Sufficient And Necessary Condition For The Convergence Of Stochastic Approximation Algorithms

Jianfeng Feng* Wenbin Liu**
*COGS, Sussex University, Brighton BN1 9QH, UK
**CBS, Kent University, UK

June 14, 2005

Abstract

We show sufficient and necessary condition for the convergence of stochastic approximation algorithms, which were proposed 50 years ago and have been widely applied to various areas and been theoretically intensively investigated as well. In the literature, only various sufficient conditions are known. The condition is simple and has a clear physical meaning.

1 Introduction

In stochastic approximation theory [2, 12, 15, 19] the convergence of the following SDE

$$dx_t = \eta(t)(-\nabla H(x_t)dt + \beta(t)dB_t) \tag{1}$$

on a probability space $(\Omega, \mathcal{F}, \mathcal{P})$ is considered where $(\eta(t), \beta(t))_{t>0}$ are both differentiable and positive functions, B_t is the N-dimensional Brownian motion and $H \in C^1(I\!R^N)$. Under certain conditions on H and

$$\int_0^\infty \eta(t)dt = \infty \tag{2}$$

$$\int_0^\infty \eta^2(t)\beta^2(t)dt < \infty \tag{3}$$

it has been proved that x_t converges to the set of all local minima of H (see for example Theorem 8.2, page 62 in [15]). The stochastic approximation theory was first proposed by Robbins and Monro [16] in around 50 years ago. The procedure defined by Eq. (1) has been successfully applied to very widespread varied applications as system identification, adaptive control, transmission systems, adaptive filtering for signal processing, several aspects of pattern recognition and most recently as neural networks (see monographs [2, 9, 12, 14, 15, 18]). During the past 50 years, various conditions on $(\eta(t), \beta(t))$ have been found to ensure the convergence of the stochastic algorithm to the set of all local minima of H. Nevertheless, the fundamental issue remains open: *what is the sufficient and necessary condition on $(\eta(t), \beta(t))$ to ensure the convergence of the stochastic algorithm defined by Eq. (1) to the set of all local minima?*

We answer the question in current paper, based upon an idea developed in [6]. It is not easy to directly deal with x_t since both the deterministic term and the noise term depend on time t. If, after some transformations, we could simplify the equation of x_t, and we know the relationship between the solution of the simpler version and x_t, we could possibly gain new insights onto the issue. More precisely, we apply the existing results of large deviation theory to our case by viewing the SDE x_t at a different time scaling so that the term $\eta(t)$ before the drift term $(-\nabla H)$ disappears but there is still a vanishing term before the Brownian motion due to its self-similarity property.

It is worthwhile to point out that Eq. (1) is a generalization of the standard simulated annealing which considers the convergence of the SDE

$$dz_t = (-\nabla H(z_t)dt + \beta(t)dB_t). \tag{4}$$

Hence results on ensuing x_t to converge to the set of global minima of H are also included.

On the other hand the stochastic approximation algorithm has been widely used to find the solution of the equation $R(x) = 0$, i.e.

$$dy_t = \eta(t)(R(x_t)dt + \beta(t)dB_t) \tag{5}$$

where $R(x)$ is, in general, not the gradient of a function. Can we introduce a proper definition of an 'optimal' solution of Eq. (5) and ensure that y_t converges to the 'optimal' solution? The question is answered as well.

Suppose that $H \in C^2(I\!R^N)$, there exist $K, c > 0$ with $|\frac{\partial H}{\partial x_i}| > c \ \forall i$ as $|x| > K$, x_t is nonexplosive and there are finitely many isolated local minima for K as $|x| < K$. The similar assumptions are true for R with finitely many isolated attractors, $R \in C^1(I\!R^N, I\!R^N)$.

2 Convergence To Local Minima

2.1 With Gradient

Denote

$$\mathcal{A}_H = \{x \in I\!R^N, \nabla H = 0\}$$

and assume that $\lim_{t\to\infty} \beta(t)\sqrt{\eta(t)}$ exists.

Theorem 1 *For all $\epsilon > 0$ and $x \in I\!R^N$,*

$$\lim_{t\to\infty} P_{0,x}(x_t \text{ in the } \epsilon-\text{neighborhood of } \mathcal{A}_H\ ) = 1 \tag{6}$$

if and only if

$$\int_0^\infty \eta(u)du = \infty \tag{7}$$

and

$$\lim_{t\to\infty} \beta(t)\sqrt{\eta(t)} = 0 \tag{8}$$

Before proving, let us point out that condition (8) is considerably weaker than condition (3). For example, when $\beta(t) = 1$ and $\eta(t) = 1/\sqrt{t}$, condition (3) is not satisfied, but condition (8) is.

