Temperature control for simulated annealing

Toyonori Munakata[1] and Yasuyuki Nakamura[2]
[1]*Department of Applied Mathematics and Physics, Kyoto University, Kyoto 606, Japan*
[2]*Department of Informatics, Nagoya University, Nagoya 464, Japan*
(Received 17 January 2001; revised manuscript received 12 June 2001; published 25 September 2001)

The optimal cooling schedule for simulated annealing is formulated to derive a differential equation for the time-dependent temperature $T(t)$. Based on this equation, the long-term behavior of $T(t)$, entropy production, and the Kullback-Leibler entropy are studied. For some simple examples, such as a many-level system and the small scale traveling salesman problem, the explicit time dependence of the temperature is obtained. Some comments are given on simulated annealing based on Tsallis statistics.

DOI: 10.1103/PhysRevE.64.046127 PACS number(s): 05.10.−a, 02.50.Ey, 02.70.Uu

Simulated annealing is a stochastic technique for searching for the (energy) minima of complex discrete or continuous systems [1]. Because of its generality and simplicity, simulated annealing has been applied to various optimization problems, such as the ground-state energy of spin-glass systems and protein folding in condensed-matter physics and the design of integrated circuits and the traveling salesman problem (TSP) in engineering [2]. The key feature of the annealing algorithm is to utilize thermal noise to allow moves that may lead to an increase of the energy and drive the system out of a local minimum. The strength of the noise, which is measured by the time-dependent temperature $T(t)$, is reduced asymptotically to zero, where the system ceases to change and takes usually a local and occasionally the global minimum state.

From both practical and theoretical standpoints, the optimal cooling schedule plays an important role in simulated annealing and there have been some studies on this issue [2–4]. Some related problems, such as the cooling schedule that ensures global optimization [5] and the residual energy as a function of the cooling rate [6], have also gathered considerable interest, reflecting the general popularity of simulated annealing as a tool for optimization. In this paper we derive a differential equation for the temperature $T(t)$ based on optimal control theory [7] and discuss some aspects of simulated annealing based on this equation.

We start from the master equation for the probability $p(\mathbf{x},t)$ of the system to be found in the state $\mathbf{x}$ at time t. Denoting by $W(\mathbf{x}\rightarrow\mathbf{x}')$ the transition rate from $\mathbf{x}$ to $\mathbf{x}'$, we have

$$\partial p(\mathbf{x},t)/\partial t=\sum_{\mathbf{x}'} D(\mathbf{x},\mathbf{x}')p(\mathbf{x}',t), \tag{1}$$

where $D(\mathbf{x},\mathbf{x}')=W(\mathbf{x}'\rightarrow\mathbf{x})$ for $\mathbf{x}\neq\mathbf{x}'$ and $D(\mathbf{x},\mathbf{x})=-\Sigma_{\mathbf{x}''(\neq\mathbf{x})}W(\mathbf{x}\rightarrow\mathbf{x}'')$. It is convenient to express Eq. (1) as

$$d\mathbf{P}(t)/dt=D(t)\mathbf{P}(t), \tag{2}$$

where $\mathbf{P}(t)$ denotes the column vector with elements $p(\mathbf{x},t)$. We will assume the detailed balance condition

$$e^{-E(\mathbf{x})/T(t)}W(\mathbf{x}\rightarrow\mathbf{x}')=e^{-E(\mathbf{x}')/T(t)}W(\mathbf{x}'\rightarrow\mathbf{x}), \tag{3}$$

where $E(\mathbf{x})$ is the energy (or cost) of the system and the time dependence of $D(t)$ comes from that of the temperature. For later use the equilibrium distribution

$$p_{\mathrm{eq}}(\mathbf{x}|T(t))=\exp[-E(\mathbf{x})/T(t)]/Z_c \tag{4}$$

at temperature $T(t)$ is introduced here.

