Convergence of simulated annealing using the generalized transition probability

1998 J. Phys. A: Math. Gen. 31 5661

(http://iopscience.iop.org/0305-4470/31/26/007)

J. Phys. A: Math. Gen. **31** (1998) 5661–5672. Printed in the UK PII: S0305-4470(98)90052-X

Convergence of simulated annealing using the generalized transition probability

Hidetoshi Nishimori and Jun-ichi Inoue

Department of Physics, Tokyo Institute of Technology, Oh-okayama, Meguro-ku, Tokyo 152-8551, Japan

Received 12 January 1998, in final form 14 April 1998

Abstract. We prove the weak ergodicity of the inhomogeneous Markov process generated by the generalized transition probability of Tsallis and Stariolo under power-law decay of the temperature. We thus have a reason to conjecture convergence of the simulated annealing processes with the generalized transition probability to the minimum of the cost function. An explicitly solvable example in one dimension is analysed in which the generalized transition probability leads to a fast convergence of the cost function to the optimal value. We also investigate how far our arguments depend upon the specific form of the generalized transition probability proposed by Tsallis and Stariolo. It is shown that a few requirements on analyticity of the transition probability are sufficient to assure fast convergence in the case of the solvable model in one dimension.

1. Introduction

Simulated annealing has been a powerful tool for combinatorial optimization problems [1–4]. To find the minimum of a cost function, one introduces a stochastic process similar to Monte Carlo simulations in statistical mechanics with a control parameter corresponding to the temperature to allow the system to escape from local minima. By gradually decreasing the temperature one searches for increasingly narrower regions in the phase space closer to the optimal state, eventually reaching the optimal state itself in the infinite-time limit.

A very important factor in such processes is the annealing schedule, or the rate of decrease of temperature. If one lowers the temperature too quickly, the system may end up in one of the local minima. On the other hand, a very slow decrease of temperature would surely bring the system to the true minimum. However, such a slow process is not useful practically. One therefore has to determine carefully how quickly to decrease the temperature in simulated annealings. On this problem, Geman and Geman [5] proved that the decrease of temperature as $T = \text{constant}/\log t$, with the proportionality constant roughly of the order of the system size, guarantees convergence to the optimal state for a wide class of combinatorial optimization problems. This inverse-log law is still too slow for most practical purposes. Nevertheless, this result serves as a mathematical background for empirical investigations by numerical methods.

There have been a few proposals to accelerate the annealing schedule by modifying the transition probabilities used in the conventional simulated annealing. Szu and Hartly [6] pointed out for a problem defined in a continuum space that occasional non-local samplings significantly improve the performance, leading to an annealing schedule inversely proportional to time $T = \text{constant}/t$. This non-local sampling corresponds to a modification

0305-4470/98/265661+12$19.50

of the generation probability (or, more precisely, the neighbourhood) to be defined later. Tsallis and Stariolo [7] proposed to modify the acceptance probability generalizing the usual Boltzmann form in addition to the generation probability (which they call the visiting distribution). Numerical investigations show faster convergence to the optimal state by annealing processes using their generalized transition probability or its modifications [7–9]. Szu and Hartley and Tsallis and Stariolo proved that the modified *generation* probability assures convergence to the optimal state under a power-law decrease of the temperature as a function of time. However, there has been no mathematically rigorous argument on the convergence under the generalized *acceptance* probability of Tsallis and Stariolo.

We prove in the present paper that the inhomogeneous Markov process generated by the generalized transition (acceptance) probability of Tsallis and Stariolo satisfies the property of weak ergodicity under an annealing schedule inversely proportional to the power of time. Rigorously speaking, weak ergodicity (which roughly means asymptotic independence of the probability distribution from the initial condition) itself does not immediately guarantee the convergence to the optimal state. Nevertheless, our result is expected to be close enough to this final goal because the probability distribution would depend upon the initial condition if the annealing schedule is not chosen appropriately.

Various definitions are given in the next section. The proof of our main theorem appears in section 3. An example of fast convergence by the generalized transition probability is discussed in section 4 for a parameter range not covered by the theorem in section 3. In section 5 we investigate whether we may further generalize the transition probability in the case of the simple model discussed in section 4. The final section is devoted to discussions on the significance of our result.

