Applied Mathematical Sciences, Vol. 2, 2008, no. 8, 387 - 395

Almost Sure Convergence and in Quadratic Mean of the Gradient Stochastic Process for the Sequential Estimation of a Conditional Expectation

A. Bennar [1], A. Bouamaine [2] and A. Namir [1]

[1] Département de Mathématiques et Informatique
Faculté des sciences Ben M'sik
Université Hassan II, Mohammedia, Casablanca, Maroc
bennar1@yahoo.fr

Ecole Nationale Supérieure d'éléctricité et de Mécanique,
Université Hassan II, Ain Chok, Casablanca, Maroc

Abstract. In this work, we present results of Almost Sure Convergence and in Quadratic Mean of the gradient stochastic process for the sequential estimation of a conditional expectation.
This work is motived by the researsh of new easy approach to estimate the conditional expectation.

Mathematics Subject Classifications: Primary 62; Secondary L20

Keywords: Stochastic approximation, Conditional expectation, stochastic gradient

1 . Introduction

Let an observable real random variable U. Let an random variable V, to values in $\mathbb{R}^k$, of law μ and ϕ a real function in $\mathbb{R}^k \times \mathbb{R}^p$, measurable.
We tries to appraise the parameter x of $\mathbb{R}^p$ such that $\phi(V, x)$ approach $E[U/V]$ in the least squares sense.

Let f the real positive function defined in $\mathbb{R}^p$ by

$$f(x) = E\left[\left(E[U/V] - \phi(V,x)\right)^2\right].$$

We look for θ that minimizes the function f.
Let's define the real positive function g in $\mathbb{R}^p$ by

$$g(x) = E\left[\left(U - \phi(V,x)\right)^2\right]$$

One has : $g(x) = f(x) + E\left[\left(U - E[U/V]\right)^2\right]$

Therefore, the problem comes back to look for θ that minimizes the function g.

We have : $\nabla_x g(x) = 2E\left[(\phi(V,x) - U)\nabla_x \phi(V,x)\right]$

To estimate θ of sequential way, we construct a stochastic gradient algorithm (See ROBBINS-MONRO[1],PARISOT[4]) (X_n) in $\mathbb{R}^p$ such that

$$X_{n+1} = X_n - a_n \nabla_x \phi(V_n, X_n))(\phi(V_n, X_n) - U_n)$$

with :

$*$ (a_n) is a sequence of positive real numbers;

$*$ $(U_1,V_1),(U_2,V_2),...,(U_n,V_n)$ is a sample of independent random variable couples and distributed identically of (U,V).

2 . Almost Sure Convergence

- Let's make the following hypotheses :

(H_1) $a_n > 0,\ \sum_1^{\infty} a_n^2 < \infty$

(H_2) there exists a and b such thats, for all $\theta = (\theta_1, \theta_2, ..., \theta_p)' \in \mathbb{R}^p$,

$$Var\left[\frac{\partial \phi(V,x)}{\partial x_i}(\phi(V,x) - U)\right] < ag(x) + b,\ for\ all\ i = 1,2,...,p.$$

(H_3) there exists $K_1 > 0$ such that, for all $x = (x_1, x_2, ..., x_p)'$,

$$\left|\frac{\partial^2 g(x)}{\partial x_i \partial x_j}\right| < K_1, \quad for\ i,j = 1,2,...,p.$$

Lemmas

Lemma 1

Under hypotheses H_1, H_2, H_3, we have :

$\cdot$ $\exists\, T$ finite positive real random variable such that $g(X_n) \xrightarrow{a.s.} T$

