ELSEVIER

Stochastic Processes and their Applications 94 (2001) 95–103

stochastic processes and their applications

www.elsevier.com/locate/spa

An adaptive simulated annealing algorithm

Guanglu Gong[a, 1], Yong Liu[b,c, *], Minping Qian[b, 2]

[a] *Department of Mathematical Sciences, Tsinghua University, Beijing, 100084, People's Republic of China*

[b] *Department of Probability and Statistics, Peking University, Beijing, 100871, People's Republic of China*

[c] *Institute of Applied Mathematics, Chinese Academy of Sciences, Beijing, 100080, People's Republic of China*

Received 23 November 1999; received in revised form 3 January 2001; accepted 3 January 2001

Abstract

In this paper, inspired by the idea of Metropolis algorithm, a new sample adaptive simulated annealing algorithm is constructed on finite state space. This new algorithm can be considered as a substitute of the annealing of iterative stochastic schemes. The convergence of the algorithm is shown.

MSC: 60G99; 65C05

Keywords: Simulated annealing; Adaptive algorithm; Recognition of handwriting Chinese characters; Estimate of spectral gap

1. Introduction

Adaptive algorithms with stochastics appear frequently in various applications, such as self-organizing learning algorithms (see Kohonen, 1984), optimization (see Kirpatrick et al., 1983), neural networks (see Hertz et al., 1991), system identification, adaptive controls (see Michalwicz, 1992) and so on. The function of these algorithms is to adjust a state vector (or monitored parameter vector) X_n for specifying the system considered, where n refers to the time of observation of the system. In most cases of applications mentioned above, the rule used to update $X_{\cdot}$ will typically be of the form

$$X_{n+1} = X_n + r_n b(X_n, \xi_n), \tag{1.1}$$

* Corresponding author. Institute of Applied Mathematics, Chinese Academy of Science, Beijing, 100080, People's Republic of China.

E-mail addresses: glgong@math.tsinghua.edu.cn (G. Gong), liuyong@amath4.amt.ac.cn, liuyong71@yahoo.com (Y. Liu), qianmp@sxx0.math.pku.edu.cn (M. Qian).

[1] Supported by NSFC of China 79970120.

[2] Supported by NSFC of China 19971005, the Doctoral program Foundation of Institution of Higher Education and 863 program.

0304-4149/01/$ - see front matter
PII: S0304-4149(01)00082-5

where r_n is a sequence of small gains and ξ_n is the input of the system at time n, either deterministic or stochastic. This model can be illustrated by practical setups. Let us take the recognition of off-line handwriting Chinese characters as an example. In this case, samples of handwriting Chinese characters are read in one by one, denoted by $\xi_1,\ldots,\xi_n,\ldots$, submitted the population of Chinese handwriting characters. Assume that this population is described by an random vector ξ, i.e. $\xi_1,\ldots,\xi_n,\ldots$ are i.i.d. copies of ξ. The bias of a candidate point x from the samples is measured by the following objective function:

$$U(x)=E\left(\min_{i\leqslant m}\|x^{(i)}-\xi\|^2\right),\quad x=(x^{(1)},\ldots,x^{(m)}).$$

Here m is the number of "standard patterns" of the handwriting Chinese characters desired to be established, e.g. $m=5000$ or 20,000. Then the set of standard patterns of size m will be represented by $x^*=(x^{(1)*},\ldots,x^{(m)*})$ – a minimizing site of the objective function $U(x)$, i.e.

$$U(x^*)=\min_x U(x).$$

Thus $x^{(1)*},\ldots,x^{(m)*}$ are the optimal representations of handwriting Chinese characters within these samples. Of course, maybe there are several $x^{(i)*}s$ standing for the same character because of the different styles of writing. Here we see that $U(x)$ takes the form of $Eg(x,\xi)$ with a function $g(x,\xi)$, of which the derivative with respect to x is not continuous. The clustering problems are much like this. And they are focused to minimize an expectation

$$U(x)=Eg(x,\xi).$$

A stochastic approach to treat the above model, well known as the self-organizing algorithm, was suggested by Kohonen (1984), where the iterative formula (1.1) is designed to find the minimum of $U(x)$.