2. Inhomogeneous Markov chain

Let us first list various definitions to fix notations (see [3, 4] for a general introduction to inhomogeneous Markov processes used in simulated annealing). We consider a problem of combinatorial optimization with the space of states denoted by $\mathcal{S}$. The size of $\mathcal{S}$ is finite. The *cost function* E is a real single-valued function on $\mathcal{S}$. We assume that the cost function is not a constant. The goal of a combinatorial optimization problem is to find the minimum (or minima) of the cost function. For this purpose we introduce the process of simulated annealing using the Markov chain generated by the *transition probability* from state $x(\in \mathcal{S})$ to state $y(\in \mathcal{S})$ at time step t:

$$G(x, y; t) = \begin{cases} P(x, y)\, A(x, y; T(t)) & (x \neq y) \\ 1 - \displaystyle\sum_{z(\neq x)} P(x, z)\, A(x, z; T(t)) & (x = y) \end{cases} \tag{2.1}$$

where $P(x, y)$ is the *generation probability*

$$P(x, y) \begin{cases} > 0 & (y \in \mathcal{S}_x) \\ = 0 & (\text{otherwise}) \end{cases} \tag{2.2}$$

with $\mathcal{S}_x$ the *neighbourhood* of x (the set of states that can be reached by a single step from x), and $A(x, y; T)$ is the *acceptance probability*. In the case of the generalized transition probability, the acceptance probability is given as [7]

$$\begin{aligned} A(x, y; T) &= \min\{1, u(x, y; T)\} \\ u(x, y; T) &= \left(1 + (q-1)\frac{E(y) - E(x)}{T}\right)^{1/(1-q)} \end{aligned} \tag{2.3}$$

where q is a real parameter. For technical reasons we have to restrict ourselves to the region $q > 1$ in this and the next sections. This acceptance probability reduces to the usual Boltzmann form in the limit $q \to 1$. The present Markov chain is *inhomogeneous*, i.e. the transition probability (2.1) depends on the time step t through the time dependence of $T(t)$.

We choose the *annealing schedule*, or the t dependence of the parameter T (the *temperature*), as

$$T(t) = \frac{b}{(t+2)^c} \qquad (b, c > 0, t = 0, 1, 2, \ldots). \tag{2.4}$$

To analyse the inhomogeneous Markov chain generated by the above transition probability, we introduce the *transition matrix* $G(t)$ with the element

$$[G(t)]_{x,y} = G(x, y; t). \tag{2.5}$$

Let us write the set of probability distributions on $\mathcal{S}$ as $\mathcal{P}$. A probability distribution $p(\in \mathcal{P})$ may be regarded as a row vector with the component $[p]_x = p(x)(x \in \mathcal{S})$. Using this matrix-vector notation, the probability distribution at time step t, starting from an initial distribution $p_0(\in \mathcal{P})$ at time s, is written as

$$p(s, t) = p_0 G^{s,t} \equiv p_0\, G(s)\, G(s+1) \cdots G(t-1). \tag{2.6}$$

The *coefficient of ergodicity* is defined as

$$\alpha(G) = 1 - \min\left\{ \sum_{z \in \mathcal{S}} \min\{G(x, z), G(y, z)\} | x, y \in \mathcal{S} \right\} \tag{2.7}$$

which is a measure of the change of states by a single step.

We shall prove in the next section the property of *weak ergodicity* for the present Markov chain, which means that the probability distribution function after sufficiently long time becomes independent of the initial condition

$$\forall s \geqslant 0 : \lim_{t \to \infty} \sup\{\| p_1(s, t) - p_2(s, t) \| \mid p_{01}, p_{02} \in \mathcal{P}\} = 0 \tag{2.8}$$

where $p_1(s, t)$ and $p_2(s, t)$ are the probability distributions with different initial conditions p_{01} and p_{02}:

$$p_1(s, t) = p_{01} G^{s,t} \tag{2.9}$$

$$p_2(s, t) = p_{02} G^{s,t}. \tag{2.10}$$

The norm is defined by

$$\| p_1 - p_2 \| = \sum_{x \in \mathcal{S}} | p_1(x) - p_2(x) |. \tag{2.11}$$

Although we focus our attention on weak ergodicity in the present paper, it may be useful as a reference to recall the definition of *strong ergodicity*:

$$\exists r \in \mathcal{P}, \forall s \geqslant 0 : \lim_{t \to \infty} \sup\{\| p(s, t) - r \| \mid p_0 \in \mathcal{P}\} = 0. \tag{2.12}$$

The following theorems give criteria for weak and strong ergodicity [3, 4]:

