Bulletin de la Société Royale des Sciences de Liège, Vol. 79, 2010, p. 131 - 140

(Manuscrit reçu le 5 juillet 2010 et accepté le 26 novembre 2010)

Consistency of Robbins Monro's algorithm within a mixing framework

Abdelnasser DAHMANI, Ahmed AIT SAIDI

Abstract

This work is devoted to the study of the consistency of Robbins-Monro's algorithm under strong mixing assumption.

Mathematics Subject Classification: 62F12, 62L20, 62J05.
Key words: stochastic approximation, convergence rate, mixing.

1 Introduction

The methodologies known under the term of stochastic approximations traces its origin from the work of Robbins and Monro [11] which studied the following problem. Let R be a function of real values and θ to be the single solution of the equation

$$R(x) = \alpha \tag{1}$$

where α is a known constant. The problem is to estimate θ. When R is a known function, we can found various numerical methods to approximate θ. A part from some general properties, Robbins and Monro considered the case where R is unknown but, for each point x, we have a random variable $\widetilde{R}(x, \xi)$ such as

$$R(x) = E\left(\widetilde{R}(x, \xi)\right) \tag{2}$$

where ξ is a random variable with zero mean. These authors argued that a recursive sequence of random variables $(X_n)_n$ which estimate θ in a consistent way, can be constructed. They show the mean square convergence of X_n to θ. Considering a weaker assumptions and within the usual framework of independent and identically distributed random errors, Blum has shown the almost sure convergence [1]. Whenever R is linear and checking the classical assumptions, Lai and Robbins argued that all least squares estimator of θ properties remain true even when the estimator of θ is obtained by the Robbins-Monro method [7]. For the nonlinear general case, a procedure of type Robbins-Monro was introduced by Venter [12] and it has been discussed in [8], and the obtained results were extended to the multivariate case by Wei [14]. In [5], Duflo has shown the almost sure convergence if the Robbins-Monro's algorithm is evaluated in $\mathbb{R}^d$.

Some theoretical results on stochastic approximation can be found in various literatures, e. g. see [13], [9] and [6].

Let us note that the independent observations are often unable to explain some phenomena, indeed, the slightly dependent observations are the most adapted to a real situation [4]. In this case and concerning different models building for the dependence of the stochastic algorithm noise, we can refer to the Brandière and Doukhan note [2].

The principal contribution of this work is to consider the least restrictive mixing sequence called also α-mixing sequence, (see [3] for more details) and to provide the almost complete $(a.co)$ convergence rate of the Robbins-Monro's algorithm.

2 Algorithm and asymptotic study

Let $(\Omega, \mathcal{F}, P)$ be the probability space and $R : \mathbb{R} \to \mathbb{R}$ to be a function known just under a measure $\widetilde{R}(x, \xi)$ with a spot of a measure error ξ. To estimate the θ root of the equation (1), Robbins and Monro [11] built their algorithm in a recursive manner using an initial value X_1 and defining by recurrence:

$$X_{n+1} = X_n - a_n \left(\widetilde{R}(X_n, \xi_n) - \alpha \right) \tag{3}$$

where $(\xi_n)_n$ is a sequence of a real random variables with zero mean and $(a_n)_n$ is a decreasing deterministic sequence to 0 such as

$$\sum_{n=1}^{+\infty} a_n = +\infty \quad \text{and} \quad \sum_{n=1}^{+\infty} a_n^2 < +\infty \tag{4}$$

and

$$\widetilde{R}(X_n, \xi_n) = R(X_n) + \xi_n.$$

Without loss of generality, let us suppose $\alpha = 0$.

Removing θ from both members of the equality (3) and using successive iterations, we obtain

$$|X_{n+1} - \theta| = \left| \prod_{k=1}^{n} \left(1 - a_k \frac{R(X_k)}{X_k - \theta} \right) \right| \left| (X_1 - \theta) - \sum_{i=1}^{n} Z_i \right| \tag{5}$$

where

$$Z_i = a_i \prod_{k=1}^{i} \left(1 - a_k \frac{R(X_k)}{X_k - \theta} \right)^{-1} \xi_i.$$

Let us introduce now the following assumptions:

H1 : The parameter θ checks a priory

$$|X_1 - \theta| \leq H < +\infty. \tag{6}$$

H2 : R is a function satisfying

$$\forall x \in \mathbb{R}, 0 < m \leq \frac{R(x)}{x - \theta} \leq M < +\infty. \tag{7}$$

H3 : We suppose that, for any $\varepsilon > 0$,

$$\varphi_n(\varepsilon) = n^{am} \exp(am\gamma)\varepsilon - H > 0 \tag{8}$$

where γ is the Euler constant.