Algorithm (1.1) has been studied broadly (see, e.g. Benveniste et al., 1990; Fang et al., 1997 and references therein). However, even in the case of $\mathrm{d}g(x,\xi)/\mathrm{d}x$ being Lipschitz continuous, it is not always successful in finding an element in the set $\underline{S}=\{z\in R^d, U(z)=\min_{x\in R^d}U(x)\}$. To avoid getting trapped in local minima, stochastic perturbation is added as follows:

$$X_{n+1}=X_n+r_nb(X_n,\xi_n)+h_n\zeta_n, \tag{1.2}$$

where ζ_n is a sequence of independently and identically distributed (i.i.d.) random variables and $b(x,\xi)$ takes the role of $\mathrm{d}g(x,\xi)/\mathrm{d}x$. As it borrows the idea of annealing process from statistical mechanics, it is well known as the simulated annealing algorithm. In various special cases, the asymptotic behavior of (1.1), (1.2) as $n\to\infty$ has been studied by Kushner (1987), Gelfand and Mitter (1991, 1993), Ljung et al. (1992), Métivier and Priouret (1987), Fang et al. (1997) and so on. Using the generalized large deviation theory of Wentzell (1990), under mild conditions on ζ_n, r_n, h_n, Fang et al. (1997) give a uniform treatment and convergence theorem of (1.2) which includes the results in Gelfand and Mitter (1991, 1993), Ljung et al. (1992). Their main result can be roughly stated as follows: If $\bar{b}(x)\equiv Eb(x,\xi_n)$ is Lipschitz continuous and has

a potential function $U(x)$ and $E\zeta_n = 0$, then under some restrictions on the behavior of $\bar{b}(x)$ and $b(x, \xi_n)$ at infinity, for $\{r_n\}$ in a wide class, one can always choose $\{h_n\}$ such that for any $\delta > 0$,

$$P_{0,y}\left\{U(X_n) < \min_{z \in R^d} U(z) + \delta\right\} \to 1,$$

uniformly for y in an arbitrary compact set F.

The above iterative algorithm (1.2) has given a satisfactory answer if X_n belongs to the Euclidean space and $\mathrm{d}g(x, \xi)/\mathrm{d}x$ is Lipschitz continuous. However, in the application of the image pattern recognition, especially in the Kohonen's self-organizing algorithm, $\mathrm{d}g(x, \xi)/\mathrm{d}x$ is always discontinuous, which almost makes (1.2) useless, since for (1.1), the convergence behavior is only known extremely roughly even for the 2-dimensional case, while (1.2) is more complicated. Fortunately, in the image pattern recognition, images are actually discretized by pixels, and we can avoid the Kohonen's model and let the state vector take values in a finite set $\mathbf{X}$ instead.

How to design an adaptive algorithm in this situation such that it converges to the global minima of $U(x) = E(b(x, \xi))$, when x takes value in a finite set? This situation often happens in the stochastic approximation and neural network framework. In general, an adaptive algorithm for a finite state space can be built in the two following equivalent forms: either by putting

$$X_{n+1} = f_n(X_n, \xi_n), \tag{1.3}$$

where f_n is a mapping from $\mathbf{X} \times \mathbb{R}$ to $\mathbf{X}$, or by defining $\{X_0, X_1, X_2 \ldots X_n \ldots\}$ as a Markov chain with a prescribed transition matrix. Similar to (1.1), $\{X_n\}$ does not always converge to a global minimum.

Inspired by the idea of Metropolis algorithm, we construct a sample adaptive algorithm as following.

Let $\mathbf{X}$ be a finite set, and $N = \mathrm{card}(X)$. Assume $\{U_n(x),\ x \in \mathbf{X}, \ldots,\ n = 1, 2 \ldots\}$ be a family of random variables on probability space $(\Omega, \mathscr{F}, \mathbf{P})$, satisfying:

(1) for any $x \in \mathbf{X}$, $\{U_n(x),\ n = 1, 2 \ldots\}$ is a sequence of independently, identically distributed random variables, standing for the sample read in;
(2) $EU_n(x) = U(x)$, and $\max_{x \in \mathbf{X}} \mathrm{Var}(U_n(x)) = D < \infty$.