$\cdot \ \sum_{1}^{\infty} a_n \|\nabla_x g(X_n)\|^2 < \infty \qquad a.s.$

Proof

Let

$$W_n = 2\nabla_x \phi(V_n, X_n)(\phi(V_n, X_n) - U_n)$$

We have :

$$E[W_n/T_n] = \nabla_x g(X_n) \ a.s.$$

T_n the sub-σ-algebra generated by the events before time n.
With $b_n = \frac{a_n}{2}$, We have $X_{n+1} = X_n - b_n W_n$
Let H the hessian of g; by the Taylor formula, there exists $0 < \mu < 1$ such that :

$$g(X_{n+1}) = g(X_n) - b_n < W_n, \nabla_x g(X_n) > + \frac{b_n^2}{2} < W_n, H_n W_n >$$

with

$H_n = H(X_n - \mu b_n W_n)$

Let $\qquad Y_n = W_n - \nabla_x g(X_n) = W_n - E[W_n/T_n]$

We have : $\qquad < W_n, \nabla_x g(X_n) >= \|\nabla_x g(X_n)\|^2 + < Y_n, \nabla_x g(X_n) >$

Under H_3, we have :

$$| < W_n, H_n W_n > | \leq \|H_n\| \|W_n\|^2 \leq 2K_1 \bigg(\|Y_n\|^2 + \|\nabla_x g(X_n)\|^2 \bigg)$$

Therefore :

$$\begin{aligned} g(X_{n+1}) \leq g(X_n) - b_n(1 - K_1 b_n)\|\nabla_x g(X_n)\|^2 \\ -b_n < Y_n, \nabla_x g(X_n) > K_1 b_n \|Y_n\|^2 \end{aligned}$$

As $\lim_{n\to\infty} a_n = 0$, we have $\quad b_n \leq \frac{1}{2K_1}$ from a certain rank.

Therefore, as $E[Y_n/T_n] = 0$, we have

$$\begin{aligned} &E[g(X_{n+1})/T_n] \leq \\ &\qquad g(X_n) - \frac{b_n}{2}\|\nabla_x g(X_n)\|^2 + K_1 b_n^2 E[\|Y_n\|^2/T_n] \ a.s. \end{aligned}$$

Let $Y_n = (Y_n^1, Y_n^2, ..., Y_n^p)'$, $\quad \nabla_x g(x) = \left(\frac{\partial g}{\partial x_1}, \frac{\partial g}{\partial x_2}, ..., \frac{\partial g}{\partial x_p} \right)'$

With the usual euclidian norm, we have :

$$\|Y_n\|^2 = \sum_{i=1}^{p}(Y_n^i)^2 = \sum_{i=1}^{p}\left(2\frac{\partial\phi(V_n,X_n)}{\partial x_i}(\phi(V_n,X_n)-U_n) - \frac{\partial g(X_n)}{\partial x_i}\right)^2$$

Therefore :

$$E[\|Y_n\|^2/T_n] = \sum_{i=1}^{p} Var\left[\frac{\partial\phi(V,X_n)}{\partial x_i}(\phi(V,X_n)-U)\right] \qquad a.s.$$

Under H_2, there exists the constants A and B such that

$$E[\|Y_n\|^2/T_n] \leq Ag(X_n) + B \qquad a.s.$$

Therefore:

$$E[g(X_{n+1})/T_n] \leq (1+K_1b_n^2)g(X_n) - \frac{b_n}{2}\|\nabla_x g(X_n)\|^2 + K_1Bb_n^2 \quad a.s.$$

Under the hypothesis H_1, and using the lemma of ROBBINS-SIEGMUND[3], we deduct that :

· $\exists\, T$ finite random positive variable such that

$$g(X_n) \xrightarrow{a.s.} T$$

· $\sum_{1}^{\infty} a_n\|\nabla_x g(X_n)\|^2 < \infty \quad a.s.$ ■

- Let's make the following hypotheses :

  (H_1') $a_n > 0,\ \sum_{1}^{\infty} a_n = \infty,\ \sum_{1}^{\infty} a_n^2 < \infty$

  (H_4) θ is a local minimum of g :

  $$\exists\, \alpha > 0 \ :\ (x \neq \theta,\ \|x-\theta\| < \alpha) \Rightarrow (g(\theta) < g(x))$$

  (H_5) θ is the unique stationary point of g :

  $$\forall\, x \in \mathbb{R}^p, (x \neq \theta) \ \Leftrightarrow\ (\nabla_x g(x) \neq 0)$$

Lemma 2 (Dubbins-Freedman[6])

Let's $(\Omega, \mathcal{A}, P)$ a probability space and T_n an increasing sequence of sub-σ-algebra of $\mathcal{A}$, Z_n' a real random variable, integrable, T_{n+1}-measurable. We suppose that the real random variable $E[Z_n'/T_n]$ is finite a.s.