We try to minimize the expectation value of the energy $\langle E\rangle(\tau)\equiv\Sigma E(\mathbf{x})p(\mathbf{x},\tau)$ at some specified time τ. For the purpose we consider the functional

$$G[\mathbf{P},T,\Lambda]\equiv\int_0^T dt\sum_{\mathbf{x}}\left\{E(\mathbf{x})\sum_{\mathbf{x}'}D(\mathbf{x},\mathbf{x}')p(\mathbf{x}',t)-\Lambda(\mathbf{x},t)\times\left[\partial p(\mathbf{x},t)/\partial t-\sum_{\mathbf{x}'}D(\mathbf{x},\mathbf{x}')p(\mathbf{x}',t)\right]\right\}. \tag{5}$$

The second term with the Lagrange multiplier $\Lambda(\mathbf{x},t)$ represents the constraint that $p(\mathbf{x},t)$ satifies Eq. (1) and the first term is reduced to $\langle E\rangle(\tau)-\langle E\rangle(0)$ which is to be minimized. The variational conditions $\delta G/\delta p(\mathbf{x},t)=0$, $\delta G/\delta T(t)=0$ lead to the following equations:

$$d\mathbf{\Lambda}(t)/dt=-D^T[\mathbf{\Lambda}(t)+\mathbf{E}], \tag{6}$$

$$[\mathbf{E}+\mathbf{\Lambda}(t)]^T[\partial D(t)/\partial T(t)]\mathbf{P}(t)=0, \tag{7}$$

where $\mathbf{\Lambda}(t)$ and $\mathbf{E}$ are the column vectors with elements $\Lambda(\mathbf{x},t)$ and $E(\mathbf{x})$, respectively, and D^T denotes the transpose of D. The condition $\delta G/\delta\Lambda(\mathbf{x},t)=0$ leads to Eq. (1). Equation (6) for the Lagrange multiplier is solved iteratively to be $\mathbf{\Lambda}(t)=-\int_0^t ds\, M^T(s,t)D^T(s)\mathbf{E}$ where we set $\mathbf{\Lambda}(t=0)=\mathbf{0}$ and the matrix $M(s,t)$ is defined by

$$M(s,t)=\sum_{n=0}^{\infty}(-1)^n\int_s^t dt_1\times\int_s^{t_1}dt_2\cdots\int_s^{t_{n-1}}dt_n D(t_n)\cdots D(t_1), \tag{8}$$

which satisfies $M(t,t)=M(s,s)=I$ (a unit matrix) and $\partial M(s,t)/\partial t=-M(s,t)D(t)$ and $\partial M(s,t)/\partial s=D(s)M(s,t)$. Inserting the solution $\mathbf{\Lambda}(t)$ into Eq. (7), we have

$$\mathbf{E}^T(t)[\partial D(t)/\partial T(t)]\mathbf{P}(t)=0, \tag{9}$$

where the time-dependent energy is defined as $\mathbf{E}(t) \equiv M^T(0,t)\mathbf{E}$ or

$$d\mathbf{E}(t)/dt=-D^T(t)\mathbf{E}(t). \tag{10}$$

The time dependence of the temperature $T(t)$ is more explicitly represented by differentiating Eq. (9) with respect to t to finally obtain

$$dT(t)/dt=\mathbf{E}^T(t)[D(t),dD(t)/dT(t)]\mathbf{P}(t)/\{\mathbf{E}^T(t) \times[d^2D(t)/dT^2(t)]\mathbf{P}(t)\}, \tag{11}$$

where $[A,B]\equiv AB-BA$. The closed set of nonlinear equations (2), (10), and (11) is the main result of this paper. It is worthwhile noting that the dynamics described by Eq. (10) is conjugate to that of $\mathbf{P}(t)$, Eq. (2), and it is readily seen that $d\{\mathbf{E}^T(t)\cdot\mathbf{P}(t)\}/dt=0$. The conjugate dynamics Eq. (10) is characterized by its non-negative eigenvalues (if our system is stable). At large time when $p(\mathbf{x}_0,t)\simeq 1$ with $\mathbf{x}_0$ the state of minimum energy, it is expected from the constancy of the inner product $\mathbf{E}^T(t)\cdot\mathbf{P}(t)$ and confirmed numerically that $E(\mathbf{x},t)$ becomes large except at $\mathbf{x}=\mathbf{x}_0$.