$U_n(x)$ is the observation value at time n. Let $S_n(x) = 1/n \sum_{k=1}^{n} U_k(x)$. By strong law of large numbers, we have

$$\mathbf{P}\left\{\lim_{n \to \infty} S_n(x) = U(x)\right\} = 1.$$

Let μ_0 be a probability measure on $\mathbf{X}$ satisfying $\mu_0(x) > 0$ for any $x \in \mathbf{X}$, and $q_0(x, y)$ be an irreducible reversible probability transition matrix on $\mathbf{X} \times \mathbf{X}$ such that μ_0 is the invariant probability measure of $q_0(x, y)$, i.e.

$$\alpha(x, y) \equiv \mu_0(x) q_0(x, y) = \alpha(y, x), \quad x, y \in \mathbf{X}.$$

Next, we define a sequence of random probability measures and random probability transition matrices: for any $\omega \in \Omega$, $\beta_n \geqslant 0$, $\beta_n \to \infty$

$$\mu_{\beta_n,\omega}(x) \equiv \frac{\mathrm{e}^{-\beta_n S_{K(n),\omega}(x)}}{Z_{\beta_n,\omega}}\mu_0(x), \quad Z_{\beta_n,\omega} = \sum_{x\in\mathbf{X}} \mathrm{e}^{-\beta_n S_{K(n),\omega}(x)}\mu_0(x),$$

$$q_{\beta_n,\omega}(x,y) \equiv \begin{cases} \exp[-\beta_n(S_{K(n),\omega}(y) - S_{K(n),\omega}(x))^+]q_0(x,y), & y \neq x, \\ 1 - \sum_{z\neq x} q_{\beta_n,\omega}(x,z), & y = x, \end{cases}$$

where $K(n)$ is a function from $\mathbb{N}$ to $\mathbb{N}$ (the set of natural numbers) satisfying $K(n) \geqslant K(n-1)$ and $\lim_{n\to\infty} K(n) = \infty$. Denote $\mathscr{B} \equiv \{\phi\colon \mathrm{X} \mapsto \mathbb{R}\}$. For any $\phi \in \mathscr{B}$, let $\mu_{\beta_n,\omega}(\phi) \equiv \sum_{x\in\mathbf{X}} \mu_{\beta_n,\omega}(x)\phi(x)$ and $\|\phi\|_\infty = \sup_{x\in\mathbf{X}} |\phi(x)|$.

We consider the coordinate process $\{Y_n,\ n = 1,2\ldots\}$: $Y_n(\bar{\omega}) = \bar{\omega}_n$, $\bar{\omega} \in \mathbf{X}^\infty$ on the coordinate space $\mathbf{X}^\infty$, and a random probability measure Q_ω on $\mathbf{X}^\infty$ such that Y_n is a Markov chain under Q_ω with probability transition matrix

$$Q_\omega[Y_n = y | Y_{n-1} = x] = q_{\beta_n,\omega}(x,y).$$

We define the above random Markov chain to be the adaptive simulated annealing algorithm. Let us give some preliminary propositions, before we state and prove the main theorem. First, we define the random operator $L_{\beta_n,\omega}$ and the Dirichlet form corresponding $L_{\beta_n,\omega}$ as follows:

$$L_{\beta_n,\omega}\phi(x) \equiv \sum_{y\in\mathbf{X}} (\phi(y) - \phi(x))q_{\beta_n,\omega}(x,y), \quad x \in \mathbf{X},$$

$$\begin{aligned} \mathbf{E}_{\beta_n,\omega}(\phi,\psi) &\equiv -\langle \phi, L_{\beta_n,\omega}(\psi)\rangle_{L^2(\mu_{\beta_n,\omega})} \\ &= \frac{1}{2Z_{\beta_n,\omega}} \sum_{x,y\in\mathbf{X}} \exp[-\beta_n(S_{K(n),\omega}(x) \vee S_{K(n),\omega}(y))] \\ &\quad \times(\phi(x) - \phi(y))(\psi(x) - \psi(y))\alpha(x,y), \end{aligned}$$

where $\phi, \psi \in \mathscr{B}$. Clearly, $-L_{\beta_n,\omega}$ is a non-negative definite operator on $L^2(\mu_{\beta_n,\omega})$. Let

$$\lambda_{\beta_n,\omega} \equiv \inf\{\mathbf{E}_{\beta_n,\omega}(\phi,\phi)\colon \|\phi\|_{L^2(\mu_{\beta_n,\omega})} = 1,\ \mu_{\beta_n,\omega}(\phi) = 0\}$$

then $\lambda_{\beta_n,\omega}$ is the gap between 0 and the rest of the spectrum of $-L_{\beta_n,\omega}$.