Proof We first prove that conditions (7) and (8) are sufficient. Let

$$s = s(t) = \int_0^t \eta(u)du \tag{9}$$

then $s(t)$ is a strictly increasing function of t and

$$\lim_{t\to\infty} s(t) = \int_0^\infty \eta(u)du = \infty \tag{10}$$

Denote its inverse function as $t = t(s)$ which is a strictly increasing function of s and

$$\lim_{s\to\infty} t(s) = \infty \tag{11}$$

since $\eta(t)$ is a positive function of t.

Define

$$v_s = x_{t(s)} \tag{12}$$

we then have the following relationship

$$\lim_{s\to\infty} P(v_s \in \mathcal{B}) = \lim_{s\to\infty} P(x_{t(s)} \in \mathcal{B}) = \lim_{t\to\infty} P(x_t \in \mathcal{B}) \tag{13}$$

for any measurable set $\mathcal{B}$.

Hence equation (1) can now be written as

$$\begin{aligned} dv_s &= dx_t \\ &= -\nabla H(v_s)\eta(t)dt + \sqrt{\eta(t)}\beta(t)\sqrt{\eta(t)}dB_t \\ &= -\nabla H(v_s)ds + \sqrt{\eta(t)}\beta(t)\sqrt{\eta(t)}dB_t \end{aligned} \tag{14}$$

In terms of the self-similarity property of the Brownian and $\eta(t)dt = ds$, we derive that

$$d\tilde{B}_s := \sqrt{\eta(t(s))}dB_{t(s)} \sim N(0, ds \cdot I) \tag{15}$$

where matrix $I = (\delta_{ij}, i, j = 1, \cdots, N)$. Therefore $\tilde{B}_s$ is still a standard Brownian on $I\!R^N$ and

$$dv_s = -\nabla H(v_s)ds + \sqrt{\eta(t(s))}\beta(t(s))d\tilde{B}_s \tag{16}$$

From results in [7] it is easily seen

$$\begin{gathered} \lim_{t\to\infty} P_{0,x}(x_t \in \text{the } \epsilon-\text{neighborhood of } \mathcal{A}_H) = \\ \lim_{s\to\infty} P_{0,x}(v_s \in \text{ the } \epsilon-\text{neighborhood of } \mathcal{A}_H) = 1. \end{gathered}$$

Now we show that conditions (7) and (8) are necessary. There are only two possibilities: (I) the time s tends to infinity but the noise term remain positive, (II) the time s does not even tend to infinity and so the noise term is certainly positive. First assume that condition (8) is not true, corresponding to the case (I), but condition (7) holds. We have

$$\lim_{t\to\infty} \sqrt{\eta(t)}\beta(t) > 0$$

which implies that

$$\lim_{s\to\infty} \sqrt{\eta(t(s))}\beta(t(s)) = s_0 > 0$$

and therefore v_s defined by Eq. (16) is ergodic within any compact set of $I\!R^N$. We thus conclude that

$$\lim_{t\to\infty} P_{0,x}(x_t \in \mathcal{B}) = \lim_{s\to\infty} P_{0,x}(v_s \in \mathcal{B}) > 0$$

for any compact set $\mathcal{B} \in I\!R^N$.

If condition (7) is not true, corresponding to the case (II), i.e.

$$t_0 = \int_0^\infty \eta(t)dt < \infty$$

we see that

$$\begin{aligned} \lim_{t\to\infty} P_{0,x}(x_t \in \mathcal{B}) &= \lim_{s\to\int_0^\infty \eta(t)dt<\infty} P_{0,x}(v_s \in \mathcal{B}) \qquad (17) \\ &= \int_{\mathcal{B}} p(x,0;y,t_0)dy > 0 \qquad (18) \end{aligned}$$

for any compact set $\mathcal{B} \in I\!R^N$ and $p(x,0;y,t)$ is the transition probability density at time t.

Now we see that the physical meaning of conditions (7) and (8) are very clear: condition (7) ensures that the time of the dynamics tends to infinity, condition (8) ensures that noise gradually vanishes.

2.2 Without Gradient

Stochastic approximation algorithm has another application: to find the solution of the equation $R(x) = 0$,

$$dy_t = \eta(t)(R(y_t)dt + \beta(t)dB_t) \qquad (19)$$

where R may not be a gradient of a function. It is obviously seen that we have the following conclusions.

Theorem 2 *For all $\epsilon > 0$ and $y \in I\!R^N$,*

$$\lim_{t\to\infty} P_{0,y}(y_t \textit{ in the } \epsilon-\textit{neighborhood of } \mathcal{A}_R\ ) = 1 \qquad (20)$$

if and only if

$$\int_0^\infty \eta(u)du = \infty \tag{21}$$

and

$$\lim_{t\to\infty} \beta(t)\sqrt{\eta(t)} = 0 \tag{22}$$

where $\mathcal{A}_R = \{x : R(x) = 0\}$

3 Convergence To Global Minima

3.1 With Gradient

First of all we present a counter example to demonstrate that condition (3) is too loose for the process x_t to converge to the global minima of H.

