On stochastic algorithms for the global optimization problem on a torus

A. Yu. Veretennikov[1,2] and E. A. Zhizhina[2]

[1] University of Leeds
[2] Institute of Information Transmission Problems, Moscow

Abstract

We consider a discrete stochastic algorithm for finding minima of a given function on the torus T^d.

1 Introduction

Let $F : R^d \mapsto R^1$ be a smooth function bounded from below, with a finite number of global minima, the latter, hence, lying in some bounded domain. A long-standing problem is to construct an algorithm for finding the minimal value of F. Assuming that we can restrict all minima on some compact, consider the problem on the torus, T^d, and assume the gradient $\nabla F =: f$ is known. This does not mean that the problem in R^d is reduced to the one on T^d, we are just saying that the problem on T^d is a reasonable simplification of the general R^d case.

We consider a discrete, gradient type algorithm defined by a Markov chain $(X_n^a,\ n = 0, 1, 2, \ldots)$

$$X_{n+1}^a = X_n^a - f(X_n^a)\, a + \sqrt{a\delta}\, W_n, \quad X_0^a = x, \tag{1}$$

on the d-dimensional torus, $X_n^a \in T^d$. Here $a > 0,\ \delta > 0$ are *small* parameters of the model, $(W_n,\ n \geq 0)$ are d-dimensional i.i.d. random variables with values on the torus T^d, generated by standard Gaussian random variables on R^d; and we assume, without loss of generality, that

$$F(x) \geq 0, \quad \text{with} \quad \min_{x \in T^d} F(x) = 0.$$

The parameter δ will be chosen in the sequel as a function of a, i.e., $\delta = \delta(a)$.

The problem was considered in many papers: we only mention the works Freidlin and Wentzell (1991) and Ljung *et al.* (1992), in which the reader can find more references. However, Freidlin and Wentzell (1991) only deals with continuous time version of the algorithm, while in Ljung *et al.* (1992) this particular setting is studied only for 1-dimensional case. We also mention just one paper Chiang *et al.* (1987) on another algorithm type, with slowly decreasing diffusion coefficient which we do not study here. To our opinion, in some sense, existing results do not give a fully satisfactory solution. The aim of this note is to make some step towards a more satisfactory solution.

2 The idea

Gradient type algorithms are often used in stochastic approximation, cf. Nevel'son and Has'minskii (1973). We briefly remind their main idea now. They exploit the idea of finding critical points of the function F; namely, assume for a moment that there is no noise, then it might be expected that the (deterministic) trajectory goes to one of the points at which f vanishes, hence, a critical point of F indeed, which, however, may well be far away from the global minimum. Stochastic

noise introduces an additional perturbation to the picture, so that having achieved a neighbourhood of a critical point, – where due to $f \approx 0$ the deterministic component of the dynamics is negligible, – the process nevertheless is able to leave this neighbourhood due to this noise, and hence, gets a chance to find (all) other critical points. Overall, this gives a hope that finally the process may well find a global minimum (minima), and leave any its small neighbourhood not often. Thus, most of the whole mass of the stationary measure of such a process should be concentrated in some small neighbourhood of global minima. This idea was investigated in a series of papers and monographs. In particular, in Freidlin's works it was shown how a continuous time algorithm works here, actually, for more general algorithms: for a certain class of functions F, although not for all, Freidlin's results give a complete answer, and we will use them. Note, however, that 'in practice' a continuous algorithm usually is to be modelled by a discrete one. Our goal in this paper is to show that under appropriate additional assumptions the algorithm (1) does provide a satisfactory answer. The approach is based on Freidlin's results established using large deviation theory, and Girsanov's transformation of measure.

3 The setting of the problem

3.1 Assumptions

(A_1) $F \in C^2(T^d)$.

(A_2) The set $\{x : \ f(x) = 0\}$ consists of a finite number of compact connected disjoint sets $K_1, \dots, K_n$.

It can be checked that the condition (**A**) from Freidlin and Wentzell (1991), section 6.2, related to more general dynamical systems is satisfied under (A_1)–(A_2).

3.2 Main result

Denote $F_\varepsilon := \{x \in T^d : \ F(x) \geq \varepsilon\}$, $\chi_\varepsilon(y) := 1\,(y \in F_\varepsilon)$, and $P_n^a(x,y)$ a transition density of X_n^a given $X_0^a = x$.

Theorem 1 *Let assumptions (A_1)–(A_2) be satisfied, and $\delta(a) \geq C|\ln a|^{-1}$, with C large enough. Then, for any $\varepsilon > 0$*

$$\lim_{\substack{a\to 0\\ \delta(a)\to 0}} \limsup_{n\to\infty} \int_{T^d} P_n^a(x,y)\chi_\varepsilon(y)dy = 0. \tag{1}$$

Here notation $\lim\limits_{\substack{a\to 0\\ \delta(a)\to 0}}$ means that both a and $\delta(a)$ tend to zero simultaneously, while $\delta(a) \geq C|\ln a|^{-1}$.

Moreover, it is plausible that there exists $c > 0$ such that

$$\sup_{x\in T^d} \limsup_{n\to\infty} \int_{T^d} P_n^a(x,y)\chi_\varepsilon(y)dy \leq c^{-1}\exp(-c/\delta(a)), \tag{2}$$

and

$$\sup_{x\in T^d} \int_{T^d} P_n^a(x,y)\chi_\varepsilon(y)dy \leq c^{-1}\left(\exp(-c/\delta(a)) + \exp(-c\,na\exp(-c^{-1}/\delta(a)))\right). \tag{3}$$

We formulate these bounds here as rather probable hypotheses.