Let $Z_n = Z'_n - E[Z'_n/T_n]$, $\quad D_n = Var[Z'_n/T_n] = E[Y_n^2/T_n]$

$$\cdot \quad If \quad \sum_n D_n < \infty \quad a.s. \quad alors \qquad \sum_n Z_n < \infty \quad p.s$$

$$\cdot \quad If \quad \sum_n D_n = \infty \quad a.s. \quad alors \quad lim_n \frac{\sum_{i=1}^n Z_i}{\sum_{i=1}^n D_i} = 0 \quad p.s.$$

Theorem

Under hypotheses $H'_1,\ H_2,\ H_3,\ H_4,\ H_5$, we have :

$$X_n \xrightarrow{a.s.} \theta^* \ or \ \|X_n\| \xrightarrow{a.s.} +\infty$$

Proof

1) Let's prove that $X_{n+1} - X_n \xrightarrow{a.s.} 0$
We have : $\quad \|X_{n+1} - X_n\| \leq a_n \|\nabla_x g(X_n)\| + a_n \|S_n\|$

with : $\quad S_n = 2\nabla_x \phi(V_n, X_n)(\phi(V_n, X_n) - U_n) - \nabla_x g(X_n)$

The lemma 1 permits to affirm that $\displaystyle\sum_1^\infty a_n \|\nabla_x g(X_n)\|^2 < \infty \quad a.s.$

what implies that $\quad a_n \|\nabla_x g(X_n)\|^2 \xrightarrow{a.s.} 0$

As $a_n \longrightarrow 0$, we have $\quad a_n \|\nabla_x g(X_n)\| \xrightarrow{a.s.} 0$

2) Let's prove that $\quad a_n \|S_n\| \xrightarrow{a.s.} 0$

Let's put $\quad Z'_n = a_n \|S_n\|, \quad Z_n = Z'_n - E[Z'_n/T_n] \quad and \quad D_n = Var[Z'_n/T_n]$

We have : $\quad D_n = a_n^2 Var[\|S_n\|/T_n] \leq a_n^2 E[\|S_n\|^2/T_n] \leq a_n^2 (Ag(X_n) + B)$

however : $g(X_n) \xrightarrow{a.s.} T$, therefore, under H'_1, we have : $\displaystyle\sum_n D_n < \infty \quad a.s.,$

the lemma 2 permits to deduct that $\displaystyle\sum_n Z_n < \infty \quad a.s.$

therefore : $\quad a_n \|S_n\| - a_n E[\|S_n\|/T_n] \xrightarrow{a.s.} 0$

Otherwise, $\quad a_n E[\|S_n\|/T_n] \leq a_n \sqrt{E[\|S_n\|^2/T_n]} \leq \sqrt{Ag(X_n) + B} \quad a.s.+$

As $\quad g(X_n) \xrightarrow{a.s.} T$, we have : $\quad a_n E[\|S_n\|/T_n] \xrightarrow{a.s.} 0$

Therefore $\quad a_n \|S_n\| \xrightarrow{p.s.} 0$, therefore : $\quad X_{n+1} - X_n \xrightarrow{a.s.} 0$

3) To prove that $\Theta_n \xrightarrow{a.s.} \theta^* \quad or \quad \|\Theta_n\| \xrightarrow{a.s.} +\infty$ we reason at $\omega \in \Omega$ fixed in the intersection a.s. convergence sets C_1, C_2, C_3 defined by :

$$C_1 = \{\omega \quad : \quad g(\Theta_n(\omega)) \quad converge\}$$

$$C_2 = \{\omega \quad : \quad \Theta_{n+1}(\omega) - \Theta_n(\omega) \longrightarrow 0\},$$

$$C_3 = \{\omega \quad : \quad \sum_1^{\infty} a_n \|\nabla_\theta g(\Theta_n(\omega))\|^2 < +\infty$$

