 0*.*

Proof. $\omega \in \Omega$ is fixed throughout the proof, belonging to the intersection of the defined *a.s.* convergence sets. Let $\epsilon > 0, \epsilon^{'} = \min(\frac{\epsilon}{2K}, \frac{\epsilon}{2})$. By H1c, $\exists \delta > 0, \exists N_1 \in \mathbb{N}, \exists K > 0 : \forall p > n > N_1, (\parallel X_p - X_n \parallel < \delta) \Longrightarrow (F_p(X_p) \leq K F_n(X_n) + \frac{\epsilon}{2})$.

$\sum_1^\infty \parallel A_n \parallel F_n(X_n) = \sum_1^\infty \parallel A_n^{\frac{1}{2}} \nabla g(X_n) \parallel^2 < \infty$. Then, by H4a, c, there exists a sequence (n_l) such that $F_{n_l}(X_{n_l}) \longrightarrow 0$. Let $N_2 \geq N_1$ such that $F_{N_2}(X_{N_2}) < \epsilon'$. Let $N_2 \leq n < p$ such that for $k \in [N_2, n], F_k(X_k) < \epsilon'$ and for $k \in [n+1, p], F_k(X_k) \geq \epsilon'$; let us prove that $F_p(X_p) < \epsilon$.

We define $I_k(\epsilon') = 1$ if $F_k(X_k) \geq \epsilon', 0$ if not.

Let $Y_n = X_n + W_n$; we have $Y_{n+1} = Y_n - A_n \nabla g(X_n) - A_n W_n$.

$Y_p - Y_n = -A_n \nabla g(X_n) - A_n W_n - \sum_{n+1}^{p-1} A_k^{\frac{1}{2}} (A_k^{\frac{1}{2}} \nabla g(X_k) + A_k^{\frac{1}{2}} W_k) I_k(\epsilon')$.

$\sum_1^\infty \left\| A_k^{\frac{1}{2}} \right\| \left\| A_k^{\frac{1}{2}} \nabla g(X_k) + A_k^{\frac{1}{2}} W_k \right\| I_k(\epsilon')$

$\leq \frac{1}{2} \sum_1^\infty I_k(\epsilon') \left\| A_k \right\| + \frac{1}{2} \sum_1^\infty \left\| A_k^{\frac{1}{2}} \nabla g(X_k) + A_k^{\frac{1}{2}} W_k \right\|^2 I_k(\epsilon')$

$\leq \frac{1}{2} \sum_1^\infty I_k(\epsilon') \frac{\left\| A_k^{\frac{1}{2}} \nabla g(X_k) \right\|^2}{F_k(X_k)} + \sum_1^\infty \left\| A_k^{\frac{1}{2}} \nabla g(X_k) \right\|^2 + \sum_1^\infty \left\| A_k \right\| \left\| W_k \right\|^2 I_k(\epsilon') < \infty$ by H3a.

Furthermore by H4b and H3a, $A_n\nabla g(X_n) + A_nW_n \longrightarrow 0, W_n \longrightarrow 0$. Then for n sufficiently large $\|Y_p - Y_n\| < \frac{\delta}{2}, \|X_p - X_n\| < \delta$; therefore $F_p(X_p) \leq KF_n(X_n) + \frac{\epsilon}{2} < K\epsilon' + \frac{\epsilon}{2} \leq \epsilon$.

If $F_{p+1}(X_{p+1}) \geq \epsilon'$, then for $k \in [n+1, p+1], F_k(X_k) \geq \epsilon'$; thus by the preceding argument $F_{p+1}(X_{p+1}) < \epsilon$; if $F_{p+1}(X_{p+1}) < \epsilon'$, replace N_2 by $p+1$ and use again the preceding argument; it follows that whatever $q \geq N_2, F_q(X_q) < \epsilon$; thus $F_q(X_q) \longrightarrow 0$ as $q \longrightarrow \infty$.

Let $a = \sup_n \frac{\lambda_{\max}(A_n)}{\lambda_{\min}(A_n)}$.

$$\|\nabla g(X_n)\|^2 \leq a\frac{\lambda_{\min}(A_n)}{\lambda_{\max}(A_n)}\|\nabla g(X_n)\|^2 \leq a\frac{\| A_n^{\frac{1}{2}}\nabla g(X_n) \|^2}{\| A_n \|} = aF_n(X_n).$$

By H4d, $\nabla g(X_n) \longrightarrow 0$ as $n \longrightarrow \infty$.