As an application of our theory we first consider an N-level system with energy $E_i=\epsilon(i-1)$ $(i=1,2,\ldots,N)$. Transitions are only between neighboring levels and the activation energy to go down (up) is $\Delta(\epsilon+\Delta)$. For $N=2$,

$$W(1\to 2)=\exp[-(\epsilon+\Delta)/T(t)],$$

$$W(2\to 1)=\exp[-\Delta/T(t)]. \tag{12}$$

In this simple model (12) we can calculate the right hand side of Eq. (11) explicitly and, when $T(t)$ becomes small enough so that $T(t)\ll\Delta$, we have

$$dT(t)/dt=\epsilon T(t)^2\exp\{-(\epsilon+\Delta)/T(t)\}/\{\Delta^2 p_2(t) -(\epsilon+\Delta)^2 p_1(t)\exp[-\epsilon/T(t)]\}. \tag{13}$$

If we further assume that annealing is successful, that is, $p_2(t)\simeq\exp[-\epsilon/T(t)]$ and $p_1(t)\simeq 1$, we have

$$dT(t)/dt=-T^2(t)\exp[-\Delta/T(t)]/(\epsilon+2\Delta), \tag{14}$$

which is solved exactly to give

$$T(t)=\Delta/\ln(t/[2+\epsilon/\Delta]). \tag{15}$$

Thus for a two-level system one has to cool the system extremely slowly, in accordance with the classical result [5]. The residual energy, the expectation of the energy at time τ, is simply given by $\langle E\rangle(\tau)=\epsilon p_2(\tau)=\epsilon[\tau/(2+\epsilon/\Delta)]^{-\epsilon/\Delta} \propto\tau^{-\epsilon/\Delta}$. This power law was derived heuristically before [6].

In view of the important role played by Tsallis statistics in optimization and simulated annealing [8], we study how the slow decay of temperature $T(t)$ is modified if one employs Tsallis instead of Gibbs statistics. According to Tsallis statistics [9], the equilibrium distribution is given by

$$p_{\rm eq}(\mathbf{x}:T)=c/[1+(q-1)E(\mathbf{x})/T]^{1/(q-1)}, \tag{16}$$

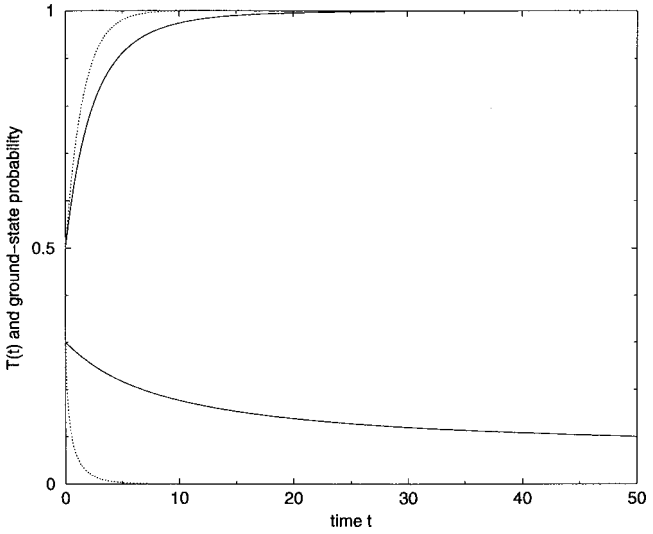

FIG. 1. Temperature $T(t)$ and $p_1(t)$, the probability to be in the ground state 1, for the Gibbs (full curves) and Tsallis (dotted curves) statistics. $T(0)=0.3$, $p_1(0)=0.5$ and $\Delta=0.25$, $\epsilon=1.0$.

with c a normalization constant. We consider the two-level system as before. To satisfy the detailed balance condition, we tentatively consider an additional level 3 with energy $E=\epsilon+\Delta$ and choose the following transition rate:

$$W(1\to 2)=p_{\rm pq}(3)/p_{\rm eq}(1)=1/[1+(q-1)(\epsilon+\Delta)/T]^{1/(q-1)}, \tag{17}$$

$$W(2\to 1)=p_{\rm eq}(3)/p_{\rm eq}(2)=\{[1+(q-1)\epsilon/T]/[1+(q-1) \times(\epsilon+\Delta)/T]\}^{1/(q-1)}. \tag{18}$$