Theorem 1 (Condition for weak ergodicity). An inhomogeneous Markov chain is weakly ergodic if and only if there exists a strictly increasing sequence of positive numbers

$$t_0 < t_1 < \cdots < t_i < t_{i+1} < \cdots$$

such that

$$\sum_{i=0}^{\infty} (1 - \alpha(G^{t_i, t_{i+1}})) = \infty. \tag{2.13}$$

Theorem 2 (Condition for strong ergodicity). An inhomogeneous Markov chain is strongly ergodic if it satisfies the following conditions:

(i) it is weakly ergodic
(ii) there exists $p_t \in \mathcal{P}(\forall t \geqslant 0)$ such that $p_t = p_t G(t)$
(iii) p_t satisfies

$$\sum_{t=0}^{\infty} \| p_t - p_{t+1} \| < \infty. \tag{2.14}$$

3. Weak ergodicity

We prove in the present section that the condition in theorem 1 is satisfied by the present inhomogeneous Markov chain generated by the generalized transition probability. The argument closely follows that for the conventional Boltzmann-type transition probability [3–5]. We need the following Lemma for this purpose.

Lemma 1 (Lower bound on the transition probability.). The elements of the transition matrix of the inhomogeneous Markov chain defined in section 2 satisfy the following bounds. For off-diagonal elements,

$$P(x, y) > 0 \quad \Rightarrow \quad \forall t \geqslant 0 : G(x, y; t) \geqslant w \left(1 + \frac{(q-1)L}{T(t)} \right)^{1/(1-q)} \tag{3.1}$$

and for diagonal elements,

$$\forall x \in \mathcal{S} - \mathcal{S}^M, \exists t_1 > 0, \forall t \geqslant t_1 : G(x, x; t) \geqslant w \left(1 + \frac{(q-1)L}{T(t)} \right)^{1/(1-q)} \tag{3.2}$$

where $\mathcal{S}^M$ is the set of locally maximum states

$$\mathcal{S}^M = \{ x \mid x \in \mathcal{S}, \forall y \in \mathcal{S}_x : E(y) \leqslant E(x) \} \tag{3.3}$$

and L denotes the maximum change of the cost function by a single step

$$L = \max\{ |E(x) - E(y)| \mid P(x, y) > 0 \} \tag{3.4}$$

and w is the minimum value of $P(x, y)$

$$w = \min\{ P(x, y) \mid P(x, y) > 0, x, y \in \mathcal{S} \}. \tag{3.5}$$

Proof. First we prove (3.1). When $E(y) - E(x) > 0$, we have $u(x, y; T(t)) \leqslant 1$ and thus

$$\begin{aligned} G(x, y; t) &= P(x, y)\, A(x, y; T(t)) \\ &\geqslant w \ \min\{1, u(x, y; T(t))\} \\ &= w\, u(x, y; T(t)) \\ &\geqslant w \left(1 + \frac{(q-1)L}{T(t)} \right)^{1/(1-q)}. \end{aligned} \tag{3.6}$$

If $E(y) - E(x) \leqslant 0$, $u(x, y; T(t)) \geqslant 1$ and therefore

$$\begin{aligned} G(x, y; t) &\geqslant w \ \min\{1, u(x, y; T(t))\} \\ &= w \\ &\geqslant w \left(1 + \frac{(q-1)L}{T(t)} \right)^{1/(1-q)}. \end{aligned} \tag{3.7}$$

We next prove (3.2). Since $x \in \mathcal{S} - \mathcal{S}^M$, there exist states $y_+ \in \mathcal{S}_x$ satisfying $E(y_+) - E(x) > 0$. For such states y_+,

$$\lim_{t\to\infty} u(x, y_+; T(t)) = 0 \tag{3.8}$$

and consequently

$$\lim_{t\to\infty} \min\{1, u(x, y_+; T(t))\} = 0. \tag{3.9}$$

Then $\min\{1, u(x, y_+; T(t))\}$ can be made arbitrarily small for sufficiently large t. More precisely, there exist $t_1 > 0$ and $0 < \epsilon < 1$ such that

$$\forall t \geqslant t_1 : \min\{1, u(x, y_+; T(t))\} < \epsilon. \tag{3.10}$$

We therefore have

$$\begin{aligned}\sum_{z\in\mathcal{S}} P(x, z)A(x, z; T(t)) &= \sum_{\{y_+\}} P(x, y_+)\min\{1, u(x, y_+; T(t))\} \\ &+ \sum_{z\in\mathcal{S}-\{y_+\}} P(x, z)\min\{1, u(x, z; T(t))\} < \sum_{\{y_+\}} P(x, y_+)\epsilon \\ &+ \sum_{z\in\mathcal{S}-\{y_+\}} P(x, z) = -(1-\epsilon)\sum_{\{y_+\}} P(x, y_+) + 1.\end{aligned} \tag{3.11}$$