For any $x, y \in \mathbf{X}$, we define a path from x to y to be any sequence of points joining x to y: $x = x_0, x_1, \ldots, x_k = y$ satisfying

$$q_0(x_{i-1}, x_i) > 0 \text{ (i.e. } q_{\beta_n}(x_{i-1}, x_i) > 0), \quad i = 1,\ldots,k.$$

Let $\Xi(x,y)$ denote the set of all the paths from x to y defined above. Let $p = (x_0, x_1, \ldots, x_k)$ be an element of $\Xi(x,y)$. For $p \in \Xi(x,y)$, we define

$$\mathrm{Elev}(p)(\omega,n) \equiv \max\{S_{n,\omega}(x_i),\ x_i \in p\},$$

$$\mathrm{Elev}(p) \equiv \max\{U(x_i),\ x_i \in p\},$$

$$H_{n,\omega}(x,y) \equiv \min\{\mathrm{Elev}(p)(\omega,n),\ p \in \Xi(x,y)\},$$

$$H(x,y) \equiv \min\{\mathrm{Elev}(p),\ p \in \Xi(x,y)\},$$

$$m_{n,\omega} \equiv \left\{ H_{n,\omega}(x,y) - S_{n,\omega}(x) - S_{n,\omega}(y) + \min_{z \in \mathbf{X}} S_{n,\omega}(z), x, y \in \mathbf{X} \right\},$$

$$m \equiv \left\{ H(x,y) - U(x) - U(y) + \min_{z \in \mathbf{X}} U(z),\ x, y \in \mathbf{X} \right\}.$$

Obviously, if $\|S_{n,\omega} - U\|_\infty < \alpha$, then $|m_{n,\omega} - m| < 4\alpha$. Due to the results of Holley and Stroock (1988), Löwe (1995) and Diaconis and Stroock (1991), we have the following proposition

Proposition 1.1. *There exists a constant $C > 0$ independent of n and ω, for any $\beta_n \geqslant 0$ such that*

$$\lambda_{\beta_n,\omega} \geqslant C\mathrm{e}^{-\beta_n m_{n,\omega}}.$$

We define $\zeta = \min_{x \in \mathbf{X}} U(x)$. Let x' satisfy $U(x') = \zeta$. Then for any $\delta > 0$, we have

Lemma 1.2. *There is a constant C_1 such that*

$$\lim_{n \to \infty} \mathbf{P}\{\omega\colon \mu_{\beta_n,\omega}\{x, U(x) > \delta + \zeta\} \leqslant C_1 \mathrm{e}^{-\beta_n \delta/4}\} = 1.$$

Proof. If $\omega \in \{\omega \in \Omega, \|S_{K(n),\omega}(x) - U(x)\|_\infty < \frac{\delta}{4}\}$, then

$$\begin{aligned}
\mu_{\beta_n,\omega}(x, U(x) > \delta + \zeta) &\leqslant \frac{\sum_{\{x:\ U(x) > \delta + \zeta\}} \mathrm{e}^{-\beta_n S_{K(n),\omega}(x)} \mu_0(x)}{\mathrm{e}^{-\beta_n S_{K(n),\omega}(x)} \mu_0(x')} \\
&\leqslant \frac{\sum_{\{x:\ S_{K(n),\omega}(x) > \delta/2 + \zeta\}} \mathrm{e}^{-\beta_n S_{K(n),\omega}(x)} \mu_0(x)}{\mathrm{e}^{-\beta_n S_{K(n),\omega}(x)} \mu_0(x')} \\
&\leqslant \frac{1}{\mu_0(x')} \mathrm{e}^{-\beta_n \delta/4}.
\end{aligned}$$

And then the lemma follows from the law of large numbers.

