SAINT PETERSBURG STATE UNIVERSITY

Mathematical methods in stochastic simulation and experimental design

Proceedings of the 2nd St.Petersburg Workshop on simulation, St.Petersburg, June 18-21, 1996

Edited by S.M. Ermakov and V.B. Melas

SAINT PETERSBURG
PUBLISHING HOUSE OF SAINT PETERSBURG UNIVERSITY
1996

Asymptotic-optimality stochastic approximation algorithms with perturbation on the input

Oleg N.Granitchine, Alexander Portnov
St.Petersburg State University

Abstract

A stochastic approximation problem is considered in the situation when the unknown regression function is measured not at the previous estimate but at its slightly exited position. Two estimation algorithms for estimate the root and the minimum point of regression function with projection is proposed. It is shown that the sequence of estimates $\{x_n\}$ obtained converges to the true value θ as sure and in the mean square sense. Sequence of estimates has asymptotic normality distribution. The new adaptive Robbins-Monro algorithm is proposed. It is provide minimum of variance of asymptotic distribution $(x_n - \theta)$.

Key words: stochastic approximation, consistency estimates, regression function, conditional mean value, perturbation noise.

1 Introduction

The main ideas of stochastic approximation was formulated by Robbins and Monro [1]. Let $x \in \mathbf{R}$ is "controllable" variable and for each x we can measure random value $Y(x)$ with distribution $P(Y(x) < y) = F_x(y)$. Let $M(x)$ is regression function $Y(x)$ on x:

$$M(x) = \int_{-\infty}^{+\infty} Y(x) dF_x(y).$$

Robbins and Monro investigated problem of finding θ unique root of regression equation $M(x) = 0$. Let $M(x)$ is increasing function. Consider recurrent sequence

$$x_{n+1} = x_n - a_n Y_n,$$

where $\{a_n\}$ are positive number and Y_n is result of measurement by x_n with distribution F_{x_n}. Robbins and Monro shown convergence x_n to θ as sure with natural proposals about $\{F_{x_n}\}$ и $\{a_n\}$. In this algorithm one considers statistic

$$z_0 = Y(x)$$

for estimate value of regression function $M(x)$. This is consistency estimate of $M(x)$

$$E_x\{z_0\} = M(x),$$

$E_x\{...\}$ conditional mean value.

The problem regression function minimum point estimation was considered by Kiefer and Wolfowits in [2]. The main idea was to solve the equation $M'(x) = 0$. For this purpose one can measure values Y in two points $x + c, x - c$ and for derivation of regression function in point x one can use statistic

$$z_1 = \frac{1}{2c}(Y(x+c) - Y(x-c)).$$

For the estimation of minimum point Kiefer and Wolfowits proposed algorithm

$$x_{n+1} = x_n - a_n z_{1_n},$$

where $\{a_n\}, \{c_n\}$ some sequences of positive numbers.

The performance of stochastic approximation algorithms depends from accuracy of estimate $M(x)$ or $M'(x)$. If we change statistics z_0 and z_1 for another we can hope to get better performance. For this purpose we need in better approximation $M(x)$ or $M'(x)$, Fabian [3] modified Kiefer and Wolfowits algorithm. They proposed to use difference approximations of high derivations with some weight. If l is number of continuous derivation of regression function $M(x)$, then Fabian's algorithm provides mean square rate of convergence as $O(n^{-\frac{l-1}{2}})$ for odd l. From computational point of view Fabian's algorithm is very complicated.

In this paper for estimation $M^{(0)}(x) = M(x)$ or $M^{(1)}(x) = M'(x)$ in Robbins-Monro or Kiefer-Wolfowits algorithms we propose to use statistics Z_0 и Z_1

$$Z_{r_n} = h_n^{-r} K_r(\zeta_n) Y(x_n + h_n\zeta_n), r = 0, 1,$$

instead z_0 or z_1. Here $\{h_n\}$ is some sequence of positive number, $\{\zeta_n\}$ is sequence of independent random values, $K_0(\zeta), K_1(\zeta)$ are some kernel function on $\mathbf{R}$. For some $\{h_n\}, \{\zeta_n\}, K_r(\zeta), r = 0, 1$ estimates Z_{r_n} are consistency estimates of $M^{(r)}(x)$:

$$E_{x_n}\{K_r(\zeta_n) h_n^{-r} Y(x_n + h_n\zeta_n\} = M^{(r)}(x_n) + o(h_n^{p-r}), r = 0, 1.$$