The diagonal element of (2.1) thus satisfies

$$\begin{aligned}G(x, x; t) &\geqslant (1-\epsilon)\sum_{\{y_+\}} P(x, y_+) \\ &\geqslant w\left(1 + \frac{(q-1)L}{T(t)}\right)^{1/(1-q)}\end{aligned} \tag{3.12}$$

where we have used that the last factor can be made arbitrarily small for sufficiently large t. □

We use the following notations in the proof of weak ergodicity. The minimum number of state transitions to reach y from x (or vice versa) is written as $d(x, y)$. One can then reach any state from x within $k(x)$ steps:

$$k(x) = \max\{d(x, y) \mid y \in \mathcal{S}\}. \tag{3.13}$$

The minimum of $k(x)$ for $x \in \mathcal{S} - \mathcal{S}^M$ is denoted as R, and the state giving this minimum value is x^*:

$$R = \min\{k(x) \mid x \in \mathcal{S} - \mathcal{S}^M\} \tag{3.14}$$

$$x^* = \arg\min\{k(x) \mid x \in \mathcal{S} - \mathcal{S}^M\}. \tag{3.15}$$

Theorem 3 (Weak ergodicity). The inhomogeneous Markov chain defined in section 2 is weakly ergodic if $0 < c \leqslant (q-1)/R$.

Proof. Consider a transition from state x to x^*. According to the definition (2.6) of the double-time transition matrix, we have

$$G^{t-R,t}(x, x^*) = \sum_{x_1,\cdots,x_{R-1}} G(x, x_1; t-R)\, G(x_1, x_2; t-R+1)\cdots G(x_{R-1}, x^*; t-1). \tag{3.16}$$

From the definitions of x^* and R, there exists at least one sequence of transitions to reach x^* from x within R steps such that

$$x \neq x_1 \neq x_2 \neq \cdots \neq x_k = x_{k+1} \cdots = x_R = x^*. \tag{3.17}$$

If we keep only such a sequence in the summation of (3.16) and use lemma 1,

$$\begin{aligned} G^{t-R,t}(x, x^*) &\geqslant G(x, x_1; t-R)\, G(x_1, x_2; t-R+1) \cdots G(x_{R-1}, x_R; t-1) \\ &\geqslant \prod_{k=1}^{R} w \left(1 + \frac{(q-1)L}{T(t-R+k-1)}\right)^{1/(1-q)} \\ &\geqslant w^R \left(1 + \frac{(q-1)L}{T(t-1)}\right)^{R/(1-q)}. \end{aligned} \tag{3.18}$$

Then the coefficient of ergodicity satisfies

$$\begin{aligned} \alpha(G^{t-R,t}) &= 1 - \min\left\{ \sum_{z \in \mathcal{S}} \min\{G^{t-R,t}(x, z), G^{t-R,t}(y, z)\} \mid x, y \in \mathcal{S} \right\} \\ &\leqslant 1 - \min\{\min\{G^{t-R,t}(x, x^*), G^{t-R,t}(y, x^*)\} \mid x, y \in \mathcal{S}\} \\ &\leqslant 1 - w^R \left(1 + \frac{(q-1)L}{T(t-1)}\right)^{R/(1-q)}. \end{aligned} \tag{3.19}$$

We now use the annealing schedule (2.4). There exists a non-negative integer k_0 such that the following inequalities hold for all $k \geqslant k_0$:

$$\begin{aligned} 1 - \alpha(G^{kR-R,kR}) &\geqslant w^R \left(1 + \frac{(q-1)L(kR+1)^c}{b}\right)^{R/(1-q)} \\ &\geqslant w^R \left(\frac{2(q-1)LR^c}{b}\left(k + \frac{1}{R}\right)^c\right)^{R/(1-q)}. \end{aligned} \tag{3.20}$$

It is clear from (3.20) that the summation

$$\sum_{k=0}^{\infty}(1 - \alpha(G^{kR-R,kR})) = \sum_{k=0}^{k_0-1}(1 - \alpha(G^{kR-R,kR})) + \sum_{k=k_0}^{\infty}(1 - \alpha(G^{kR-R,kR})) \tag{3.21}$$

diverges if c satisfies $0 < c \leqslant (q-1)/R$. This proves weak ergodicity according to theorem 1. □

Remark 1. The arguments developed in sections 2 and 3 break down for $q < 1$. For instance, the argument of the outer parentheses on the right-hand side of (2.3) becomes negative for sufficiently small T if $E(y) - E(x) > 0$ and $q < 1$. The acceptance probability is regarded as vanishing in such a case in numerical calculations [7, 8]. However, it is difficult to modify the present proof to adopt this convention used in numerical investigations. Theorem 3 does not exclude the possibility that the present Markov chain is weakly ergodic for $q < 1$ or that it is strongly ergodic for arbitrary q.