Main Theorem. *We choose $\beta_n \to \infty$ and $0 \leqslant \beta_n - \beta_{n-1} \leqslant 1/cn$, $c > m$, take η satisfying $m/c < \eta < 1$ and choose $K(n)$ satisfying*: (1) $[K(n) - K(n-1)]/K(n) < M/n^\eta$, $n = 1, 2 \ldots$, *where M is a constant*; (2) $\sum_{n=1}^{\infty} n^\eta / K(n) < \infty$. *Then for any $\delta > 0$ and $\varepsilon > 0$, there exists n_0, and for any $n > n_0$ there exist constants C_1, $C_{2,\varepsilon}$ and $v < 0$ such that*

$$\mathbf{P}\left\{ Q_\omega(U(Y_n) > \delta + \min_{x \in \mathbf{X}} U(x)) < C_1 n^{-\delta/4c} + C_{2,\varepsilon} n^v \right\} \geqslant 1 - \varepsilon.$$

Remark 1.1. It is necessary in practical computation that $c, v, C_1, C_{2,\varepsilon}$, and β_n are independent of ω.

Remark 1.2. For convenience, we assume that $Q_\omega[Y_0 = x_0] = 1$, $x_0 \in \mathbf{X}$, for any $\omega \in \Omega$. In fact, this assumption does not influence the convergence of our algorithm.

Remark 1.3. For a given η, we can take a $\alpha > 0$ satisfying $(m + 4\alpha)/c < \eta < 1$ since $m/c < \eta < 1$. Thus by the following proof of the main theorem, we are able to take $v \in ((m + 4\alpha)/c - \eta, 0)$.

Remark 1.4. Actually, we can choose $K(n) = n^2$, then $K(n)$ satisfies the conditions of the theorem.

Our algorithm is designed in view of discrete time and finite state space. It may be considered as a substitute of the annealing of iterative stochastic schemes (1.2) in case of finite state space and can be hopefully extended to the denumerable state case with some necessary modifications.

Comparing our SA algorithm with (1.2), the function of probability transition matrix $q_0(x, y)$ is analogous to the distribution of the artificial noise ζ_n in (1.2), and that $\exp[-\beta_n(S_{K(n),\omega}(y) - S_{K(n),\omega}(x))^+]$ is analogous to h_n in (1.2).

In order to bring the information into full play, we use the mean of observation values S_n. Because our algorithm converges finally to some set in $\mathbf{X}$ while mean values even do not belong to the sample space or the state space $\mathbf{X}$, that is different from some adaptive learning algorithms (see Kohonen, 1984), in which the mean values are often considered as the cluster points.

2. Proof of the Main Theorem

Just because we borrow the ideas of proof of Holley and Stroock (1988), Götze (1992) and especially quote a general framework of Frigerio and Grillo (1993), we only give a short sketch of our proof.

Lemma 2.1. *Let* $A_n \equiv \{\omega, 1/K(n) \sum_{j=K(n-1)+1}^{K(n)} U_j < 1/n^\eta\}$, *then* $\lim_{k\to\infty} \mathbf{P}\{\bigcap_{n>k} A_n(x)\} = 1$.

Proof. Because of the hypotheses on $K(n)$, we have

$$\lim_{k\to\infty} \mathbf{P}\left\{\bigcap_{n>k} A_n(x)\right\} \geqslant 1 - \lim_{k\to\infty} \sum_{n>k} \mathbf{P}\{A_n^c(x)\}$$

$$\geqslant 1 - 2 \lim_{k\to\infty} \sum_{n>k} \frac{n^{2\eta}}{K(n)^2} \sum_{j=K(n-1)+1}^{K(n)} EU_n^2(x)$$

$$\geqslant 1 - 2ME U_1^2(x) \lim_{k\to\infty} \sum_{n\geqslant k} \frac{n^\eta}{K(n)} = 1. \qquad \square$$

For α and for any $\varepsilon > 0$, by Hajek–Renyi inequality, we can take $K_{\alpha,\varepsilon} > 0$ such that for any $k > K_{\alpha,\varepsilon}$, $x \in \mathbf{X}$, $\mathbf{P}\{\omega, \sup_{j\geqslant k} |S_j(x) - U(x)| \leqslant \alpha\} \leqslant \frac{1}{\alpha^2}(D/k + D\sum_{j=k+1}^{\infty} 1/j^2) \leqslant \varepsilon/4N$.