H4 : The distributed variables queues ξ_i check the condition of uniform decrease, that is, for any $p > 2$,

$$\forall t > 0 \ , \ P\left[|\xi_i| > t\right] \leq t^{-p}. \tag{9}$$

H5 : We assume that the coefficients of the α-mixing sequence $(\xi_n)_n$ satisfy the following arithmetic decay condition :

$$\exists d \geq 1, \exists b > 0, \ \alpha(n) \leq dn^{-b}. \tag{10}$$

H6 : The condition of arithmetically decrease (10) is satisfied for any b value such as

$$\exists \delta > 0, \ \frac{4b + p(3 - b)}{(b+1)\, p} + \delta \leq am - 1. \tag{11}$$

At last, we notice that if the ξ_i random errors are α-mixing then the Z_i random variables are also strongly mixing with mixing coefficients lower or equal than those of the sequence $(\xi_i)_i$.

We can know state the following result :

Theorem 1. *Under the assumptions (H1)–(H6) and if* $0 < aM < \frac{1}{2}$ *then for any real* b *positive such as*

$$b > \frac{(2-am)q}{\nu_0} \text{ with } \nu_0 \in]0,1[\text{ and } q \text{ is such as } \frac{2}{p} + \frac{1}{q} = 1, \tag{12}$$

we have:

$$X_{n+1} - \theta = O(\sqrt{\frac{\log n}{n^{am}}}) \quad a.co. \tag{13}$$

Proof. We have

$$\log \prod_{k=1}^{n} \left(1 - a_k \frac{R(X_k)}{X_k - \theta}\right) \leq \sum_{k=1}^{n} -\frac{am}{k} = -am\left(\log n + \gamma_n\right) \tag{14}$$

where γ_n is defined by the relation $\gamma_n = \sum_{i=1}^{n} \frac{1}{i} - \log n = \gamma + (\psi(n+1) - \log n)$ where $\psi(.)$ is the digamma function.

It is obvious to show that $\gamma_n - \gamma_{n-1} = \log\left(1 - \frac{1}{n}\right) + \frac{1}{n} < 0$, for any $n > 1$. This leads to the well known result where the sequence γ_n decrease to the Euler constant γ, let

$$\gamma_n > \gamma = \lim_{n \to +\infty} \left\{ \sum_{i=1}^{n} \frac{1}{i} - \log n \right\} = 0.577215... \tag{15}$$

From this relation, we obtain

$$\log \prod_{k=1}^{n} \left(1 - a_k \frac{R(X_k)}{X_k - \theta}\right) \leq -am\left(\log n + \gamma\right) \tag{16}$$

This makes it possible to conclude

$$\prod_{k=1}^{n} \left(1 - a_k \frac{R(X_k)}{X_k - \theta}\right) \leq n^{-am} \exp\left(-am\gamma\right) \tag{17}$$

and then, using (5) and the assumptions (H1) and (H4), we deduce that

$$P\left[|X_{n+1}-\theta|>\varepsilon\right]\leq P\left[\left|\sum_{i=1}^{n}Z_i\right|>\varphi_n\left(\varepsilon\right)\right]. \tag{18}$$

On the other hand, for a rather large natural integer n we have

$$\frac{H}{n^{am}\exp(am\gamma)}<\frac{\varepsilon}{2}$$

which gives

$$\varphi_n\left(\varepsilon\right)=n^{am}\exp(am\gamma)\left(\varepsilon-\frac{H}{n^{am}\exp(am\gamma)}\right)>\frac{\varepsilon}{2}n^{am}, \tag{19}$$

so, taking $\lambda=\frac{\varepsilon}{8}$,

$$P\left[|X_{n+1}-\theta|>\varepsilon\right]\leq P\left[\left|\sum_{i=1}^{n}Z_i\right|>4\lambda n^{am}\right]. \tag{20}$$