We have the 4 following possibilities :

1) $0 = \liminf \|\Theta_n(\omega) - \theta^*\| < \limsup \|\Theta_n(\omega) - \theta^*\|$

2) $0 < \liminf \|\Theta_n(\omega) - \theta^*\| \leq \limsup \|\Theta_n(\omega) - \theta^*\| < +\infty$

3) $0 < \liminf \|\Theta_n(\omega) - \theta^*\| \leq \limsup \|\Theta_n(\omega) - \theta^*\| = +\infty$

4) $\|\Theta_n(\omega) - \theta^*\| \longrightarrow 0 \quad or \quad \|\Theta_n(\omega)\| \longrightarrow +\infty$

Let's prove that the first three possibilities are contradictory with hypotheses of the theorem.

First Case : $0 = \liminf \|\Theta_n(\omega) - \theta^*\| < \limsup \|\Theta_n(\omega) - \theta^*\|$

As $\limsup \|\Theta_n(\omega) - \theta^*\| > 0 \quad and \quad \liminf \|\Theta_n(\omega) - \theta^*\| = 0,$ it exists an infinity of vectors Θ_n such that : $\frac{\varepsilon}{2} \leq \|\Theta_n(\omega) - \theta^*\| \leq \varepsilon$

By the Bolzano-Weistrass theorem, it exists a point of accumulation of the sequence Θ_n, $\|\theta_0\|$ that verifies : $\frac{\varepsilon}{2} \leq \|\theta_0\| \leq \varepsilon$

We can then extract of sequence (Θ_n) a subsequence (Θ_{n_k}) such that :
$\lim_k \Theta_{n_k} = \theta_0 + \theta^*$. As $g(\Theta_n)$ converges and $g(.)$ is continue, we have :
$\lim_k g(\Theta_{n_k}) = g(\theta_0 + \theta^*)$
Since $\liminf \|\Theta_n(\omega) - \theta^*\| = 0$, we can extract a subsequence (Θ_{n_l}) such that : $\lim_l \Theta_{n_l} = \theta^*$. And therefore $\lim_n g(\Theta_n) = \lim_l g(\Theta_{n_l}) = g(\theta^*)$.
Therefore $g(\theta_0 + \theta^*) = g(\theta^*)$.
What is absurd with the hypothesis H_4 of the theorem.

Second Case : $0 < \liminf \|\Theta_n(\omega) - \theta^*\| \leq \limsup \|\Theta_n(\omega) - \theta^*\| < +\infty$

Let's prove that : $\liminf \|\nabla_\theta g(\Theta_n)\| > 0$.
Suppose that $\liminf \|\nabla_\theta g(\Theta_n)\| = 0$. then, it exist an integer-subsequence (n_k) such that $\lim_k \nabla_\theta g(\Theta_{n_k}) = 0$.

By the hypothesis H_5, the sequence (Θ_{n_k}) converges toward θ^*, then $\liminf \|\Theta_n(\omega) - \theta^*\| = 0$. What is absurd.
However: As $\sum_n a_n \|\nabla_\theta g(\Theta_n)\|^2 < \infty \quad and \quad \sum_n a_n = \infty$, we have $\liminf \|\nabla_\theta g(\Theta_n)\| = 0$. It is contradictory.

Third Case : $0 < \liminf \|\Theta_n(\omega) - \theta^*\| \leq \limsup \|\Theta_n(\omega) - \theta^*\| = +\infty$

We get a contradiction while using the hypothesis H_5 and the Dubbins-Freedmann-Lemma[6].

Therefore : $\Theta_n \xrightarrow{a.s.} \theta^* \quad or \quad \|\Theta_n\| \xrightarrow{a.s.} +\infty.$ ■

3 . Quadratic Mean Convergence

• Let's make the following hypotheses :

(H_8) $\phi(x,\theta)$, $\nabla_x\phi(x,\theta)$ are uniformly bounded in x and θ.