Now, we take $K'_{\alpha,\varepsilon} > K_{\alpha,\varepsilon}$ such that for any $K(k) > K'_{\alpha,\varepsilon}$, $\mathbf{P}\{\bigcap_{n\geqslant k} A_n(x)\} > 1 - \frac{\varepsilon}{4N}$.

Lemma 2.2. *If* $\omega \in \bigcap_{x\in \mathbf{X}}\{(\bigcap_{K(n)\geqslant K'_{\alpha,\varepsilon}} A_n(x)) \cap \{\omega, \sup_{K(n)\geqslant K'_{\alpha,\varepsilon}} |S_{K(n),\omega}(x) - U(x)| < \alpha\}\}$, *then there exists a constant* $M_{1,\varepsilon} > 0$ *such that for any* $K(n) > K'_{\alpha,\varepsilon}$

$$\|S_{K(n),\omega}\|_\infty \leqslant M_{1,\varepsilon} \quad \textit{and} \quad \|S_{K(n),\omega} - S_{K(n-1),\omega}\|_\infty < \frac{M_{1,\varepsilon}}{n^\eta}.$$

Proof. Obviously, $\|S_{K(n),\omega}\|_\infty < M'$, where M' is a constant. Furthermore,

$$\begin{aligned}
&|S_{K(n),\omega}(x) - S_{K(n-1),\omega}(x)| \\
&\leqslant \left|\left(\frac{1}{K(n)} - \frac{1}{K(n-1)}\right)\sum_{j=1}^{K(n-1)} U_j(x)\right| + \left|\frac{1}{K(n)}\sum_{j=K(n-1)+1}^{K(n)} U_j(x)\right| \\
&\leqslant \left|\frac{K(n-1)-K(n)}{K(n)} S_{K(n-1),\omega}(x)\right| + \left|\frac{1}{K(n)}\sum_{j=K(n-1)+1}^{K(n)} U_j(x)\right| \\
&\leqslant \frac{M'M}{n^\eta} + \frac{1}{n^\eta} \leqslant \frac{M_{1,\varepsilon}}{n^\eta}. \qquad \square
\end{aligned}$$

Using Hajek–Renyi inequality again, we obtain for J_ε large enough,

$$\mathbf{P}\left\{\max_{K(n)\leqslant K'_{\alpha,\varepsilon}} \frac{1}{K(n)}\left|\sum_{j=1}^{K(n)}(U_j(x) - U(x))\right| > J_\varepsilon\right\} \leqslant \frac{D}{J_\varepsilon^2}\sum_{j=1}^{\infty}\frac{1}{j^2} < \frac{\varepsilon}{4N}.$$

If $\omega \in \bigcap_{x\in \mathbf{X}}\{\max_{K(n) < K'_{\alpha,\varepsilon}} \frac{1}{K(n)} |\sum_{j=1}^{K(n)}(U_j(x) - U(x))| \leqslant J_\varepsilon\}$, then there exists a constant $M_{2,\varepsilon} > 0$ such that $\|S_{K(n),\omega}(x)\|_\infty < M_{2,\varepsilon}$, for $K(n) \leqslant K'_{\alpha,\varepsilon}$. Moreover, it is easy to show that there exist the constants $\bar{C}, M_{3,\varepsilon}$ satisfying $0 < \bar{C} \leqslant C$ and $M_{3,\varepsilon} > 0$ such that

$$\lambda_{\beta_n,\omega} \geqslant \bar{C}\mathrm{e}^{-\beta_n(m+4\alpha)} \quad \text{and} \quad \|S_{K(n),\omega} - S_{K(n-1),\omega}\| < \frac{M_{3,\varepsilon}}{n^\eta} \text{ for } K(n) \leqslant K'_{\alpha,\varepsilon}.$$

