Applied Mathematical Sciences, Vol. 2, 2008, no. 25, 1219 - 1228

Some Applications of Stochastic Gradient Processes

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, B.P. 7955 Casablanca, Maroc

[2] Ecole Nationale Supérieure d'éléctricité et de Mécanique
Equipe de Recherche : Architecture des Systèmes
Université Hassan II, Ain Chok, B.P. 8118 Casablanca, Maroc

Abstract

We consider a stochastic gradient process, which is used for the estimation of a conditional expectation : $X_{n+1} = X_n - a_n \nabla_x \phi(V_n, X_n)(\phi(V_n, X_n) - U_n)$. We give one theorem of almost sure convergence and one theorem of mean quadratic convergence. Several applications are given : linear estimation of a conditional expectation, sequential estimation of law mixture parameters in classification, estimation of an observable function in random points, estimation of a function $h(x) = E[Z(x)]$, estimation of a linear regression parameters, estimation of baysian discriminant function.

Mathematics Subject Classification: Primary: 62; Secondary: L20

Keywords: Stochastic approximation, conditional expectation, stochastic gradient

1. Introduction

We consider a random vector X_n in $\mathbb{R}^p$ defined by :

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

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 with the same probability law that (U, V).

$*$ $\phi(.,.)$ is a real known measurable function in $\mathbb{R}^k \times \mathbb{R}^p$.

In the following, $\langle .,. \rangle$ and $\|.\|$ are respectively the usual inner product and norm in $\mathbb{R}^k$; A' denotes the transposed matrix of A, $\lambda_{\min}(B)$ the smallest eigenvalue of B ; the abbreviation $a.s.$ means almost surely and $q.m.$ means quadratic mean.

2. Convergence

2.1 Almost Sure Convergence

• Let's make the following hypotheses :

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

(H_1') $a_n > 0$, $\sum_1^\infty a_n = \infty$, $\sum_1^\infty a_n^2 < \infty$

(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 $K > 0$ such that, for all $x = (x_1, x_2, ..., x_p)'$,

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

(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 :

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

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

See [3] ■

2.2 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.

Theorem

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

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

Proof

See [3] ■

3. Applications

3.1 Séquential Estimation of a Conditional Expectation by the linear model

Let $\rho_1, \rho_2, ..., \rho_p$ p functions of q real variables, measurable, known.

Let's put $\rho = (\rho_1, \rho_2, ..., \rho_p)'$

to appraise θ that minimizes $E\left[(E[U/V] - x'\rho(V))^2\right]$, We consider the stochastic approximation process (X_n) in $\mathbb{R}^p$ by :

$$X_{n+1} = X_n - a_n\rho(V_n)(\rho'(V_n)X_n - U_n)$$

Where $(U_1, V_1), (U_2, V_2), ..., (U_n, V_n)$ is a sample of (U, V)formed of independent random variables and distributed identically.

Let's make hypotheses

(H_6) $\rho_1(V), \rho_2(V), ..., \rho_p(V)$ are linearly independent.

(H_7) Moments of order 4 of the vector $(\rho_1(V), \rho_2(V), ..., \rho_p(V), U)$ exists.

(H_8) X_1 is an random variable such that $E[\|X_1\|^2] < \infty$

Corollary

Under hypotheses H_1', H_6, H_7, H_8, we have : $X_n \xrightarrow{a.s.} \theta$

Proof

Let ϕ the real function of $\mathbb{R}^q \times \mathbb{R}^p$ defined by :

$$\phi(V, x) = x'\rho(V) = \sum_{j=1}^{p} x_j\rho_j(V)$$

For $j = 1, 2, ..., p$, $\dfrac{\partial\phi(V, x)}{\partial x_j} = \rho_j(V)$, we have : $\nabla_x\phi(V, x) = \rho(V)$

Let : $A = E[\rho(V)\rho'(V)]$

Under H_7, the matrix A is symmetrical definite positive, therefore inversible.
Then :
θ^* is solution unique of the equation

$$\nabla_x g(x) = 2E[\rho(V)(\rho'(V)x - U)] = 0$$

We have : $$\theta^* = A^{-1}E[\rho(V)U]$$

• Let's prove that the hypothesis H_2 is verified.

We have

$$g(x) = E[U^2] + \|x\|^2 - 2 < x, \theta^* >_A = E[U^2] + \|x - \theta^*\|_A^2 - \|\theta^*\|_A^2$$

$$\geq c\|x - \theta^*\|_A^2 + d \quad (c = \lambda_{min(A)} \text{ and } d = E[U^2] - \|\theta^*\|_A^2)$$

$$\geq \frac{1}{2}c\|x\|^2 - c\|x\|^2 + d \geq e\|x\|^2 + f \quad (e = \frac{c}{2} \text{ and } f = d - c\|x\|^2)$$

Therefore, for $i = 1, 2, ..., p$, we have:

$$Var\left[\frac{\partial\phi(V,x)}{\partial x_i}(\phi(V,x) - U)\right] = Var\left[\rho_i(V)(x'\rho(V) - U)\right]$$

$$\leq E[\rho_i^2(V)(x'\rho(V) - U)^2]$$

$$\leq a\|x\|^2 + b \quad (a = 2E[\rho_i^2(V)\|\rho(V)\|^2],\ b = 2E[\rho_i^2(V)U^2])$$

$$\leq Ag(x) + B \quad (A = \frac{1}{e}\ ,\ B = b - \frac{af}{e})$$

• Let's prove that the hypothesis H_3 is verified.