Using (9), we show that, we can found positive constant M_1 such as, for a given $p>2$,

$$\exists p>2,\forall t>0\ ,\ P\left[|Z_i|>t\right]\leq M_1t^{-p}. \tag{21}$$

Thus, applying directly the Fuk-Nagaev exponential inequality given by Rio ([10], formula (6.19a)), to strongly mixing variables Z_i we have,

$$P\left[|X_{n+1}-\theta|>\varepsilon\right]\leq 4\left(1+\frac{\left(\lambda n^{am}\right)^2}{rs_n^2}\right)^{\frac{-r}{2}}+4Cnr^{-1}\left(\frac{r}{\lambda n^{am}}\right)^{\frac{(b+1)p}{b+p}} \tag{22}$$

for any $\varepsilon>0$ and $r\geq 1$

with

$$C=2pM_1\left(2p-1\right)^{-1}\left(2^bd\right)^{\frac{p-1}{b+p}}$$

and

$$s_n^2=\sum_{i=1}^{n}\sum_{j=1}^{n}\left|cov(Z_i,Z_j)\right|=\sum_{i=1}^{n}var\left(Z_i\right)+\sum_{i=1}^{n}\sum_{j\neq i}\left|cov(Z_i,Z_j)\right|. \tag{23}$$

On the one hand, under the assumptions (H2) and by virtue of the inequality

$$\log\left(1-x\right)\geq -x-x^2$$

we have

$$\begin{aligned}\prod_{k=1}^{i}\left(1-a_k\frac{R\left(X_k\right)}{X_k-\theta}\right) &\geq \prod_{k=1}^{i}\left(1-\frac{aM}{k}\right)\\ &\geq \left(1-aM\right)i^{-aM}e^{-(aM)^2}. \end{aligned} \tag{24}$$

From this relation and according to (21), we obtain

$$\exists M_2<+\infty : EZ_i^2\leq M_2 \quad \text{and} \quad var\left(Z_i\right)\leq C_1 i^{2(aM-1)} \tag{25}$$

with $C_1=\left(\frac{a}{1-aM}\right)^2 M_2\exp(2a^2M^2)$. As $a<\frac{1}{2M}$, we deduce that

$$\sum_{i=1}^{n} var\left(Z_i\right)\leq\sum_{i=1}^{n}\frac{C_1}{i^{2(1-aM)}}\leq DC_1 \tag{26}$$

since it is a partial sum of a convergent sequence with positive terms.

On the other hand, for $i\neq j$,

$$\left|cov(Z_i,Z_j)\right|\leq C_2 i^{aM-1}j^{aM-1}\left|E(\xi_i\xi_j)\right| \tag{27}$$

with $C_2=\left(\frac{a}{1-aM}\right)^2\exp(2a^2M^2))$. From the relation (9), we can use the Davydov-Rio inequality given by Rio (2000, formula (1.12c)) to obtain:

$$\left|E(\xi_i\xi_j)\right|\leq 2q\left(\alpha\left(|i-j|\right)\right)^{\frac{1}{q}} \tag{28}$$

and, then

$$\left|cov(Z_i,Z_j)\right|\leq 2qC_2 i^{aM-1}j^{aM-1}\left(\alpha\left(|i-j|\right)\right)^{\frac{1}{q}}. \tag{29}$$

Applying a second time the Davydov-Rio inequality to Z_i variables and using (21), we obtain

$$\forall i\neq j,\ \left|cov(Z_i,Z_j)\right|\leq 2qM_1^{2/p}\left(\alpha\left(|i-j|\right)\right)^{\frac{1}{q}} \tag{30}$$

since the mixing coefficients of the sequence $\left(Z_i\right)_i$ are lower or equal than those of the sequence $\left(\xi_i\right)_i$. Making together (29) and (30), we have

$$\sum_{i=1}^{n}\sum_{i\neq j}|cov(Z_i,Z_j)| \leq \sum_{i=1}^{n}\sum_{|i-j|\leq u_n} 2qC_2 i^{aM-1}j^{aM-1}\left(\alpha\left(|i-j|\right)\right)^{\frac{1}{q}}$$
$$+\sum_{i=1}^{n}\sum_{|i-j|>u_n}^{n} 2qM_1^{2/p}\left(\alpha\left(|i-j|\right)\right)^{\frac{1}{q}} \tag{31}$$
$$\leq \sum_{i=1}^{n}\frac{1}{i^{2(1-aM)}}\sum_{k=1}^{n}2qC_2\left(\alpha\left(k\right)\right)^{\frac{1}{q}}+2n^2qM_1^{2/p}\left(\alpha\left(u_n\right)\right)^{\frac{1}{q}}$$