Let $d(X_n, S)$ be the distance of X_n to the set S of stationary points of g. Suppose that $d(X_n, S) \nrightarrow 0$. If the sequence (X_n) is contained in C, there exist $\epsilon > 0$, a subsequence $(X_{n'})$ and $\theta \in \mathbb{R}^p$ such that $X_{n'} \longrightarrow \theta$ and $d(\theta, S) > \epsilon$; now if ∇g is continuous in C, $\nabla g(X_{n'}) \longrightarrow \nabla g(\theta) = 0$, a contradiction; thus $d(X_n, S) \longrightarrow 0$. ■

2.3 Proof of the theorem

Proof. By H1b and H4a, as $\left\|A_n^{\frac{1}{2}}\nabla g(X_n)\right\|^2 = \langle A_n\nabla g(X_n), \nabla g(X_n)\rangle$, we have:

$$g(X_{n+1}) - g(X_n) \leq -\left\|A_n^{\frac{1}{2}}\nabla g(X_n)\right\|^2 - \langle A_nV_n, \nabla g(X_n)\rangle + 2L\|A_n\nabla g(X_n)\|^2 + 2L\|A_nV_n\|^2.$$

By H2a, b, as $\langle A_nE[V_n \mid T_n], \nabla g(X_n)\rangle \leq \frac{1}{2}\left\|A_n^{\frac{1}{2}}E[V_n \mid T_n]\right\|^2 + \frac{1}{2}\left\|A_n^{\frac{1}{2}}\nabla g(X_n)\right\|^2$

$a.s.$ and $\|A_n \nabla g(X_n)\|^2 \leq \|A_n\| \left\|A_n^{\frac{1}{2}} \nabla g(X_n)\right\|^2$, we have $a.s.$:

$$\mathrm{E}[g(X_{n+1}) \mid T_n] \leq g(X_n)(1+\tfrac{1}{2}B_n+2LD_n)+\tfrac{1}{2}C_n+2LE_n-(\tfrac{1}{2}-2L\|A_n\|)\left\|A_n^{\frac{1}{2}} \nabla g(X_n)\right\|^2$$

Applying the Robbins-Siegmund lemma (1971) gives the conclusion of (a).

By H2a, b, since $g(X_n)$ converges $a.s.$, we have $\sum_1^\infty \left\|A_n^{\frac{1}{2}} E\left[V_n \mid T_n\right]\right\|^2 < \infty$ and $\sum_1^\infty E\left[\|A_n V_n\|^2 \mid T_n\right] < \infty$ $a.s.$ Apply part (a) of the theorem to the process (W_n), with $g(W_n) = \frac{1}{2}\|W_n\|^2, \nabla g(W_n) = W_n$ and $L = \frac{1}{2}$; it follows that $\|W_n\|$ and $\sum_1^\infty \left\|A_n^{\frac{1}{2}} W_n\right\|^2$ converge $a.s.$ By H4c, d and given that $\left\|A_n^{\frac{1}{2}} W_n\right\|^2 \geq \lambda_{\min}(A_n)\|W_n\|^2$, it follows that W_n converges $a.s.$ to 0. We apply then lemma 2 to obtain (b). ■

3 Second almost sure convergence theorem

Let $Y_n = X_n + W_n$; we have $Y_{n+1} = Y_n - A_n \nabla g(X_n) - A_n W_n$. Suppose:

(H3b) $\sum_1^\infty \left\|A_n^{\frac{1}{2}} W_n\right\|^2 < \infty$ $a.s.$

We give another theorem where assumptions H2 on (V_n) are replaced by assumptions H3a, b on (W_n), but where assumption H1b on g is replaced by

(H1b') ∇g is L-Lipschitz continuous $(\forall x_1, x_2, \|\nabla g(x_1) - \nabla g(x_2)\| \leq L\|x_1 - x_2\|)$.

Theorem 3 *(a) Assume H1a, b', H3b and H4 a, b, d hold; then* $g(Y_n), \sum_1^\infty \left\|A_n^{\frac{1}{2}} \nabla g(Y_n)\right\|^2$ *and* $\sum_1^\infty \left\|A_n^{\frac{1}{2}} \nabla g(X_n)\right\|^2$ *converge a.s.*

(b) Under the assumptions of (a), if moreover H3a, H4c hold, then we have the same conclusions as in part (b) of theorem 1.