If we take the limit $q\to 1$ we recover the Gibbsian result Eq. (12). The right hand side of Eq. (11) is calculated easily as in the Gibbsian case and after lengthy calculations we have, corresponding to Eq. (14),

$$dT(t)/dt=-\alpha T(t), \tag{19}$$

with $\alpha=[\epsilon/(\epsilon+\Delta)]^{1/(q-1)}(q-1)/(2-q)$. Thus for $1<q<2$ we have an exponential decay of the optimal temperature. This rapid decay of $T(t)$, compared with the logarithmic decay in the Gibbsian case, can be profitably used in simulated annealing [8]. In Fig. 1 we show the optimal $T(t)$ and the ground-state probability $p_1(t)$ for the Gibbsian case (12) and the Tsallis case (17) and (18). The temperature variation is found to precisely follow the theoretical predictions (15) and (19) for large time $t\gg 1$.

Here we note that the exponential asymptotic cooling (19) is different from the well-known power law asymptotic cooling for generalized simulated annealing [8(a)] and we briefly touch upon this point. In Ref. [10] it is shown that the generalized simulated annealing with power law cooling [8(a)] satisfies the (weak) ergodicity property. That is, the final state arrived at by the simulated annealing is independent of the initial distribution function. On the other hand, in our paper we do not claim that we give a sufficient or ergodicity condition for successful simulated annealing. Instead, we

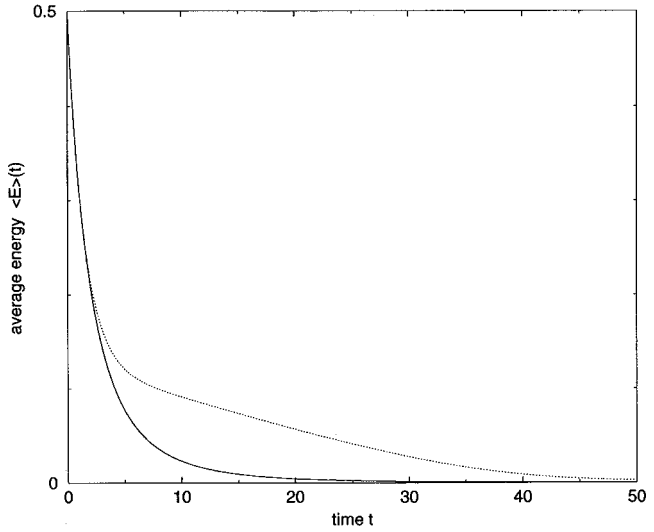

FIG. 2. Average energy $\langle E\rangle(t)$ from the optimal temperature $T(t)$ (full curve) and from the linear temperature $T_L(t)=T(0)-at$ (dotted curve) for a two-level system ($\Delta=0.25$, $\epsilon=1.0$).

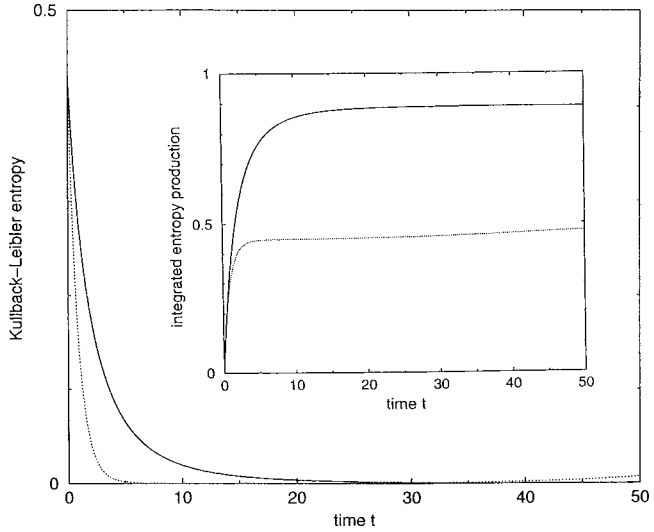

FIG. 3. Kullback-Leibler entropy $S_{\mathrm{KL}}(t)$ from the optimal $T(t)$ (full curve) and from the linear variation $T_L(t)=T(0)-at$ (dotted curve). Inset: Integrated entropy production $\int_0^t dt'\sigma(t')$ from the optimal $T(t)$ (full curve) and from the linear variation $T_L(t)=T(0)-at$ (dotted curve). The system and the initial conditions are the same as those in Fig. 2.

show that, if the initial distribution and other conditions for the system parameters happen to be properly chosen, and consequently if the calculation is successful in the sense that $T(t)$ becomes very small for large t, then the temperature variation should be inverse logarithmic and exponential for Gibbsian and generalized simulated annealing, respectively. In this sense the exponential behavior of $T(t)$ [Eq. (9)] is not contradictory with the power law [8(a),10].