Remark 2. The condition for weak ergodicity given in theorem 1 is similar to the condition of 'infinite often in time' used to show convergence under the generalized generation probability in continuum space [6, 7].

Remark 3. Theorem 3 with the annealing schedule (2.4) does not immediately mean a fast convergence of the expectation value of the cost function. We have proved only the convergence in the sense of weak ergodicity, not a *fast* convergence of the expectation value of the cost function. See section 6 for detailed discussions on this point.

4. Case of $q < 1$

It is instructive to investigate a simple solvable model with the parameter $q < 1$ because the general analysis in the previous section excluded this range of q for technical reasons. The one-dimensional model discussed by Shinomoto and Kabashima [10] is particularly suited for this purpose.

They considered the thermal diffusion process of an object in a one-dimensional space. The object is located on one of the discrete positions $x = ai$, with i an integer, and is under the potential $E(x) = x^2/2$. Hoppings to neighbouring positions $i+1$ and $i-1$ take place if thermal fluctuations allow the object to climb over the barriers with height B for the process $i \to i-1$ and height $B + \Delta_i$ for $i \to i+1$, where Δ_i is the difference of the potentials at neighbouring locations $\Delta_i = E(a(i+1)) - E(ai) = ax + a^2/2$. A point to notice here is that the potential barriers are not symmetric around a given point; it has a fixed value B to the left and a position-dependent value $B + \Delta_i$ to the right. The latter barrier may be larger or smaller than the former depending upon the position x of the object. Nevertheless, since we will take the continuum limit $a \to 0$ with fixed x, the difference between the two barriers, Δ_i, becomes very small compared to the fixed value B.

By adaptively optimizing the temperature at each given time, Shinomoto and Kabashima found that the energy (the expectation value of the potential at the position of the object) decreases as $B/\log t$. The optimum annealing schedule $T_{\rm opt}(t)$ was shown to have this same asymptotic behaviour as a function of t. We show in the present section that the generalized transition probability with $q = \frac{1}{2}$ leads to a much faster convergence of the energy.

It should be noted that the analysis of the present section is not an application of the general theorem in the previous section. For example, q is less than 1 here, the number of possible states is not finite (i runs from $-\infty$ to ∞), and the optimal annealing schedule will turn out to be t^{-1}, not t^{-c}. The purpose of the present section is to show the existence of a case, independently of theorem 3, where the generalized transition probability yields a much faster decrease of the temperature and energy.

The problem is defined by the master equation describing the time evolution of the probability P_i that the object is at the ith position at time t:

$$\frac{{\rm d}P_i}{{\rm d}t} = \left(1 + (q-1)\frac{B}{T}\right)^{1/(1-q)} P_{i+1} + \left(1 + (q-1)\frac{B+\Delta_{i-1}}{T}\right)^{1/(1-q)} P_{i-1} - \left(1 + (q-1)\frac{B+\Delta_i}{T}\right)^{1/(1-q)} P_i - \left(1 + (q-1)\frac{B}{T}\right)^{1/(1-q)} P_i. \tag{4.1}$$

We now assume that q satisfies $q = 1 - (2n)^{-1}$ ($n = 1, 2, 3, \ldots$) so that $1/(1-q)$ in the above expression is an even number, which assures that the transition probabilities are positive semi-definite at any temperature.

It is straightforward to show that this master equation reduces to the following Fokker–Planck equation in the continuum limit $a \to 0$

$$\frac{\partial P}{\partial t} = \gamma(T)\frac{\partial}{\partial x}(xP) + D(T)\frac{\partial^2 P}{\partial x^2} \tag{4.2}$$

where

$$\gamma(T) = \frac{1}{T}\left(1 + (q-1)\frac{B}{T}\right)^{q/(1-q)} \tag{4.3}$$

$$D(T) = \left(1 + (q-1)\frac{B}{T}\right)^{1/(1-q)}. \tag{4.4}$$

We have rescaled the time unit by $1/a^2$ as in [10].