or also

$$\sum_{i=1}^{n}\sum_{j\neq i}|cov(Z_i,Z_j)| \leq 2DC_2D_1+2n^2qM_1^{2/p}\left(\alpha\left(u_n\right)\right)^{\frac{1}{q}} \tag{32}$$

with $D_1 = \sum_{k=1}^{n} q\left(\alpha\left(k\right)\right)^{\frac{1}{q}}$. So, taking $u_n = [n^{\nu_0}]$, the hooks indicating the whole part, and using (10), (12), (26) and (32) we obtain, for n rather large:

$$s_n^2 = o(n^{am}) \tag{33}$$

since

$$\frac{2n^2q\left(\alpha\left(u_n\right)\right)^{\frac{1}{q}}M_1^{2/p}}{n^{am}} \leq \frac{2qd^{\frac{1}{q}}M_1^{2/p}}{n^{am-2+\nu_0\frac{b}{q}}} \longrightarrow 0. \tag{34}$$

So, taking into account (33), we have the inequality

$$P\left[|X_{n+1}-\theta|>\varepsilon\right] \leq K_1+K_2 \tag{35}$$

with $K_1 = 4\left(1+\frac{\lambda^2 n^{am}}{r}\right)^{\frac{-r}{2}}$ and $K_2 = 4Cnr^{-1}\left(\frac{r}{\lambda n^{am}}\right)^{\frac{(b+1)p}{b+p}}$.

Taking $\lambda = \frac{\rho}{4}\sqrt{n^{-am}\log n}, \rho>0$, we obtain the convergence rate.

For a suitably chosen r such as $r = K\left(\log n\right)^2$, we obtain

$$K_1 = 4\left(1+\frac{\rho^2\log n}{16r}\right)^{\frac{-r}{2}} \leq K\exp(-\rho^2\frac{\log n}{32}) = Kn^{-\frac{\rho^2}{32}} \tag{36}$$

where K indicates a generic positive constant. With regard to K_2, we have

$$
\begin{aligned}
K_2 &= 4Cnr^{-1}r^{\frac{p(b+1)}{b+p}}\lambda^{-\frac{p(b+1)}{b+p}}n^{-\frac{amp(b+1)}{b+p}} \\
&= 4Cnr^{-1+\frac{p(b+1)}{b+p}}\left(\frac{\rho}{4}\right)^{-\frac{p(b+1)}{b+p}}n^{am\frac{p(b+1)}{2(b+p)}}\left(\log n\right)^{\frac{-p(b+1)}{2(b+p)}}n^{-\frac{amp(b+1)}{b+p}} \\
&= 4Cnr^{-1+\frac{p(b+1)}{b+p}}\left(\frac{\rho}{4}\right)^{-\frac{p(b+1)}{b+p}}\left(\log n\right)^{\frac{-p(b+1)}{2(b+p)}}n^{-\frac{amp(b+1)}{2(b+p)}} \qquad (37) \\
&= 4Cnr^{-1+\frac{p(b+1)}{b+p}}\left(\frac{\rho}{4}\right)^{-\frac{p(b+1)}{b+p}}\left(\log n\right)^{\frac{-p(b+1)}{2(b+p)}}\left(nn^{am-1}\right)^{-\frac{p(b+1)}{2(b+p)}} \\
&= 4Cnr^{-1+\frac{p(b+1)}{b+p}}\left(\frac{\rho}{4}\right)^{-\frac{p(b+1)}{b+p}}\left(\log n\right)^{\frac{-p(b+1)}{2(b+p)}}n^{-\frac{p(b+1)}{2(b+p)}}n^{-(am-1)\frac{p(b+1)}{2(b+p)}}.
\end{aligned}
$$

since $r = K\left(\log n\right)^2$, we have

$$K_2 = 4C\left(\frac{\rho}{4}\right)^{-\frac{p(b+1)}{b+p}}\left(\log n\right)^{\frac{b(3p-4)-p}{2(b+p)}}n^{\frac{(b(2-p)+p)}{2(b+p)}}n^{-(am-1)\frac{p(b+1)}{2(b+p)}}. \qquad (38)$$