We give here two comments: one is on how $\langle E\rangle(t)$ depends on the temperature variation and the other on how the optimal $T(t)$ changes as N increases from 2. For the two-level system considered above $\langle E\rangle(t)$ obtained from the optimal $T(t)$ and from the linear variation $T_L(t)=T(0)-at$, with $T(0)=0.5$ and a determined from $T_L(\tau=50)=T(\tau=50)$, are compared in Fig. 2. Here we took $p_1(0)=0.5$ for both cases. We note that the optimal $T(t)$ depends only on the initial probability distribution $p(\mathbf{x},t=0)$ and temperature $T(0)$, and it is reasonable that $\langle E\rangle(t)$ from the optimal $T(t)$ is always smaller than that from the linear one. For an N-level system (up to $N\simeq 10^3$), we found that the optimal $T(t)$ behaves nearly the same as in the two-level system for long times $t>50$. However, in the range $0<t<50$ cooling becomes slower as N becomes larger. A more complicated many-level system, a small scale TSP, will be touched upon later.

Next we consider the entropy production $\sigma(t)$ and the Kullback-Leibler entropy $S_{\mathrm{KL}}(t)$ in the process of simulated annealing. Introducing the statistical entropy $S_{\mathrm{st}}(T)$ by $S\equiv -\Sigma p(\mathbf{x},t)\ln p(\mathbf{x},t)$, we know that the entropy production $\sigma(t)dt$ in time dt, defined below, is non-negative [11]:

$$dS_{\mathrm{st}}(t)=d\langle E\rangle(t)/T(t)+\sigma(t)dt. \tag{20}$$

The Kullback-Leibler entropy $S_{\mathrm{KL}}(t)$ [12], which is also non-negative, is defined here as a measure of the distance between the actual distribution $p(\mathbf{x},t)$ and the equilibrium distribution at temperature $T(t)$, Eq. (4),

$$S_{\mathrm{KL}}(t)=\Sigma p(\mathbf{x},t)\ln[p(\mathbf{x},t)/p_{\mathrm{eq}}(\mathbf{x}:T(t))]. \tag{21}$$

$\sigma(t)$ and $S_{\mathrm{KL}}(t)$ are related to each other through

$$S_{\mathrm{KL}}(t)-S_{\mathrm{KL}}(0)=-\int_0^t dt'\sigma(t')-\int_0^t \delta\langle E\rangle \times(t')\frac{dT(t')}{dt'}\Big/T^2(t'), \tag{22}$$

where $\delta\langle E\rangle(t)\equiv\langle E\rangle(t)-\langle E\rangle_{\mathrm{eq}}(t)$ with $\langle E\rangle_{\mathrm{eq}}(t)\equiv\Sigma E(\mathbf{x})p_{\mathrm{eq}}(\mathbf{x}:T(t))$. Since it is expected that $dT(t)/dt<0$ and $\delta\langle E\rangle(t)>0$ in a general annealing process, we see that the first and the second terms of Eq. (22) are negative and positive, respectively. In simulated annealing, it is desirable that the actual distribution $p(\mathbf{x},t)$ precisely follows the equilibrium distribution (4) and in this sense the optimal annealing process represented by Eq. (11) is expected to keep $S_{\mathrm{KL}}(t)$ small. We note that if $p(\mathbf{x},t)\simeq p_{\mathrm{eq}}(\mathbf{x}:T(t))$ for any t, it is proved that $\sigma(t)=0$ and also $\delta\langle E\rangle(t)=0$; thus $dS_{\mathrm{KL}}(t)/dt=0$ (as it should). However, in actual simulated annealing the cooling rate $|dT(t)/dt|$ cannot be zero and the system has to go through an irreversible process in which $\sigma(t)>0$. The question we would like to address is how $S_{\mathrm{KL}}(t)$ is kept small, whether the degree of irreversibility represented by $\sigma(t)$ is made small or the positivity of $\sigma(t)$ is used in order to counteract the second term of Eq. (22). We compare in Fig. 3 $S_{\mathrm{KL}}(t)$ for the two temperature variations represented in Fig. 2. Also shown (inset) is the integrated entropy production $\int_0^t\sigma(t')dt'$, from which we notice that entropy production $\sigma(t)$ is larger for optimal control. For $t<20$, $S_{\mathrm{KL}}(t)$ is definitely larger for the optimal case than the linear case. However, the linear $S_{\mathrm{KL}}(t)$ begins to increase