We denote

$$\begin{aligned}
B(K'_{\alpha,\varepsilon}) \equiv \bigcap_{x\in \mathbf{X}} \Bigg\{ &\left\{\omega\colon \max_{K(n)<K'_{\alpha,\varepsilon}} \frac{1}{K(n)}\left|\sum_{j=1}^{K(n)}(U_j(x)-U(x))\right| \leqslant J_\varepsilon\right\} \\
&\cap \left\{\omega\colon \sup_{K(n)\geqslant K'_{\alpha,\varepsilon}} \frac{1}{K(n)}\left|\sum_{j=1}^{K(n)}(U_j(x)-U(x))\right| \leqslant \alpha\right\} \cap \left(\bigcap_{K(n)\geqslant K'_{\alpha,\varepsilon}} A_n(x)\right)\Bigg\}.
\end{aligned}$$

Summing up the Lemma 2.2 and the above formula, we have

Lemma 2.3. *If* $\omega \in B(K'_{\alpha,\varepsilon})$, *then there exists a constant* $M_{4,\varepsilon}$ (*independent of* ω) *such that*

(1) $\|S_{K(n),\omega}\|_\infty < M_{4,\varepsilon}$, $n = 1, 2\ldots$;
(2) $\lambda_{\beta_n,\omega} \geqslant \bar{C}\mathrm{e}^{-\beta_n(m+4\alpha)}$, $n = 1, 2\ldots$;
(3) $\|S_{K(n),\omega} - S_{K(n-1),\omega}\| < M_{4,\varepsilon}/n^\eta$, $n = 1, 2\ldots$.

For $\phi \in \mathscr{B}$, we denote

$$Q_{n,\omega}(y) = \sum_{x \in \mathbf{X}} q_{\beta_n,\omega}(x,y) Q_{n-1,\omega}(x), \quad Q_{n,\omega}(\phi) = \sum_{x \in \mathbf{X}} Q_{n,\omega}(x)\phi(x), \quad n = 1,2\ldots,$$

where Q_0 is similar to that in Remark 1.2.

By the Theorem 2.5 in Frigerio and Grillo (1993), we have

Proposition 2.4. *If $\omega \in B(K'_{\alpha,\varepsilon})$, then for any $f \in \mathscr{B}$ we have*

$$\left| \sum_{x \in \mathbf{X}} f(x) Q_{n,\omega}(x) - \sum_{x \in \mathbf{X}} f(x) \mu_{\beta_n,\omega}(x) \right| \leqslant C_{2,\varepsilon} \|f\|_\infty n^v,$$

where $C_{2,\varepsilon}$ is a constant.

Proof of the Main Theorem. Let $A_\delta \equiv \{x\colon U(x) > \delta + \zeta\}$. Due to Lemma 2.2, it implies for any $\varepsilon > 0$, there exists $n_0 > K'_{\alpha,\varepsilon}$, such that for any $n > n_0$

$$\mathbf{P}\{\mu_{\beta_n,\omega}(A_\delta) < C_1 n^{-\delta/4c}\} > 1 - \frac{\varepsilon}{4}.$$

Let $\hat{B}_n(K'_{\alpha,\varepsilon}) \equiv \{\omega, \mu_{\beta_n,\omega}(A_\delta) < k n^{-\delta/4c}\} \cap B(K'_{\alpha,\varepsilon})$. If $\omega \in \hat{B}_n(K'_{\alpha,\varepsilon})$, then

$$Q_{n,\omega}(A_\delta) \leqslant \mu_{\beta_n,\omega}(A_\delta) + |Q_{n,\omega}(A_\delta) - \mu_{\beta_n,\omega}(A_\delta)| < C_1 n^{-\delta/4c} + C_{2,\varepsilon} n^v.$$

Hence we have

$$\mathbf{P}\left\{ Q_\omega \left(U(Y_n) > \delta + \min_{x \in \mathbf{X}} U(x) \right) < C_1 n^{-\delta/4c} + C_{2,\varepsilon} n^v \right\}$$
$$\geqslant \mathbf{P}\{\hat{B}_n(K'_{\alpha,\varepsilon})\} > 1 - \varepsilon. \qquad \square$$

If $U_n(x)$, $x \in \mathbf{X}$, $n = 1,2,\ldots$ are bounded random variables, i.e. for any n, x and $\omega \in \Omega$, there exists a constant $\bar{M}$, such that $|U_{n,\omega}(x) - U(x)| < \bar{M}$, then we have the following corollary.