Example 1 Let $\eta(t) = \frac{1}{t}, \beta(t) = 1$ and so conditions (2) and (3) are both fulfilled. For simplicity we assume that $N = 1$. Define $t = e^s$ or $s = \log(t)$ and

$$u_s = y_{e^s} \tag{23}$$

Eq. (1) becomes

$$du_s = -\frac{1}{e^s}H'(u_s)e^s ds + \frac{1}{\sqrt{e^s}} \cdot \frac{1}{\sqrt{e^s}} dB_{e^s} \tag{24}$$

Since

$$\frac{1}{\sqrt{e^s}} dB_{e^s} \sim N(0, \frac{1}{e^s}e^s ds) = N(0, ds) \tag{25}$$

we denote

$$d\tilde{B}_s = \frac{1}{\sqrt{e^s}} dB_{e^s} \tag{26}$$

which is again a standard Brownian motion. Therefore

$$du_s = -H'(u_s)ds + \frac{1}{\sqrt{e^s}} d\tilde{B}_s \tag{27}$$

For the SDE above it is well-known that u_s will be trapped at some local minima with a positive probability since the noise vanishes too fast at a rate of $e^{-s/2}$[1, 3, 4, 10, 11, 17].

The following theorem shows which kind of function $(\eta(t), \beta(t))_{t\geq 0}$ we can choose to ensure the convergence of x_t to the global minima of H. Let γ be a constant which is larger than the critical value γ_0 in the simulated annealing for function H (see [1]). Denote $\mathcal{B}_H$ as the set of global minima of H. We then have

Theorem 3 *For all $\epsilon > 0$ and $x \in I\!R^N$,*

$$\lim_{t\to\infty} P_{0,x}(x_t \text{ in the } \epsilon - \text{neighbourhood of } \mathcal{B}_H) = 1$$

where

$$\begin{cases} \int_0^\infty \eta(u)du &= \infty \\ \beta(t) &= \dfrac{\gamma}{\sqrt{\eta(t)\log(\int_0^t \eta(u)du + 2)}} \end{cases} \tag{28}$$

In particular, when $\beta(t) = 1$, $\eta(t)$ is the unique solution of the following equation

$$\eta'(t) = -\eta^3(t)/(\gamma^2 \int_0^t \eta(u)du + 2)$$

with initial condition $\eta(0) = \gamma^2/(\log 2)$

It is easily seen that condition (28) is stronger than condition (8).

In fact, the theorem is proved in [6], for the completeness of our paper we include it here.

3.2 Without Gradient

First of all let us define the meaning of convergence to global minima of

$$dx_t = \eta(t)(R(x_t)dt + \beta(t)dB_t) \tag{29}$$

Assume that $A_i, i = 1\cdots, n$ are all attractors of $dx/dt = R(x)$. For two attractors A_i, A_j, the action functional is defined by

$$T(A_i, A_j) = \inf_\phi\{S_{0t}(\phi) : \phi \in C^1([0,t], I\!R^N), \phi_0 \in A_i, \tag{30}$$

$$\phi_t \in A_j, t > 0\} \tag{31}$$

and (see for example, Eq. (3) in [5])

$$S_{0t}(\phi) = \frac{1}{2}\int_0^t ||\dot{\phi}_s - R(\phi_s)||^2 ds$$

When $\eta(t) = 1$ and $\beta(t) = \sigma$, a positive constant, intuitively, the action functional of a path ϕ measures the difficulty for the system to go along the path ϕ from A_i to A_j and the conditional probability of such an event is

$$p(A_j|A_i)|_\phi \sim \exp(-S_{0t}(\phi)/(2\sigma^2))$$

Therefore the larger the action functional of a path, the more difficult for the system to go along it from A_i to A_j. Furthermore, since the probability is exponentially proportional to σ^2, the system will always find paths which minimize the action functional to go from A_i to A_j, provided that σ^2 is small enough. In other words, with probability one, the system goes along the optimal paths from A_1 to A_2 and

$$p(A_j|A_i) \sim \exp(-T(A_i, A_j)/(2\sigma^2))$$

When $R = -\nabla H$, $T(A_i, A_j)$ is the height of the lowest energy barriers between two attractors, as discussed in the previous subsection. Define

$$T(A_i) = \min_{j=1,\cdots,n, j\neq i}\{T(A_i, A_j)\}$$

and

$$\mathcal{B}_R = \{A_i, T(A_i) = \max\{T(A_j), j = 1, \cdots, n\}\}$$

we then have the following theorem.