By virtue of the condition (11), we obtain

$$K_2 \leq 4C\left(\frac{\rho}{4}\right)^{-\frac{p(b+1)}{b+p}}\left(\log n\right)^{\frac{b(3p-4)-p}{2(b+p)}}n^{-1-\frac{\delta p(b+1)}{2(b+p)}}. \qquad (39)$$

Consequently, it exists $\widetilde{d} > 0$ such as

$$K_2 \leq Kn^{-1-\widetilde{d}}, \qquad (40)$$

so, for $\varepsilon = 2\rho\sqrt{n^{-am}\log n}$ and for ρ sufficiently large, we have

$$P\left[\left|X_{n+1}-\theta\right| > 2\rho\sqrt{n^{-am}\log n}\right] \leq Kn^{\frac{-\rho^2}{32}} + Kn^{-1-\widetilde{d}} \leq Kn^{-1-\widetilde{d}}. \qquad (41)$$

The right-hand side of the latter inequality is a convergent series. So, (41) leads to the result. $\square$

Application. Finding a root of a regression function.

A typical example is $\widetilde{R}\left(X,\xi\right) = R\left(X\right) + \xi$. By conditioning with respect to X and moving to the expectation, we can write:

$$E(\widetilde{R}\left(X,\xi\right) \mid X) = E\left(R\left(X\right) \mid X\right) + E(\xi \mid X).$$

Bulletin de la Société Royale des Sciences de Liège, Vol. 79, 2010, p. 131 - 140

Assuming that $E(\xi \mid X) = 0$, we have

$$E(\widetilde{R}\,(X,\xi) \mid X) = E\,(R\,(X) \mid X)\,. \tag{42}$$

Thus, the search for the root function

$$R(x) = E(\widetilde{R}\,(X,\xi) \mid X = x) = E\,(R\,(X) \mid X = x)$$

reduces to that of the regression function $R\,(X)$ on X. It is therefore possible to use the stochastic algorithm of Robbins-Monro to find the root of a unimodal regression function. To characterize the strong mixing (α-mixing) random errors ξ_i, it suffices to consider an autoregressive model of order 1

$$\xi_i = \phi\xi_{i-1} + v_i$$

where v_i is a Gaussian white noise process and $|\phi| < 1$. This situation mainly occurs in time series models.

References

[1] J.R. Blum, Approximation methods which converge with probability one, Ann. Math. Stat., 25 (1954), 382–386

[2] P. Doukhan, O. Brandière, Dependent noise for stochastic algorithms, C. R. Acad. Sci. Paris, Ser. I 337 (7) (2003) pp. 473–476.

[3] P. Doukhan, Mixing : Properties and examples. Lect. notes in statistic, 80, Springer-Verlag. Berlin, 1994.

[4] P. Doukhan, S. Louhichi, A new weak dependence condition and applications to moment inequalities, Stochastic Process and their applications 84, (1999), pp.313-342.

[5] M. Duflo, Méthodes récursives aléatoires. Masson 1990.

[6] A.N. Korostelev, Procédures stochastiques récurrentes : propriétés locales. Naouka, Moscou, 1984 (en russe).

[7] T.L. Lai, H. Robbins, adaptive design in regression and control, Proc., Nat., Acad., Sci., USA, 75, (1978), pp. 586-587.

[8] T.L. Lai, H. Robbins, adaptive design in regression and stochastic approximation, Ann. Statist., 7, (1979), pp. 1196-1221.

[9] M.B. Nevelson, R.Z. Hasminskii, Stochastic approximation and recursive estimation, Amer., Math., Soc., Providence, R.I., 1973.

[10] E. Rio, Théorie asymptotique des processus faiblement dépendants. In Mathématiques & Applications, 31, Springer-Verlag, Berlin Heidelberg, 2000.

[11] H. Robbins, S. Monro. A stochastic approximation method. Ann. Math. Stat., 22. N°1. (1951), p.400-407.

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

[13] M.T. Wasan, Stochastic approximation. Cambridge. University press, 1969.

[14] C.Z. Wei, Multivariate adaptive stochastic approximation, Ann. of Stat., Vol. 15, N°3, (1987) , pp. 1115-1130.

Abdelnasser DAHMANI, Ahmed AIT SAIDI
Laboratory of Applied Mathematics
University of Bejaia, Algeria
E-mail : a_dahmany@yahoo.fr