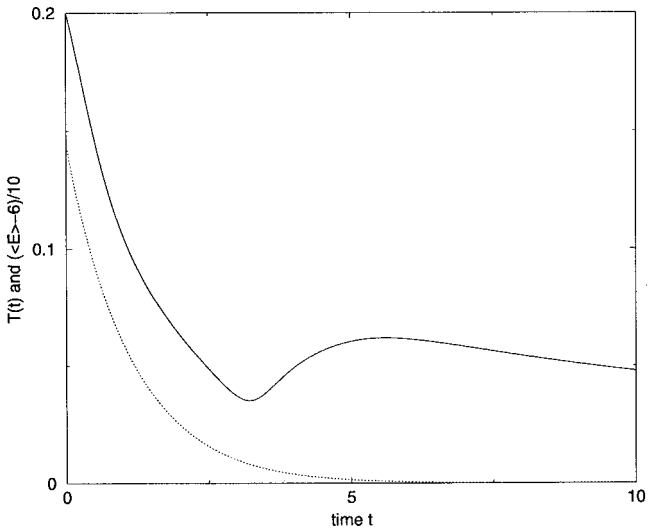

FIG. 4. $T(t)$ (full curve) and $\langle E\rangle(t)$ (dotted curve) for the TSP with $M=6$ cities. Note that actually $[\langle E\rangle(t)-E_{\min}]/10$ is plotted instead of $\langle E\rangle(t)$.

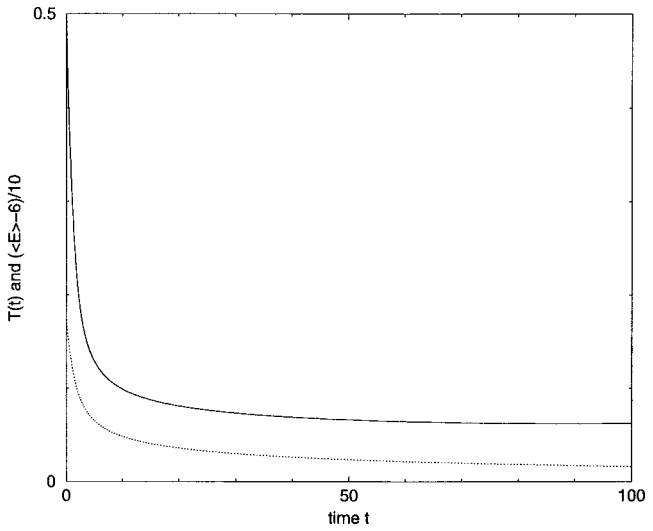

FIG. 5. The same as Fig. 4 except that the activation energy $\Delta=0.3$ is introduced and the initial temperature $T(0)$ is slightly higher.

around $t\simeq 30$, signaling that the function $p(\mathbf{x},t)$ deviates from the equilibrium one, Eq. (4). In passing it is noted that for the N-level model this increase of $S_{\mathrm{KL}}(t)$ seems to be a general tendency for cooling schemes other than the optimal one.