Theorem 4 *For all $\epsilon > 0$ and $x \in I\!R^N$,*

$$\lim_{t\to\infty} P_{0,x}(x_t \text{ in the } \epsilon - \text{neighbourhood of } \mathcal{B}_R) = 1$$

where

$$\begin{cases} \int_0^\infty \eta(u)du &= \infty \\ \beta(t) &= \dfrac{\gamma}{\sqrt{\eta(t)\log(\int_0^t \eta(u)du + 2)}} \end{cases} \tag{32}$$

In particular, when $\beta(t) = 1$, $\eta(t)$ is the unique solution of the following equation

$$\eta'(t) = -\eta^3(t)/(\gamma^2\int_0^t \eta(u)du + 2)$$

with initial condition $\eta(0) = \gamma^2/(\log 2)$

4 Discussion

We have presented a sufficient and necessary condition for the convergence of the stochastic approximation algorithm which was proposed about 50 years ago. In the literature, only various sufficient conditions are known. We have also considered the condition to ensure the algorithm converges to the global minima. Finally we have extended the results to a dynamical system without gradient.

A dynamical systems with a random perturbation has been the central theme for mathematics and physics for centuries. However, majority, if not all, studies are concentrated on a dynamical system with a fixed potential. Certainly many problems arising from practical applications have a potential depending on time. Our results here could be seen as one step forward to investigate such a system: the potential depends on time, but in a simple way, i.e. taking the form of $\eta(t)H$.

References

[1] Albeverio, S., Feng, J. , and Qian, M. (1995). Role of noises in neural networks, *Physical Review E*, **52**, 6593-6606.

[2] Benveniste, A., Métivier, M. & Priouret, P.(1990), *Adaptive algorithms and stochastic approximations*, Berlin: Springer-Verlag.

[3] Catoni, O.(1992), *Rough large deviation estimates for simulated annealing: application to exponential schedules*, Ann. Prob. **20** pp.1109-1146.

[4] Chiang, T.S. & Chow, Y.(1988), On the convergence rate of annealing processes, *SIAM J. Control Optim.* **26** pp.1455-1470.

[5] Smelyanskiy V.N., and Dykman M.I. (1997), Optimal control of large fluctuations *Phys. Rev. E.* **55**: 2516-2521.

[6] Feng J., Georgii H.-O., and Brown D. (2000) Convergence to global minima for a class of diffusion processes *Physica A* **276**, 465-476.

[7] Freidlin, M. I. & Wentzell, A.D.(1984), *Random perturbations of dynamic system*, Berlin: Springer-Verlag.

[8] Ginebra, J. & Clayton, M.K.(1995), Response surface bandits, *J. R. Statist. Soc. B* **57** 771-784.

[9] Goodwin, G.C. & Sin, K.(1984), *Adaptive Filtering, Prediction, and Control,* Prentice Hall, Englewood Cliffs, New Jersey.

[10] Hajek, B. (1988), *Cooling schedules for optimal annealing,* Math. Oper. Res., **13**. pp. 311-329.

[11] Hwang, C-R & Sheu, S-J(1990), *Large-Time behavior of perturbed diffusion Markov processes with applications to the second eigenvalue problem for Fokker-Planck operators and simulated annealing*, Acta Applicandae Mathematicae **19** pp. 253-295.

[12] Kushner, H. & Clark, D.(1978) *Stochastic approximation methods for constrained and unconstrained systems,* New York: Springer-Verlag.

[13] van Laarhoven, P. J. M. & Aarts, E. H. L.(1987), *Simulated annealing: theory and applications*, Reidel, Dordrecht: D. Reidel Publishing Company.

[14] Ljung, L. & Soderström, T.(1983), *Theory and Practice of Recursive Identification*, MIT Press, Cambridge.

[15] Nevel'son, M.B. & Has'minskii, R.Z.(1976), *Stochastic approximation and recursive estimation* , Providence, Rhode Island: Translation of Math. Monograph 47, Amer. Math. Soc. .

[16] Robbins H. and Monro S. (1951) A stochastic approximation method. Ann math. Stat., *22* 400-407.

[17] Sirlantize K., Feng J., and Liu W. (2002) Novel gradient algorithms for stochastic optimization. (Submitted to SIAM journals)

[18] Winkler,G. (1994), *Image analysis, random fields and dynamic Monte Carlo methods. An introduction to mathematical aspects,* Berlin: Springer-Verlag .

[19] Yin, G. & Yin, K.(1994), *Asymptotically optimal rate of convergence of smoothed stochastic recursive algorithms,* Stochastics and Stochastics Reports, **47** pp. 21-46.