We give in section 4 sufficient conditions to ensure H3a such as $V_n \longrightarrow 0$ $a.s.$ or $\sum_1^\infty A_n V_n$ converges $a.s.$ The preceding theorem can be used particularly in cases where the random errors V_n are correlated. See among others the remarks 1.3 and 1.13b of Walk (Ljung, Pflug, Walk, 1992) for $A_n = a_n I$.

Proof. By H1b' and H4a, we have

$$\begin{aligned} g(Y_{n+1}) - g(Y_n) \quad &\leq \quad L \left\| A_n \nabla g(X_n) + A_n W_n \right\|^2 \\ &\quad - \left\langle A_n^{\frac{1}{2}} (\nabla g(X_n) - \nabla g(Y_n)) + A_n^{\frac{1}{2}} W_n, A_n^{\frac{1}{2}} \nabla g(Y_n) \right\rangle - \left\| A_n^{\frac{1}{2}} \nabla g(Y_n) \right\|^2 \end{aligned}$$

The first term on the right-hand of the inequality is bounded up by

$L(\|A_n(\nabla g(X_n) - \nabla g(Y_n))\| + \|A_n W_n\| + \|A_n \nabla g(Y_n)\|)^2$

$\leq 2L(L+1)^2 \left\| A_n \right\|^2 \left\| W_n \right\|^2 + 2L \left\| A_n \right\| \left\| A_n^{\frac{1}{2}} \nabla g(Y_n) \right\|^2$.

The second term is bounded up by $(L+1)^2 \left\| A_n \right\| \left\| W_n \right\|^2 + \frac{1}{2} \left\| A_n^{\frac{1}{2}} \nabla g(Y_n) \right\|^2$.

It follows that $a.s.$:

$$E\left[g(Y_{n+1}) \mid T_n \right] \leq g(Y_n) + (L+1)^2 \left\| A_n \right\| (2L \left\| A_n \right\| + 1) \left\| W_n \right\|^2 - (\frac{1}{2} - 2L \left\| A_n \right\|) \left\| A_n^{\frac{1}{2}} \nabla g(Y_n) \right\|^2$$

By H3b and H4a, d, as $\left\| A_n^{\frac{1}{2}} W_n \right\|^2 \geq \lambda_{\min}(A_n) \left\| W_n \right\|^2$, we have $\sum_1^\infty \left\| A_n \right\| \left\| W_n \right\|^2 < \infty$ $a.s.$

By H1a and H4b, we may then apply the Robbins-Siegmund lemma: $g(Y_n)$ and $\sum_1^\infty \left\| A_n^{\frac{1}{2}} \nabla g(Y_n) \right\|^2$ converge $a.s.$ We deduce that $\sum_1^\infty \left\| A_n^{\frac{1}{2}} \nabla g(X_n) \right\|^2$ converges $a.s.$

By H1b' and H4d, H1c holds. We may then apply lemma 2 and obtain part (b) of the theorem. ■

4 Convergence lemma of W_n

Suppose:

(H2a') $\sum_1^\infty \left\| A_n^{\frac{1}{2}} E\left[V_n \mid T_n\right] \right\|^2 < \infty \ a.s.$

(H2b') $\sum_1^\infty E\left[\|A_n V_n\|^2 \mid T_n\right] < \infty \ a.s.$

(H2c) $V_n \longrightarrow 0 \ a.s.$

(H2d) $\sum_1^\infty A_n V_n$ converges $a.s.$

(H4b') There exists $l > 0$ such that $\sup_n \lambda_{\max}(A_n) < l \ \ a.s.$

Lemma 4 *(a) Assume H2a', b' and H4a, b' with $l = \frac{1}{2}$, c, d hold; then W_n converges a.s. to 0 and $\sum_1^\infty \left\| A_n^{\frac{1}{2}} W_n \right\|^2$ converges a.s.*

(b) Assume H2c, or H2d, and H4a, b' with $l = 1$, c, d hold; then W_n converges a.s. to 0.

Proof. The part (a) has been already proved in the proof of theorem 1.

For the part (b), $\omega \in \Omega$ is fixed throughout the proof, belonging to the intersection of the defined $a.s.$ convergence sets. Denote $C_l = I - A_l$. We have:

$W_{n+1} = C_n W_n + A_n V_n = \sum_{j=1}^n \prod_{l=j+1}^n C_l A_j V_j$, with $\prod_{l=n+1}^n C_l = I$.