For $i = 1, 2, ..., p$, We have $\dfrac{\partial g(x)}{\partial x_i} = 2E[(x'\rho(V) - U)\rho_i(V)]$

Therefore : for $i, j = 1, 2, ..., p$, we have

$\dfrac{\partial^2 g(x)}{\partial x_i \partial x_j} = 2E[\rho_j(V)\rho_i(V)]$, that doesn't depend of x.

Hypotheses of the theorem 2.1 are verified, therefore :

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

• Let's prove that that we can not have $\|X_n\| \xrightarrow{a.s.} +\infty$

Indeed : as $\sum_1^\infty a_n \| \nabla_x g(X_n)\|^2 < \infty$ *a.s.* (see [3] Lemma 2.1) and $\sum_n a_n = +\infty$, there exists an sub-sequence of integers (n_l) such that $\|\nabla_x g(X_{n_l})\| \xrightarrow{a.s.} 0$

Besides : $\nabla_x g(x) = 2E[\rho(V)(\rho'(V)x - U)] = E[\rho(V)(\rho'(V)] = 2A(x - \theta^*)$

Therefore : $\|\nabla_x g(X_n)\|^2 \geq 4\lambda_{min(A)}^2 \|X_n - \theta^*\|^2 \qquad (\lambda_{min(A)} > 0)$

Therefore : If $\|X_n\| \xrightarrow{p.s.} +\infty$ then $\|\nabla_x g(X_n)\| \xrightarrow{a.s.} +\infty$. What is absurd. ■

3.2 Sequential Estimation of parameters of a law mixture in

classification

Let $X_1, X_2, ..., X_n, ...$ a sample of X formed of random variables independent, distributed identically of law μ and to values in $\mathbb{R}^q$ definied on a probability space $(\Omega, \mathcal{A}, P)$.

We suppose that μ is a mixture of laws : $\mu = \sum_{j=1}^{r} p_j \mu_j$ with :

$$\forall j,\ p_j \geq 0,\ \sum_{j=1}^{r} p_j = 1 \text{ et } \mu_j \text{ is a law of probability on } \mathbb{R}^q.$$

Let F $(resp.\ F_j)$ the function of distribution of μ $(resp.\ \mu_j)$

We have : $\forall x \in \mathbb{R}^q, \quad F(x) = \sum_{j=1}^{r} p_j F_j(x)$

For $j = 1, 2, ..., r$, We suppose that p_j depends a parameter β_j and F_j of a multidimentional parameter m^j.

Let $\theta = (\beta_1, \beta_2, ..., \beta_r, m^1, m^2, ..., m^r)'$

Let p the dimension of θ and let Φ the real function of $\mathbb{R}^p \times \mathbb{R}^q$, measurable such that :

$$\Phi(x, \theta) = \sum_{j=1}^{r} \pi(\beta_j) \Psi(m^j, x)$$

We suppose that functions π et Ψ are known verifying :

$$\forall j,\ \pi(\beta_j) \geq 0,\ \sum_{j=1}^{r} \pi(\beta_j) = 1$$

$$\forall j,\ \Psi(m^j, .) \text{ is a function of distribution in } \mathbb{R}^q.$$

We wish to determine the parameter θ of $\mathbb{R}^p$ such that $\Phi(x, \theta)$ approach $F(x)$ in the least square sense.

Let $f(\theta) = E\left[(\Phi(X, \theta) - F(X))^2\right]$

We look for θ^* such that the function f is minimal for $\theta = \theta^*$.

Let Z a random variable in $\mathbb{R}^q$ of law μ and Y the indicatory function defined by :

$$Y = \mathcal{I}_Z(X) = \begin{cases} 1 & if\ x \in \mathcal{D}_z \\ 0 & else \end{cases}$$

with $\mathcal{D}_z = \{x \in \mathbb{R}^q \ : \ x \geq z\}$

and : $x' = (x^1, ..., x^q) \qquad z' = (z^1, ..., z^q)$

$$x \geq z \Leftrightarrow \forall j = 1, 2, ...q, \quad x^j \geq z^j$$

We have : $E[\mathcal{I}_Z(x)] = P(Z \leq x) = F(x)$

$$E[Y/X] = E[\mathcal{I}_Z(X)/X] = F(X)$$

Therefore $f(\theta) = E\Big[(E[Y/X] - \Phi(X, \theta))^2\Big]$

Let $g(\theta) = E[(Y - \Phi(X, \theta))^2] = E[(\mathcal{I}_Z(X) - \Phi(X, \theta))^2]$