Here p is some index of regression function $M(x)$ (in particular, $p = l$ if $l \in \mathbf{N}$ and $M(x)$ has $l-1$ derivations satisfies Lipchits conditions). Such a stochastic approximation algorithm with additive perturbation noises $h_n \zeta_n$ has recently been investigated by Polyak and Thybakov [4] for independent measurement noises $Y_n - M(x_n + h_n \zeta_n)$ and by Granitchine [5],[6],[7] for some special cases with dependent measurement noises. Polyak and Tsybakov have shown that this algorithm had an optimum minimax rate of convergence in wide variety of algorithms. This algorithm has mean square rate of convergence $O(n^{-\frac{p-1}{p}})$.

We propose to consider two new algorithms as Robbins-Monro and Kiefer-Wolfowits. This algorithms provide convergence as sure with high rate and asymptotic normality of distribution of random values $x_n - \theta$. We calculate the asymptotic variance of $x_n - \theta$. This approach leads to way of optimal choosing of kernel function $K_r(\zeta)$. The last section deal with new adaptive Robbins-Monro algorithm, which every time use $K_0(\zeta)$ and $K_1(\zeta)$ for estimation function regression root θ and $M^{(1)}(\theta)$. This algorithm has optimal perfomance: minimum variance of asymptotic normal distribution $x_n - \theta$.

2 Differentiation kernel

Let $\{p_m(u)\}$ is some system of orthogonal polinoms on some interval $[-\gamma, \gamma]$ with degree below l and weight function $\psi(u) \geq 0, \gamma > 0$. Then

$$\int_{-\gamma}^{\gamma} \psi(u) p_i(u) p_j(u) du = a_i \delta_{i,j}, \int_{-\gamma}^{\gamma} \psi(u) du = 1, \tag{1}$$

for $i, j = 1, ..., l$, where $\delta_{i,j}$ is equal 1, if $i = j$, and 0 if $i \neq j$, $a_i = \int_{-\gamma}^{\gamma} \psi(u) p_i^2(u) du$ is some constants.

Define the functions $K_r(u), r = 0, 1$ on interval $[-\gamma, \gamma]$ as linear combination of polinoms $p_m, m = 1, ..., l$

$$K_r(u) = \sum_{m=0}^{l} \frac{p_m^{(r)}(0)}{a_m} p_m(u). \tag{2}$$

We can see

$$\int_{-\gamma}^{\gamma} \psi(u) K_r(u) u^q du = \delta_{q,r}, \tag{3}$$

for any $q \in \mathbf{Z}, q \leq l$.

Let function f has l times continuous derivations near point x_0 on $\mathbf{R}$. We have

$$f(x_0 + cu) = \sum_{i=0}^{l} \frac{f^{(i)}(x_0)}{i!} (cu)^i + o(u^l).$$

Consider integral representation of function f with kernel K_r

$$\frac{1}{c^2} < f(x_0+cu), K_r(u) >= \frac{1}{c^2}\int_{-\gamma}^{\gamma}\psi(u)K_r(u)f(x_0+cu)du \qquad (4)$$

We can obtain

$$\frac{1}{c^2} < f(x_0+cu), K_0(u) >= f(x_0) + \int_{-\gamma}^{\gamma}\psi(u)K_0(u)o(u^l)du, \qquad (5)$$

$$\frac{1}{c^2} < f(x_0+cu), K_1(u) >= f^{(1)}(x_0) + \int_{-\gamma}^{\gamma}\psi(u)\frac{K_1(u)}{c}o(u^l)du. \qquad (6)$$

Equations 5 and 6 shows the main idea of new stochastic approximation algorithms listed below.

Note

$$K_0(0) = \int_{-\gamma}^{\gamma}\psi(u)K_0(u)^2du, \qquad (7)$$

$$K_1^{(1)}(0) = \int_{-\gamma}^{\gamma}\psi(u)K_1(u)^2du. \qquad (8)$$

We can use Legendre's or Chebuchev's polinoms to build kernel functions $K_r(u), r=0,1,$ for example. The values $K_0(0)$ and $K_1'(0)$ have importance role in calculation variance of asymptotic distribution $x_n-\theta$ We have for Legendre's polinoms