Denote $a_{nj} = \prod_{l=j+1}^n \|C_l\| \|A_j\|, j = 1, ..., n-1, a_{nn} = \|A_n\|$.

By H4a, b', c, d, $a_{nj} < \prod_{l=j+1}^n (1 - \lambda_{\min}(A_l))$ and

$$\ln a_{nj} < -\sum_{l=j+1}^n \lambda_{\min}(A_l) \longrightarrow -\infty, a_{nj} \longrightarrow 0 \ as \ n \longrightarrow \infty.$$

Furthermore, there exists $a > 0$ such that

$$a_{nj} \leq a \prod_{l=j+1}^{n}(1-\lambda_{\min}(A_l))\lambda_{\min}(A_j) = a(\prod_{l=j+1}^{n}(1-\lambda_{\min}(A_l)) - \prod_{l=j}^{n}(1-\lambda_{\min}(A_l))).$$

Then $\sum_{j=1}^{n} a_{nj} \leq a$.

1) $\|W_{n+1}\| \leq \sum_{j=1}^{n} a_{nj} \|V_j\|$. By H2c, $W_n \longrightarrow 0$.

2) Denote $S_n = \sum_{j=1}^{n} A_j V_j, S_0 = 0, B_j = (\prod_{l=1}^{j} C_l)^{-1}$.

$$\begin{aligned} W_{n+1} &= B_n^{-1} \sum_{j=1}^{n} B_j A_j V_j = B_n^{-1} \sum_{j=1}^{n} B_j (S_j - S_{j-1}) \\ &= S_n + B_n^{-1} \sum_{j=2}^{n} (B_{j-1} - B_j) S_{j-1} = S_n - \sum_{j=2}^{n} \prod_{l=j+1}^{n} C_l A_j S_{j-1}. \end{aligned}$$

By H2d, let $S = \lim_{n\longrightarrow\infty} S_n$. As $\sum_{j=2}^{n} \prod_{l=j+1}^{n} C_l A_j = I - \prod_{l=2}^{n} C_l$, we have:

$$W_{n+1} = S_n - S - \sum_{j=2}^{n} \prod_{l=j+1}^{n} C_l A_j (S_{j-1} - S) + \prod_{l=2}^{n} C_l S.$$

$$\|W_{n+1}\| \leq \|S_n - S\| + \sum_{j=2}^{n} a_{nj} \|S_{j-1} - S\| + \prod_{l=2}^{n} \|C_l\| \|S\|.$$

By H4a, b', c, d, $\prod_{l=2}^{n} \|C_l\| \longrightarrow 0$ as $n \longrightarrow \infty$; by H2d, $W_n \longrightarrow 0$. ■

5 Corollary

Suppose $V_n = V_n^1 + V_n^2 + V_n^3 + V_n^4$. Define for i=1, 2, 3, 4, W_n^i such that $W_{n+1}^i = (I - A_n)W_n^i + A_n V_n^i$ and $W_n = \sum_{i=1}^{4} W_n^i$.

Corollary 5 *Assume that H2a', b' hold for V_n^1, H2c for V_n^2, H2d for V_n^3 and H3a for V_n^4; assume furthermore H1a, b', H3b, H4a, b, c, d hold; then $\nabla g(X_n)$ converges a.s. to 0.*

In the case where $A_n = a_n I$, this corollary can be compared to theorem 1.2a of Walk (Ljung, Pflug, Walk, 1992) who makes a more restrictive assumption, $\sum_1^\infty a_n \left\|V_n^1\right\|^2 < \infty$, than H2a', b'.

References

Dippon, J. (1998), Globally convergent stochastic optimization with optimal asymptotic distribution, *J. Appl. Proba.* **35**, 395-406.

Ljung, L., Pflug, G., Walk, H. (1992), *Stochastic Approximation and Optimization of Random Systems*, DMV Seminar **17** (Birkhäuser).

MacQueen, J. (1967), Some methods for classification and analysis of multivariate observations, *Proc. 5th Berkeley Symp. Prob. Stat.* pp. 281-297.

Nevel'son, M.B., Has'minskii, R.Z. (1973), *Stochastic Approximation and Recursive Estimation*, Transl. of Math. Monographs **47** (AMS, Providence, Rhode Island).

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

Spall, J.C. (2003), *Introduction to Stochastic Search and Optimization* (Wiley, Hoboken, New Jersey).