Some Applications of Stochastic Gradient Processes

A. Bennar¹, A. Bouamaine² and A. Namir¹

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’électricité 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 $A$, $\lambda_{\text{min}}(B)$ 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 that, for all $\theta = (\theta_1, \theta_2, ..., \theta_p)' \in \mathbb{R}^p$,

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

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

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

$(H_4)$ $\theta^*$ is a local minimum of $g$ :

$$\exists \alpha > 0 : (x \neq \theta^*, \|x - \theta^*\| < \alpha) \Rightarrow (g(\theta^*) < g(x))$$
Applications of stochastic gradient processes

\( H_5 \) \( \theta^* \) is the unique stationary point of \( g \):
\[
\forall x \in \mathbb{R}^p, (x \neq \theta^*) \iff (\nabla_x g(x) \neq 0)
\]

**Theorem**

Under hypotheses \( H'_1, H_2, H_3, H_4, H_5 \), we have:
\[
X_n \xrightarrow{a.s.} \theta \text{ or } \|X_n\| \xrightarrow{\text{a.s.}} +\infty
\]

**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 \( \theta \).

\( H_9 \) 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 $\rho_1, \rho_2, \ldots, \rho_p$ be functions of $q$ real variables, measurable, known.

Let's put $\rho = (\rho_1, \rho_2, \ldots, \rho_p)'$ to appraise $\theta$ 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 $(U_1, V_1), (U_2, V_2), \ldots, (U_n, V_n)$ is a sample of $(U, V)$ formed of independent random variables and distributed identically.

Let's make hypotheses

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

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

$(H_8)$ $X_1$ is a 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 $\phi$ 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 $j = 1, 2, \ldots, p$, $\frac{\partial \phi(V, x)}{\partial x_j} = \rho_j(V)$, 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 invertible. Then:

$\theta^*$ 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
\[ g(x) = E[U^2] + \|x\|^2 - 2 < x, \theta^* > = E[U^2] + \|x - \theta^*\|^2_A - \|\theta^*\|^2_A \]
\[ \geq c\|x - \theta^*\|^2_A + d \quad (c = \lambda_{\text{min}(A)} \text{ and } d = E[U^2] - \|\theta^*\|^2_A) \]
\[ \geq \frac{1}{2}c\|x\|^2 - c\|x\|^2 + d \geq e\|x\|^2 + f \quad (e = \frac{c}{2} \text{ and } f = d - c\|x\|^2) \]

Therefore, for $i = 1, 2, ..., p$, we have:
\[ \text{Var} \left[ \frac{\partial \phi(V, x)}{\partial x_i} (\phi(V, x) - U) \right] = \text{Var} \left[ \rho_i(V)(x'\rho(V) - U) \right] \]
\[ \leq E[\rho_i^2(V)(x'\rho(V) - U)^2] \]
\[ \leq a\|x\|^2 + b \quad (a = 2E[\rho_i^2(V)\|\rho(V)\|^2], \ b = 2E[\rho_i^2(V)U^2]) \]
\[ \leq Ag(x) + B \quad (A = \frac{1}{e}, \ B = b - \frac{af}{e}) \]

Let’s prove that the hypothesis $H_3$ is verified.

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

Therefore: for $i, j = 1, 2, ..., p$, we have \[ \frac{\partial^2 g(x)}{\partial x_i \partial x_j} = 2E[\rho_j(V)\rho_i(V)], \text{ that doesn’t depend of } x. \]

Hypotheses of the theorem 2.1 are verified, therefore:
\[ X_n \xrightarrow{a.s.} \theta^* \quad \text{or} \quad \|X_n\| \xrightarrow{a.s.} +\infty \]

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

Indeed: as \[ \sum_{i=1}^{\infty} a_n \| \nabla_x g(X_n) \|^2 < \infty \ a.s. \ (\text{see } [3] \text{ Lemma 2.1}) \]
\[ \sum_{i=1}^{\infty} a_n = +\infty, \text{ there exists an sub-sequence of integers } (n_i) \text{ such that} \]
\[ \| \nabla_x g(X_{n_i}) \| \xrightarrow{a.s.} 0 \]

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

Therefore: \[ \| \nabla_x g(X_n) \|^2 \geq 4\lambda_{\text{min}(A)}^2\|X_n - \theta^*\|^2 \quad (\lambda_{\text{min}(A)} > 0) \]

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 \( \mu \) and to values in \( \mathbb{R}^q \) defined on a probability space \( (\Omega, \mathcal{A}, P) \).

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

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

Let \( F \) (resp. \( F_j \)) the function of distribution of \( \mu \) (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 \( \beta_j \) and \( F_j \) of a multidimensional parameter \( m_j \).

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

Let \( p \) the dimension of \( \theta \) and let \( \Phi \) 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 \( \pi \) et \( \Psi \) are known verifying :

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

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

We wish to determine the parameter \( \theta \) 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 \( \theta^* \) such that the function \( f \) is minimal for \( \theta = \theta^* \).

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

\[
Y = I_Z(X) = \begin{cases} 
1 & \text{if } x \in \mathcal{D}_z \\
0 & \text{else}
\end{cases}
\]
with \( D_z = \{ x \in \mathbb{R}^q : x \geq z \} \) and:
\[
x' = (x^1, \ldots, x^q) \quad z' = (z^1, \ldots, z^q)
\]
\( x \geq z \iff \forall j = 1, 2, \ldots, q, \ x^j \geq z^j \)

We have:
\[
E[I_Z(x)] = P(Z \leq x) = F(x)
\]
\[
E[Y/X] = E[I_Z(X)/X] = F(X)
\]
Therefore
\[
f(\theta) = E\left[ (E[Y/X] - \Phi(X, \theta))^2 \right]
\]
Let
\[
g(\theta) = E[(Y - \Phi(X, \theta))^2] = E[(I_Z(X) - \Phi(X, \theta))^2]
\]
The problem of estimation of \( \theta^* \) that minimizes \( f \) becomes to look for \( \theta^* \) such that \( g \) is minimal for \( \theta = \theta^* \)

We have:
\[
\nabla g(\theta) = 2E\left[ \nabla_\theta \Phi(X, \theta) \left( \Phi(X, \theta) - I_Z(X) \right) \right]
\]
For estimate \( \theta^* \) by sequential scheme, we define the process \( (\Theta_n) \) in \( \mathbb{R}^p \) by:
\[
\Theta_{n+1} = \Theta_n - a_n \nabla_\theta \Phi(X_n, \Theta_n) \left( \Phi(X_n, \Theta_n) - I_Z(X_n) \right)
\]
with \( (X_1, Z_1), (X_2, Z_2), \ldots, (X_n, Z_n) \) a sample of \( (Z, X) \) formed of random independent variables, distributed identically and \( (a_n) \) is a sequence of positive real numbers.

**Corollary**

Under \( H_1, H_2, H_3, H_4, H_5 \), we have \( \Theta_n \overset{a.s.}{\longrightarrow} \theta^* \text{ or } \| \Theta_n \| \overset{a.s.}{\longrightarrow} +\infty \)

**Proof**

It is a consequence of the previous theorem. \( \blacksquare \)

### 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 combination of functions \( \Upsilon^i(X), i = 1, 2, \ldots, p \), 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


Received: November 19, 2007