$$K_0(0) = \sum_{m=0}^{[\frac{l+1}{2}]}[\frac{(2m-1)!!}{2m!!}]^2(4m+1),$$

$$K_1^{(1)}(0) = \frac{1}{\gamma^2}\sum_{m=0}^{[\frac{l-1}{2}]}(4m+3)(1+\frac{1}{2})^2(1+\frac{1}{4})^2...(1+\frac{1}{2m})^2,$$

and for Chebuchev's polinoms

$$K_0(0) = 1+\frac{1}{2}[\frac{l+1}{2}], K_1'(0) = \frac{1}{\gamma^2}2([\frac{l-1}{2}]+1)^2(\frac{3}{4}[\frac{l-1}{2}]([\frac{l-1}{2}]+2)+1),$$

[...] is entire function.

3 Convergence and asymptotic normality

Let all random values define on some fixed probability space (Ω, F, P).

Let regression function $M(x)$ define on some compact set $\Theta \in \mathbf{R^N}$. It has l times continuous derivations on Θ and $M^l(x)$ which satisfy Hoelder's conditions with some constant $\alpha, 0 < \alpha \le 1$ so that

$$M(x_0 + t) = \sum_{m=0}^{l} \frac{M^{(m)}(x_0)}{m!} t^m + o(|t|^p|), \tag{9}$$

where $p = l + \alpha$.

Т е о р е м а 1 *Let random sequences $\{x_n\}$ is "own" design of an experiment and $\{Y_n\}$ is result of measurements(or observations), $E\{x_1\} < \infty$, $\{\zeta_n\}$ is perturbation noises, the sequence of independent random values with same distributions on some interval $[-\gamma, \gamma](0 < \gamma < \infty)$ with distribution density $\psi(u)$, h and a are some positive constants, the real design of an experiment is determined by summa $x_n + \frac{h}{n^{\frac{1}{2p}}}\zeta_n$ and $E_{F_n}\{Y_n\} = M(x_n + \frac{h}{n^{\frac{1}{2p}}}\zeta_n)$; $F_N = \sigma\{x_1, Y_1 - M(x_1 + \frac{h}{1}\zeta_1), ..., Y_{N-1} - M(x_{N-1} + \frac{h}{(N-1)^{\frac{1}{2p}}}\zeta_{N-1}), \zeta_1, ..., \zeta_{N-1}\}$ is σ-algebra, random values ζ_N and $Y_1 - M(x_1 + \frac{h}{1}\zeta_1), ..., Y_{N-1} - M(x_{N-1} + \frac{h}{(N-1)^{\frac{1}{2p}}}\zeta_{N-1})$ are independent,$N = 1, 2, ...$, measurement noises are satisfied*

$$E\{(Y_n - M(x_n + \frac{h}{n^{\frac{1}{2p}}}\zeta_n))^2\} \le \sigma^2, E\{(Y_n - M(x_n + \frac{h}{n^{\frac{1}{2p}}}\zeta_n))^2\} \to \sigma^2(\theta)$$

as $x_n \to \theta$, there is some positive constant $\lambda > 0$ so that for any $q > 0$

$$E\{(Y_n - M(x_n + \frac{h}{n^{\frac{1}{2p}}}\zeta_n))^2 \mathbf{1}_{\{(Y_n - M(x_n + \frac{h}{n^{\frac{1}{2p}}}\zeta_n))^2 \ge qn^\lambda\}}\} \to 0$$

as $n \to \infty$,($\mathbf{1}_{\{...\}}$ is indicator function).

For this conditions we have

1) If $l \ge 0$, regression equation $M(x) = 0$ has the unique root on Θ in the point θ,

$$M^{(1)}(\theta) > \frac{1}{2a} \tag{10}$$

there is $B > 0, D > 0$ so that for any $x \in \mathbf{R}$

$$|M(x)| \le B + D|x| \tag{11}$$

for any positive ϵ

$$inf_{\epsilon < |x-\theta| < \epsilon^{-1}}\{sign((x - \theta)M(x))\} > 0, \tag{12}$$

then estimates $\{x_n\}$ *which formed by*

$$x_{n+1} = P_{\Theta}\{x_n - \frac{a}{n}K_0(\zeta_n)Y_n\} \tag{13}$$

($P_{\Theta}\{...\}$ *is projection operator) satisfy convergence* $x_n \to \theta$ *as sure and random value* $(x_n - \theta)n^{\frac{1}{2}}$ *has asymptotically normality distribution with mean value* 0 *and variance*