Our aim is to find the fastest possible asymptotic decrease of the expectation value of the potential defined by

$$y = \int \mathrm{d}x\, E(x)\, P(x,t) \tag{4.5}$$

by adaptively changing T as a function of time. Differentiating both sides of the definition (4.5) and using the Fokker–Planck equation (4.2), we obtain the following equation describing the time evolution of y:

$$\frac{\mathrm{d}y}{\mathrm{d}t} = -2\gamma(T)\, y + D(T). \tag{4.6}$$

The temperature is adaptively optimized by extremizing this right-hand side with respect to T, yielding

$$\begin{aligned} T_{\mathrm{opt}} &= \frac{2yB + (1-q)B^2}{2y + B} \\ &= (1-q)B + 2qy + \mathrm{O}(y^2). \end{aligned} \tag{4.7}$$

The evolution equation (4.6) then has the asymptotic form

$$\frac{\mathrm{d}y}{\mathrm{d}t} = -\frac{2}{B}\left(\frac{2q}{1-q}\right)^{q/(1-q)} y^{1/(1-q)}. \tag{4.8}$$

The solution is

$$y = B^{q/(1-q)}\left(\frac{1-q}{2q}\right)^{1/q} t^{-(1-q)/q}. \tag{4.9}$$

The optimum annealing schedule (4.7) is now

$$T_{\mathrm{opt}} \sim (1-q)B + \text{constant} \times t^{-(1-q)/q}. \tag{4.10}$$

The asymptotic behaviour of the average position can be calculated in the same way. The result is

$$\begin{aligned} \langle x \rangle &= \int \mathrm{d}x\, x\, P(x,t) \\ &\sim \text{constant} \times t^{-1/2q}. \end{aligned} \tag{4.11}$$

From (4.9) we see that the fastest decrease of the energy is achieved when $q = \frac{1}{2}$. With this value of q,

$$y \sim \tfrac{1}{4} B\, t^{-1} \tag{4.12}$$

$$T_{\mathrm{opt}} \sim \tfrac{1}{2} B + \tfrac{1}{4} B\, t^{-1} \tag{4.13}$$

$$\langle x \rangle \sim \text{constant} \times t^{-1}. \tag{4.14}$$

It may be useful to remark that the non-vanishing value $(1-q)B$ of the temperature (4.10) in the infinite-time limit does not cause troubles. What is required is not an asymptotically vanishing value of the temperature, but that the probability distribution does not change with time in the infinite-time limit. This condition is satisfied if $T=(1-q)B$ as is apparent from (4.2) with (4.3) and (4.4).

The results (4.12) and (4.13) show asymptotic relaxations proportional to t^{-1} which is much faster than those for the conventional transition probability, $B/\log t$ [10]. This result of course depends upon the specific structure of the one-dimensional model. We are not claiming to have shown that the generalized transition probability with $q<1$ always gives a faster decrease of the temperature and energy.

5. More general transition probability

A natural question may arise on how far the arguments in the previous sections depend on the specific form of the acceptance probability (2.3). We investigate this problem for the one-dimensional model treated in the preceding section.

The master equation is now generalized to

$$\frac{\mathrm{d}P_i}{\mathrm{d}t}=f\left(\frac{B}{T}\right)P_{i+1}+f\left(\frac{B+\Delta_{i-1}}{T}\right)P_{i-1}-f\left(\frac{B+\Delta_i}{T}\right)P_i-f\left(\frac{B}{T}\right)P_i. \tag{5.1}$$

The same Fokker–Planck equation (4.2) is derived in the limit $a\to 0$ with the following parameters:

$$\gamma(T)=-\frac{1}{T^2}f'\left(\frac{B}{T}\right) \tag{5.2}$$

$$D(T)=f\left(\frac{B}{T}\right). \tag{5.3}$$

The expectation value of the potential obeys the same evolution equation as in (4.6):

$$\frac{\mathrm{d}y}{\mathrm{d}t}=-2\gamma(T)\,y+D(T)=\frac{2y}{T}f'\left(\frac{B}{T}\right)+f\left(\frac{B}{T}\right)\equiv\mathcal{L}\left(\frac{1}{T}\right). \tag{5.4}$$