4 Sketch of the proof

We will show only the sketch of the proof of the assertion (1) here, and the calculus could be also used at least for deriving (2).

1. Note that $\mathbf{E}_x\mathbf{1}(X_n^a \in F_\varepsilon) = \int_{T^d} P_n^a(x,y)\chi_\varepsilon(y)dy$. Consider an Itô process $\widetilde{X}_t^a$ satisfying an *SDE on the torus* T^d, which coincides with X_t^a at times $t = na$:

$$d\widetilde{X}_t^a = -f\left(\widetilde{X}_{[t/a]a}^a\right)\,dt + \sqrt{\delta(a)}\,dW_t, \quad \widetilde{X}_0^a = x, \tag{1}$$

where W_t is a d-dimensional Wiener process, $[b]$ an integer part of $b \in R$. Then it suffices to show that for any $\varepsilon > 0$ and $\alpha > 0$ there exist a_0 and T_δ such that for any $0 < a < a_0$ and any $t > T_\delta$ we have

$$\sup_{x\in T^d} \mathbf{E}_x\mathbf{1}(\widetilde{X}_t^a \in F_\varepsilon) < \alpha. \tag{2}$$

2. Along with the process (1), consider a diffusion process X_t^δ of the form

$$dX_t^\delta = -f\left(X_t^\delta\right)\,dt + \sqrt{\delta}\,dW_t, \quad X_0 = x. \tag{3}$$

We will estimate $\sup_{x\in T^d}\mathbf{E}_x\mathbf{1}(X_t^\delta \in F_\varepsilon)$, and using Girsanov's formula, derive a similar bound for $\widetilde{X}_t^a$. The stationary distribution of the process (3) has a Gibbsian form $\mu_\delta(dx) = Z_\delta^{-1}e^{-\frac{2}{\delta}F(x)}\mu_0(dx)$, where Z_δ is the normalizing factor, μ_0 the uniform distribution on T^d. It is readily shown that $\mu_\delta(F_\varepsilon) \to 0, \quad \delta \to 0$. Moreover, it follows from Freidlin and Wentzell (1991) section 6.2, that is, for C_0 large enough, – depending on certain characteristics of the system, – and any $\varepsilon > 0$ and $\alpha > 0$, there exists a_0 such that for any $0 < a < a_0$ and $t \geq T(\delta(a)) = \exp(C_0/\delta)$ we have,

$$\sup_{x\in T^d} \mathbf{E}_x\mathbf{1}(X_t^\delta \in F_\varepsilon) < \alpha, \tag{4}$$

and the same assertion is also true for *any* initial distribution of X_0^δ.

3. Let $t' \geq 0$ and $t = t' + T(\delta(a))$. From Girsanov's formula applied on the interval $[t', t]$ and Hölder's inequality it follows for any Borel $A \subset T^d$,

$$\mathbf{E}_x\mathbf{1}(\widetilde{X}_t^a \in A) \leq e^{kT(\delta(a))\frac{a}{\delta(a)}}\,\mathbf{E}_x\left(\mathbf{E}_x\left(\mathbf{1}(X_t^\delta \in A) \mid X_{t'}^\delta = \widetilde{X}_{t'}^a\right)\right)^{\frac{1}{p}}, \tag{5}$$

where p, q are any positive satisfying the relation $p^{-1} + q^{-1} = 1$, and $k > 0$; the assumption $f \in C^1$ is used in this calculus which we drop here. The inner expectation in the right hand side of (5) is understood as conditional expectation given $X_{t'}^\delta$ which starts at time t' from the 'initial value' $\widetilde{X}_{t'}^a$ (otherwise there is a question whether the condition has a positive measure; most probably it has not). Because of (4), the inner expectation with $A = F_\varepsilon$ does not exceed α. Hence, we get a version of (2), with $\alpha^{1/p}$ instead of α, – which does not make any difference, – and a factor $e^{kT(\delta(a))\frac{a}{\delta(a)}}$.

4. This suffices for the assertion (1) to be valid for any $t \geq T_\delta := T(\delta(a))$ if the latter value satisfies the inequality $T(\delta(a)) \leq C\delta(a)/a$ with some $C > 0$. For this, $\delta(a) \geq 2\,C_0\,|\ln a|^{-1}$ is sufficient. So, (2) and (1) follow. $\diamond$

Acknowledgements: The authors thank the RFBR project 01-01-0444 and INTAS 99-0590 for support; the second author thanks Leeds University for hospitality during their visit in February 2004.

References

Freidlin, M. I. and Wentzell, A. D. (1991). *Random perturbations of dynamical systems*, 2nd edition. New York, Springer.

Ljung, L., Pflug, G., and Walk, H. (1992). *Stochastic approximation and optimization of random systems*. Basel, Birkhäuser.

Chiang, T.-S., Hwang, C.-R., and Sheu, S.-J. (1987). Diffusion for global optimization in R^n. *SIAM J. Control and Optimization*, **25**, 737-753.

Nevel'son, M. B., and Has'minskii, R. Z. (1973). *Stochastic approximation and recursive estimation.* Providence, R.I., AMS.