Corollary 2.5. *For any $\varepsilon > 0$ and $\varepsilon' > 0$, there exists n_0, if $n > n_0$, then*

$$\mathbf{P}\{\omega, Q_{n,\omega}(A_\delta) < \varepsilon'\} > 1 - \varepsilon.$$

3. Unsolved problems

According to the referee's suggestions, we use $S_{K(n)}$ at the nth iteration of the SA algorithm rather than S_n in our original manuscript. This idea leads to the expected condition $c > m$ on the speed of decrease of the temperature in the main theorem, which become more delicate. As the referee pointed out, the question of the optimal choice of $K(n)$ should be observed. How can we choose $K(n)$ such that the speed of convergence is as fast as possible for given ε, δ. Another question is how to judge whether this SA algorithm with the random energy has sufficiently closed to a required set in a limited time and determine when it should be stopped.

These two questions are both valuable to be considered in practical computation.

However, in practical applications, it is difficult to obtain a satisfactory answer to the above problems by the ideas and tools used in the present paper. Although we guess that some large deviation results of Markov chains could be applied to solve this sort of problem, we are not sure how to apply them to this kind of random energy model.

Acknowledgements

We would like to thank the referee for his valuable comments, which were a great incentive to improve our paper.

References

Benveniste, A., Métiviet, M., Priouret, P., 1990. Adaptive Algorithms and Stochastic Approximations. Springer, Berlin.

Diaconis, P., Stroock, D., 1991. Geometric bounds for eigenvalues of Markov chains. Ann. Appl. Probab. 1 (1), 36–61.

Frigerio, A., Grillo, G., 1993. Simulated annealing with time-dependent energy function. Math. Z. 213, 97–116.

Fang, H.T., Gong, G.L., Qian, M.P., 1997. Annealing of iterative stochastic schemes. SIAM J. Control Optim. 35 (6), 1886–1907.

Gelfand, S.B., Mitter, S.K., 1991. Recursive stochastic algorithm for global optimization in R^d. SIAM J. Control Optim. 29, 999–1018.

Gelfand, S.B., Mitter, S.K., 1993. Metropolis-type annealing algorithms for global optimization in R^d. SIAM J. Control Optim. 31, 111–131.

Götze, F., 1992. Rate of convergence of simulated annealing processes, preprint paper.

Hertz, J., Krogh, A., Palmar, G.R., 1991. Introduction to the Theory of Neural Computation. Santa Fa Inst. Studies in the Science of Complexity. Addison-Wesley, Reading, MA.

Holley, R.A., Stroock, D.W., 1988. Simulated annealing via Sobolev inequalities. Comm. Math. Phys. 115, 553–569.

Kirpatrick, S., Gelatt, C.D., Vecchi, M.P., 1983. Optimization by simulated annealing. Science 220, 671–680.

Kohonen, T., 1984. Self-organization and Associative Memory. Springer, New York.

Kushner, H.J., 1987. Asymptotic global behavior for stochstic approximation and diffusions with slowly decreasing noise effects: global minimization via Monte Carlo. SIAM J. Appl. Math. 47, 169–185.

Ljung, L., Pflug, G., Walk, H., 1992. Stochastic approximation and optimization of random system. Birhäuser-verlag, Basel.

Löwe, M., 1995. Simulated annealing with time-dependent energy function via Sobolev inequalities. Stochastic Process. Appl. 63, 221–233.

Métivier, M., Priouret, P., 1987. Thérémes de convergence presque sûre pour une classe d'algorithmes stochastiques á pas décroissant. Probab. Theory Related Fields 74, 403–428.

Michalwicz, Z., 1992. Genetic Algorithms + Datastuctuers = Evolution Programs. Springer, Berlin.

Wentzell, A.D., 1990. Limit Theorems on Large Deviations for Markov Stochastic Processes. Kluwer Academic Publishers, Dordrecht.