The problem of estimation of θ^* that minimizes f becomes to look for θ^* such that g is minimal for $\theta = \theta^*$

We have : $\nabla g(\theta) = 2E\Big[\nabla_\theta \Phi(X, \theta)\Big(\Phi(X, \theta) - \mathcal{I}_Z(X)\Big)\Big]$

For estimate θ^* by séquential schem, we define the process (Θ_n) in $\mathbb{R}^p$ by :

$$\Theta_{n+1} = \Theta_n - a_n \nabla_\theta \Phi(X_n, \Theta_n)\Big(\Phi(X_n, \Theta_n) - \mathcal{I}_{Z_n}(X_n)\Big)$$

with $(X_1, Z_1), (X_2, Z_2), ..., (X_n, Z_n)$ a sample of (Z, X) formed of random independent variables, distributed identically and (a_n) is asequence of positive real numbers.

Corollary

Under H_1, H_2, H_3, H_4, H_5, we have $\Theta_n \xrightarrow{a.s.} \theta^* \quad or \quad \|\Theta_n\| \xrightarrow{a.s.} +\infty$

Proof

It is a consequence of the previous theorem. ■

3.3 Estimation of an observable function in random points.

Let h a real function of m real variables. Let an random variable X to values in $\mathbb{R}^m$. We suppose that we can observe the real random variable $h(X)$.

We have $E[h(X)/X] = h(X)$. We can approach $h(X)$ by a linear combinaison of functions $\Upsilon^i(X), i = 1, 2, ..., p$, and estimate so the function h by using the general method of the gradient, with $Y = h(X)$.

3.4 Estimation of an function $h(x) = E[Z(x)]$.

Let a family of real random variable $\{Z(x), x \in \mathbb{R}^m\}$; Let $E[Z(x)] = h(x)$ and let an random random variable X to values in $\mathbb{R}^m$. We suppose that we can observe the real random variable $Z(X)$.

We have $E[Z(X)/X] = h(X)$. We can approach $h(X)$ by a linear

combinaison of functions $\Upsilon^i(X), i = 1, 2, ..., p$, and estimate so the function h by using the general method of the gradient, with $Y = Z(X)$.

3.4 Estimation of a linear regression parameters.

The most direct application of the stochastic gradient method is the linear regression. Y is the explained random variable, $X^1, X^2, ..., X^m$ are the explanatory random variables, that constitute the variable $X \in \mathbb{R}^m$. We approach $E[Y/X^1, X^2, ..., X^m]$ by a linear combinaison of $\Upsilon^i(X)$, $i = 1, 2, ..., p$.

3.5 Estimation of baysian discriminant function.

We distinguish r classes $C_1, C_2, ..., C_r$ in a set of individuals. To classe a new individual, we measure m variables $X^1, X^2, ..., X^m$, that constitute the variable $X \in \mathbb{R}^m$.

The utilization of the ordering baysian rule requires the knowledge of probabilities to posteriori $P(C_i/X)$.

References

[1] 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).

[2] A. Bennar, J.M. Monnez, *Almost sure convergence of a stochastic Approximation process in a convex set.* International Journal of Applied Mathematics, vol. 20, n° 5, pp : 713-722 (2007).

[3] A. Bennar, A. Bouamaine, A. Namir, *Almost sure Convergence and in quadratic mean of the gradient stochastic process for the sequential estimation of a conditional expectation.* Applied Mathematical Sciences, vol. 2, no. 8, pp : 387-395, (2008).

[4] A. BOUAMAINE, *Méthodes d' approximation stochastique en Analyse des Données.* Thèse de doctorat d'etat, Université Mohamed V, (1996).

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

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

[7] D.A. FREEDMAN, L.E. DUBINS, *A sharper from of the Borel-Cantelli lemma and the strong law.* A.M.S., vol. 36, pp : 800-807.

[8] J.M. MONNEZ, *Etude d'un processus general multidimensionnel d'approximation stochastique sous contraintes convexes. Applications a l'estimation statistique.* Thèse de doctorat d'Etat es Sciences Mathématiques, Université de Nancy I, (1982).

[9] J.M. MONNEZ, *Almost Sure Convergence of Stochastic Gradient Process with Matrix Step Sizes.* Statistics and Probability Letters, 76, pp : 531-536, (2006).

[10] J.P. PARISOT,*Optimisation stochastique : le processus de KIEFFER-WOLFOFITZ, essai de synthese et quelques complements.* Thèse de Doctorat de troisième cycle, Université de Nancy 1 (1981).

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

[12] H. ROBBINS, S. MONRO, *A stochastic approximation method.* A.M.S., vol 22, pp : 400-407, (1951).

[13] J.H. VENTER, *On convergence of the Kiefer-Wolfowitz process. Approximation procedure.* A.M.S., vol. 38, pp : 1031-1036.

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

Received: November 19, 2007