(H_9) It exists two real positives functions h and h' defined in $\mathbb{R}^p$ such that :

$\forall\theta,\ \theta' \in \mathbb{R}^p,\ \forall x \in \mathbb{R}^q,$

$$|\phi(x,\theta) - \phi(x,\theta')| \leq h(x)\|\theta - \theta'\|$$

$$\|\nabla_\theta\phi(x,\theta) - \nabla_\theta\phi(x,\theta')\| \leq h'(x)\|\theta - \theta'\|$$

$$E[h(X)] < \infty;\ E[h'(X)] < \infty$$

(H_{10}) Y is a real random bounded variable.

Lemma (Braverman[8])

Let, for all n, M_n, a_n, b_n a real positives numbers such that :

$$\forall n, \quad M_{n+1} \leq M_n + a_n + b_n,$$

$$\sum_1^\infty a_n M_n < \infty, \quad \sum_1^\infty b_n < \infty, \quad \sum_1^\infty a_n = +\infty, \quad a_n \longrightarrow 0$$

Then, we have $\lim\limits_{n\to+\infty} M_n = 0$.

Theorem

Under hypotheses $H_1', H_3, H_8, H_9, H_{10}$,

$$\text{we have :} \quad \nabla_\theta g(\Theta_n) \xrightarrow{a.s.} 0 \quad and \quad \nabla_\theta g(\Theta_n) \xrightarrow{q.m.} 0$$

Proof

i) Let's show that it exists a positive real number A such that :

$$\forall \quad \theta_1, \theta_2, \quad \|\nabla_\theta g(\theta_1) - \nabla_\theta g(\theta_2)\| \leq A\|\theta_1 - \theta_2\|$$

We have :

$(\phi(x,\theta_1) - y)\nabla_\theta\phi(x,\theta_1) - (\phi(x,\theta_2) - y)\nabla_\theta\phi(x,\theta_2)$

$= (\phi(x,\theta_1) - y)(\nabla_\theta\phi(x,\theta_1) - \nabla_\theta\phi(x,\theta_2)) + \nabla_\theta\phi(x,\theta_2)(\phi(x,\theta_1) - \phi(x,\theta_2))$

Therefore, under H_8, H_9, H_{10}, we have :

$$\|\nabla_\theta g(\theta_1)-\nabla_\theta g(\theta_2)\| \leq 2E\big[(\phi(X,\theta_1)-y)\nabla_\theta\phi(X,\theta_1)-(\phi(X,\theta_2)-y)\nabla_\theta\phi(X,\theta_2)\big]$$
$$\leq A\|\theta_1-\theta_2\|$$

ii) Let's show that it exists two real constants $c_1,\ c_2$ such that :

$$\|\nabla_\theta g(\Theta_{n+1}\|^2 \leq \|\nabla_\theta g(\Theta_n\|^2 + c_1 a_n + c_2 a_n^2$$

We have :$\Theta_{n+1} = \Theta_n - \frac{a_n}{2}W_n$, with $W_n = 2(\phi(X_n,\Theta_n) - Y_n)\nabla_\theta\phi(X_n,\Theta_n)$

Therefore,
$$\|\nabla_\theta g(\Theta_{n+1}\|^2 = \|\nabla_\theta g(\Theta_n - \frac{a_n}{2}W_n)\|^2$$
$$\leq \|\nabla_\theta g(\Theta_n - \frac{a_n}{2}W_n) - \nabla_\theta g(\Theta_n)\|^2 + \|\nabla_\theta g(\Theta_n)\|^2$$
$$+2\big\|\nabla_\theta g(\Theta_n - \frac{a_n}{2}W_n) - \nabla_\theta g(\Theta_n)\big\|\|\nabla_\theta g(\Theta_n)\|$$
$$\leq \|\nabla_\theta g(\Theta_n\|^2 + A^2\frac{a_n^2}{4}\|W_n\|^2 + 2a_n\|W_n\|\big\|\nabla_\theta g(\Theta_n)\big\|$$

Therefore, under $H_8,\ H_{10}$, it exists two real positives numbers c_1, c_2 such that :

$$(\star) \qquad \|\nabla_\theta g(\Theta_{n+1}\|^2 \leq \|\nabla_\theta g(\Theta_n)\|^2 + c_1 a_n + c_2 a_n^2$$

iii) Let's show that : $\lim_{n\to\infty} \big\|\nabla_\theta g(\Theta_n)\big\|^2 = 0 \quad a.s.$