Minimization of $\mathcal{L}(v)(v=1/T)$ with respect to T for given y leads to

$$2vyB\,f''(Bv)+(2y+B)\,f'(Bv)=0. \tag{5.5}$$

The solution of this equation for v gives the optimal annealing schedule

$$\frac{1}{T_{\mathrm{opt}}}=v=g(y). \tag{5.6}$$

Assuming analyticity of $g(y)$ as $y\to 0$, we write (5.6) as

$$v=c_1+c_2y+\mathrm{O}(y^2). \tag{5.7}$$

It is required that the system stops its time evolution as $y\to 0$ and $v\to c_1$. We then have $f(Bc_1)=0$ from (5.4) assuming c_1 is finite. (This condition of $c_1<\infty$ is not satisfied by the conventional Boltzmann-type acceptance probability in which $1/v=T\to 0$ ($c_1\to\infty$) as $y\to 0$.) It is also necessary that the minimization condition (5.5) is satisfied in the same limit, leading to $f'(Bc_1)=0$. These two conditions on f and f' are satisfied if $f(Bv)$ $(=f(Bc_1+Bc_2y))$ and its derivative behave for small y as

$$f(Bv)\sim c_3y^k \qquad f'(Bv)\sim -c_4y^{k-1} \tag{5.8}$$

where $k > 1$ and $c_3, c_4 > 0$. The minus sign in front of c_4 comes from the observation that an increase of the inverse temperature $v = 1/T$ means a decrease of the energy y and therefore the differentiations by v and y should be done with the opposite sign (i.e. $c_2 < 0$).

The evolution equation (5.4) then has a form

$$\frac{\mathrm{d}y}{\mathrm{d}t} = -c_5 y^k \tag{5.9}$$

with positive c_5 if $2c_1c_4 > c_3$. This equation is solved as

$$y = (c_5(k-1)t)^{-1/(k-1)} \tag{5.10}$$

which shows a power decay of the expectation value of the cost function.

It is useful to set a restriction on k as in the preceding section for q. The following acceptance probability for $y \to 0$ should be positive for any x:

$$f\left(\frac{B+\Delta_{i-1}}{T}\right) \sim f(Bc_1 + Bc_1\Delta_{i-1}) \sim \left(\tfrac{1}{2}a^2 - ax\right)^k \tag{5.11}$$

where we have used (5.8). This requirement is satisfied if k is a positive even number $k = 2n$. The energy (5.10) then decays as

$$y \sim t^{-1}, t^{-1/3}, t^{-1/5}, \ldots, \tag{5.12}$$

the same formula as in the preceding section. In fact the argument in section 4 is recovered if we choose

$$f(v) = (1 + (q-1)v)^{1/(1-q)}. \tag{5.13}$$

In this way the fast decrease of the energy has been obtained for a very general acceptance probability distribution function satisfying certain analyticity conditions.

6. Discussions

We have proved weak ergodicity of the inhomogeneous Markov process generated by the generalized transition probability under certain conditions on the parameters. For technical reasons we were unable to prove strong ergodicity, or more strongly, convergence to the optimal distribution function. We could not show that the condition (2.14) of theorem 2 is satisfied by the present inhomogenous Markov chain. However, weak ergodicity alone already means that asymptotically the state of the system becomes independent of the initial condition, and it is most likely that such an asymptotic state is the optimal one as mentioned in section 1.

It may useful to discuss how the parameter c appearing in equation (2.4) and theorem 3 depends upon the structure of the problem by taking the example of a 'golf-course-like' energy landscape where there are large flat parts and very narrow regions of low-energy states. The quantity R defined in equation (3.14) may be large in the 'golf-course-like' landscape because the number of elements in the set $\mathcal{S} - \mathcal{S}^M$ is small and thus the chance to find smaller $k(x)$ is limited. Nevertheless, R is still finite and it is possible to find a path mentioned in equation (3.17). For large R, c will be small as is seen in the inequality in the statement of theorem 3. This leads to a slower convergence according to the annealing schedule $T \sim 1/(t+2)^c$. Thus, as one would expect naively, the rate of convergence slows down in such a case although the convergence itself is assured.

Intuitively, if there is a large flat portion in the energy landscape, the particle under stochastic motion will be migrating around in that region for some extended time. The particle, however, will eventually (within a finite amount of time) hit a region with a lower

energy and will move into that region if the annealing schedule is chosen appropriately as stated in theorem 3. If the lowest-energy state also has a flat structure in the energy landscape (the ground-state degeneracy), the particle may continue to follow a free random walk there. This does not cause a problem because the weak ergodicity claims convergence of the probability distribution: the probability distribution P in the above situation is $P = 1/K$, where K is the number of states (degeneracy) in the ground state. In other words, the probability distribution has converged into the uniform distribution over the ground state even though the actual motion of the particle continues.