$$\frac{a^2\sigma^2(\theta)K_0(0)}{2aM^{(1)}(\theta) - 1}, \tag{14}$$

2) If $l \geq 1$, *regression function* $M(x)$ *has the unique minimum point on* Θ *in the point* θ,

$$M^{(2)}(\theta) > \frac{p-1}{2pa} \tag{15}$$

there is $B' > 0, D' > 0$ *so that for any* $x', x" \in \mathbf{R}$

$$|M^{(1)}(x') - M^{(1)}(x")| \leq D'|x' - x"|, |M^{(1)}(\theta)| \leq B' \tag{16}$$

then estimates $\{x_n\}$ *which formed by*

$$x_{n+1} = P_{\Theta}\{x_n - \frac{a}{n}\frac{h}{n^{\frac{1}{2p}}}K_1(\zeta_n)Y_n\} \tag{17}$$

satisfy convergence $x_n \to \theta$ *as sure and random value* $(x_n - \theta)n^{\frac{p-1}{2p}}$ *has asymptotically normality distribution with mean value* 0 *and variance*

$$\frac{a^2(\sigma^2(\theta) + M(\theta))K_1^{(1)}(0)}{h^2(2aM^{(2)}(\theta) - \frac{p-1}{p})}. \tag{18}$$

4 Adaptive storage of Robbins-Monro algorithm

Т е о р е м а 2 *Let all conditions of part 1 theorem 3 are hold.*

If θ *is the unique root of regression equation* $M(x) = 0$ *on* Θ *and there is two positive constants* $s^+ > s^- > 0$ *so such*

$$s^- \leq M^{(1)}(\theta) \leq s^+, \tag{19}$$

$P_S\{...\}$ *is projection operator on set* $S = [s^-, s^+]$,

then estimates $\{x_n\}$ *which formed by*

$$x_{n+1} = P_{\Theta}\{x_n - \frac{1}{n}\frac{1}{s_n}K_0(\zeta_n)Y_n\} \tag{20}$$

where $\{s_n\}$ *is sequence of random values*

$$s_{n+1} = P_S\{\frac{1}{n}\sum_{i=1}^{n}\frac{i^{\frac{1}{2p}}}{h}K_1(\zeta_i)Y_i\}, \tag{21}$$

satisfy convergence $x_n \to \theta$ *as sure and* $s_n \to M^{(1)}(\theta)$ *as sure, random value* $(x_n-\theta)n^{\frac{1}{2}}$ *has asymptotically normality distribution with mean value* 0 *and variance*

$$\frac{\sigma^2(\theta)K_0(0)}{(M^{(1)}(\theta))^2} \tag{22}$$

and random value $(s_n - M^{(1)}(\theta))n^{\frac{p-1}{2p}}$ *has asymptotically normality distribution with mean value* 0 *and variance*

$$\frac{p\sigma^2(\theta)K_1^{(1)}(0)}{p+1}. \tag{23}$$

Note, expressions 22 and 23 are minimum of possible in wide range of similar algorithms. In accordance expressions 14(22) and 18(23) we have one way to choice kernels $K_r(u), r = 0, 1$. For example we can calculate variance for Legendre's and Chebushev's polinoms or we can study dependence between γ and variance of asymptotically distribution.

References

[1] Robbins H., Monro S. A stochastic approximation method. *Ann. Math. Statist. 22*,pp. 400-407, 1951.

[2] Kiefer J., Wolfowits J. Stochastic estimation of the maximum of a regression function. *Ann. Math. Statist. 23*,pp. 452-466, 1952.

[3] Fabian V., Stochastic approximation of minima with improved asymptotic speed. *Ann. Math. Statist. 38*,pp. 191-200, 1967.

[4] Polyak B.T., Tsybakov A.B. Stochastic estimation of the maximum of a regression function. *Problems Inform. Transmission 26*, pp. 126-133, 1990.

[5] Granitchin O.N. Stochastic Approximation with sample perturbations in the input *Automat. Remote Control 53*, pp. 97-104, 1992.

[6] Granitchin O.N. Unknown function minimum point estimation under dependent noises. *Problems Inform. Transmission 28*, pp. -, 1992.

[7] Granitchin O.N. Stochastic Approximation under Dependent Noises, Detecting Signals and Adaptive Control *Approximation, Probability, and Related Fields*, pp. 247-271, Plenum, 1994.