Up to now we have been mainly concerned with a rather simple $N\ (=2)$ -level system, which underlies many complex systems [13]. In order to make contact with Monte Carlo simulations, we apply the optimal cooling schedule to a small scale TSP with $M=6$ cities put at $(0, \pm 0.5)$, $(1, \pm 0.5)$, and $(2, \pm 0.5)$ with the minimum path length $E_{\min}=6$. For the neighborhood structure of the paths we employ the Lin-Kernighan [14] 2-opt transition. Then $\frac{1}{6}$ of the elements of the transition matrix W or $D(t)$ in Eq. (2), which is an $N\times N$ matrix with $N=(M-1)!/2=60$, turn out to be nonzero. In Fig. 4 we plot the optimal $T(t)$ and the average energy obtained from the initial condition $T(0)=0.2$ and $p(i,t=0)\propto\exp[-E_i/T(0)]$. The energy (=path length) is seen to be rapidly approaching $E_{\min}$, although $T(t)$ has a bump around $t=5$. In this case we imposed no activation energy in the transition to states with lower energy (i.e., we used the Metropolis algorithm) in contrast to the N-level system studied before. If we increase the initial temperature $T(0)$, we cannot get a *physical* solution, because the initial (equilibrium) distribution becomes more diffuse and goes out of the basin of the desirable attractor of Eqs. (2), (10), and (11). To cope with this situation we introduced the activation energy $\Delta=0.3$ in the transition to states with lower energy. The result is shown in Fig. 5 for $T(0)=0.5$, where $T(t)$ and also $\langle E\rangle(t)$ change more slowly compared with Fig. 4 but are attracted to the desirable stationary state of the coupled equations (2), (10), and (11). Thus introducing a nonzero Δ seems to extend the basin of the physical attractor with inevitable slow cooling. Here we comment on the utility of our approach in connection with realistic optimization problems. Usually, simulated annealing for optimization problems involves many parameters, the values of which we can choose to achieve good performance. Before embarking on simulated annealing for the large scale problem at hand, one can guess desirable ranges of each parameter by solving the differential equation (11) for a small scale problem, as we did for the case of the TSP.

In this paper we studied temperature control for simulated annealing and derived a first order differential equation (11) supplemented by Eqs. (2) and (10), whose properties were studied for some simple model systems such as an N-level system and a small scale TSP. Our concern here was directed to conceptual rather than practical aspects of the simulated annealing methods. It is hoped that further studies of the set of equations (11), (2), and (10) will shed more light on technical aspects of simulated annealing.

One of the authors (T.M.) expresses his gratitude to S. Ikemoto for useful discussions at an early stage of this study.

[1] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, Science **220**, 671 (1983).

[2] P. van Laarhoven and E. H. L. Aarts, *Simulated Annealing: Theory and Applications* (Kluwer, Dordrecht, 1992).

[3] K. H. Hoffmann and P. Salamon, J. Phys. A **23**, 3511 (1990).

[4] T. M. Heskes, E. T. P. Slijpen, and B. Kappen, Phys. Rev. E **47**, 4457 (1993).

[5] S. Geman and D. Geman, IEEE Trans. Pattern Anal. Mach. Intell. **6**, 721 (1984).

[6] D. A. Huse and D. S. Fisher, Phys. Rev. Lett. **57**, 2203 (1986); Y. Kabashima and S. Shinomoto, J. Phys. Soc. Jpn. **60**, 3993 (1991).

[7] H. Tolle, *Optimization Methods* (Springer, Berlin, 1975).
[8] (a) C. Tsallis and D. A. Stariolo, Physica A **233**, 395 (1996); (b) I. Andricioaei and J. E. Straub, J. Chem. Phys. **107**, 9117 (1997); U. H. E. Hansmann and Y. Okamoto, Phys. Rev. E **56**, 2228 (1997).
[9] C. Tsallis, J. Stat. Phys. **52**, 479 (1988). For a review covering the area of computational methods and simulation, see *Nonextensive Statistical Mechanics and Its Applications*, Vol. 560 of *Lecture Notes in Physics*, edited by S. Abe and Y. Okamoto (Springer, Berlin, 2001).
[10] H. Nishimori and J. Inoue, J. Phys. A **31**, 5661 (1998).
[11] T. Munakata, A. Igarashi, and T. Shiotani, Phys. Rev. E **57**, 1403 (1998).
[12] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).
[13] S. A. Langer and J. P. Sethna, Phys. Rev. Lett. **61**, 570 (1988).
[14] S. Lin and B. W. Kernighan, Oper. Res. **21**, 498 (1973).