Let us also comment on computational complexity. The time t_1 necessary for the temperature (2.4) to reach a small specified value δ is obtained by solving the relation $b/t_1^c \sim \delta$ ($c = (q-1)/R$) for t_1:

$$t_1 \sim \exp\left(\frac{k_1 N}{q-1} \log \frac{b}{\delta}\right). \tag{6.1}$$

Here we have set $R = k_1 N$ with N the system size because R defined in (3.14) is roughly of this order of magnitude in many cases. For example, in the problem of spin glasses, one can reach any spin configuration by flipping at most N spins. The corresponding time for the conventional simulated annealing is

$$t_2 \sim \exp\left(\frac{k_2 N}{\delta}\right) \tag{6.2}$$

which has been obtained from $k_2 N/\log t_2 \sim \delta$. A comparison of (6.1) with (6.2) reveals that the coefficient of N in the exponent has been reduced from $1/\delta$ to $\log 1/\delta$ by using the generalized transition probability. In this sense, $t_1 \ll t_2$. Since we have proved theorem 3 under very general conditions on the system (which would include problems with NP completeness), it is not possible to find an algorithm to reach a low-temperature state in polynomial time. The best we could achieve is an improvement of the coefficient in the exponent.

One should be careful that the rapid decrease of the temperature does not immediately mean a rapid decrease of the cost function. This aspect can be checked by comparing the acceptance probability (2.3) at $T = \delta$

$$u_1(T = \delta) \sim \left(\frac{\delta}{(q-1)\Delta E}\right)^{1/(q-1)} \tag{6.3}$$

with the corresponding one for the conventional transition

$$u_2(T = \delta) \sim \exp(-\Delta E/\delta). \tag{6.4}$$

Since $u_1(\delta) \gg u_2(\delta)$ if $\Delta E/\delta \gg 1$, we see that the generalized transition probability at a given temperature has a larger value to induce transitions into states with high values of the cost function than in the case of the conventional one at the same temperature. Thus the expectation value of the cost function may be larger under the generalized transition probability than under the conventional Boltzmann form at the same temperature if one waits sufficiently long until thermal equilibrium is reached. This phenomenon has actually been observed in a numerical investigation under a slightly different (but essentially similar) situation [9].

Therefore, if the expectation value of the cost function is observed in a numerical simulation to indeed decrease rapidly under the generalized transition probability, it would be due to not only the rapid decrease of the temperature but also because the relaxation time is shorter. The conventional transition probability may give a larger possibility for the system to stay longer in local minima with high values of the cost function. A mathematical

analysis of this property of quick relaxation by the generalized transition probability is beyond the scope of the present paper. However, one may naively expect it to arise from the larger probability of climbing over high barriers as discussed above.

It should be remarked that theorem 3 with the annealing schedule (2.4) does not give a practically useful prescription of simulated annealing. In actual numerical simulations one rarely uses such annealing schedules as (2.4) obtained from worst-case estimates. Even exponentially fast decreases of temperature often give satisfactory results in the conventional and generalized methods (see [8] and references in [7]). The significance of theorem 3 is that convergence (in the sense of weak ergodicity) as has also been proved with the annealing schedule (2.4) under the generalized transition (acceptance) probability where only empirical numerical investigations have been carried out without a mathematical guarantee of convergence under any annealing schedule.

Acknowledgments

We thank Dr Naoki Kawashima, Dr Toshiyuki Tanaka, Dr Tsuyoshi Chawanya, Professor Constantino Tsallis and Professor Masatoshi Shiino for discussions and comments. One of the authors (JI) acknowledges support from the Junior Research Associate Programme of RIKEN.

References

[1] Kirkpatrick S, Gelett C D and Vecchi M P 1983 *Science* **220** 671
[2] Cerny V 1985 *J. Optim. Theor. Appl.* **45** 41
[3] Uesaka Y 1993 *Mathematical Foundations of Neurocomputing* (Tokyo: Kindai Kagakusha) ch 4 and references therein (in Japanese)
[4] Aarts E H L and Korst J 1989 *Simulated Annealing and Boltzmann Machines: a Stochastic Approach to Combinatorial Optimization and Neural Computing* (New York: Wiley) ch 3 and references therein
[5] Geman S and Geman D 1984 *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-6** 721
[6] Szu H and Hartley R 1987 *Phys. Lett.* **122A** 157
[7] Tsallis C and Stariolo D A 1996 *Physica* **233A** 395
[8] Penna T J P 1995 *Phys. Rev.* E **51** R1
[9] Andricioaei I and Straub J E 1996 *Phys. Rev.* E **53** R3055
[10] Shinomoto S and Kabashima Y 1991 *J. Phys. A: Math. Gen.* **24** L141