The previous lemma affirms that: $\sum_1^\infty a_n\|\nabla_x g(X_n)\|^2 < \infty \qquad a.s.$

Then, we can apply the Braverman-Lemma with :

$$M_n = \big\|\nabla_\theta g(\Theta_n)\big\|^2 \quad \text{and} \quad b_n = c_2 a_n^2$$

iv) Let's show that : $\lim_{n\to\infty} E\left[\big\|\nabla_\theta g(\Theta_n)\big\|^2\right] = 0$

We have $E\big[g(\Theta_{n+1})/T_n\big] \leq (1+K_1a_n^2)g(\Theta_n) - \frac{a_n}{2}\|\nabla_\theta g(\Theta_n)\|^2 + K_2Ba_n^2$

(See Proof of Lemma 1)

As $E[g(\Theta_{n+1})] = E[E[g(\Theta_{n+1})/T_n]]$, we have :

$$E[g(\Theta_{n+1})] \leq (1+K_1a_n^2)E[g(\Theta_n)] - \frac{a_n}{2}E\left[\|\nabla_\theta g(\Theta_n)\|^2\right] + K_2Ba_n^2$$

By the Robbins-Siegmund-Lemma[3], under H_1', we have the almost sure convergence of $E[g(\Theta_n)]$ and of $\sum_1^\infty a_n E\left[\|\nabla_\theta g(\Theta_n)\|^2\right]$, in addition, according to relation $(\star)$, we have :

$$E\left[\|\nabla_\theta g(\Theta_{n+1}\|^2\right] \leq E\left[\|\nabla_\theta g(\Theta_n)\|^2\right] + c_1 a_n + c_2 a_n^2$$

We apply the Braverman-Lemma[8], with :

$$M_n = E\left[\|\nabla_\theta g(\Theta_n)\|^2\right] \quad \text{and} \quad b_n = c_2 a_n^2$$

Therefore : $\nabla_\theta g(\Theta_n) \xrightarrow{q.m.} 0$ ■

References

[1] H. ROBBINS-MONRO. *A stochastic approximation method.* A.M.S. , 1951 , Vol 22, 400-407.

[2] A. BENNAR. *Approximation stochastique : Convergence dans le cas de plusieurs solutions et étude de modèles de corrélations.* Thèse de doctorat de 3ème cycle, Université de Nancy I, 1985.

[3] H. ROBBINS, D. SIEGMUND. *A convergence theorem for nonnegative almost upermartingales and some applications.* Optimizing methods in statics , edited by J.S. RUSTAGI , Academic Press , New York , 1971, 233-257 .

[4] J.P. PARISOT. *Optimisation stochastique : le processus de Kiefer-Wolfowitz. Essai de synthèse et quelques compléments.* Thèse de doctorat de 3ème cycle, Université de Nancy I, 1981.

[5] J.H. VENTER. *On convergence of the Kiefer-Wolfowitz process. Approximation procedure.* A.M.S. Vol.38, p.1031-1036.

[6] D.A. FREEDMAN, L.E. DUBINS. *A sharper from of the Borel-Cantelli lemma and the strong law.* A.M.S. Vol.36, p.800-807.

[7] A. BOUAMAINE. *Méthodes d' approximation stochastique en Analyse des Données.* Thèse de doctorat dètat, Université Mohamed V, 1996.

[8] BRAVERMAN E.M., ROZONOER L. T.(1969). *Convergence of trandom process in learning machines theory. Part I and II.* Automation and Remote Control. Vol 30,pp. 44-64 and 386-402.

[9] J.C. SPALL. *Introduction to Stochastic Search and Optimizing* Wiley, Hoboken, New Jersey(2003).

[10] J. DIPPON. *Globally convergent stochastic optimization with optimal asymptotic distribution.* J. Appl.Proba.35,(1998) pp: 395-406.

Received: July 30, 2007