Role of noises in neural networks

Sergio Albeverio
Mathematisches Institut, Ruhr-Universität Bochum, Postfach 102148, D-44780 Bochum 1, Germany and Bielefeld—Bochum Stochastics, Sonderforschungsbereich 237 Bochum-Essen-Düsseldorf, Centro di Ricerche in Fisica e Matematica, Locarno, Switzerland

Jianfeng Feng*
Mathematisches Institut, Universität München, Theresienstrasse 39, D-80333, München, Germany

Minping Qian
Department of Probability and Statistics, Peking University, Beijing 100871, China
(Received 10 January 1995)

The important role played by noise in retrieving memories in a network is discussed. Two-stage annealing is proposed to retrieve memories stored in a neural network. The network, undergoing two-stage annealing, behaves like a nonergodic system, and the noise helps the network to select memory inside a local region in which the initial stimuli drop. Theoretical and numerical results of two-stage annealing are presented, and some further possible applications are pointed out.

PACS number(s): 87.10.+e, 05.40.+j

I. INTRODUCTION

The human brain is extremely superior to a digital computer at many tasks. It is not only much faster than any artificial intelligence (AI) system, but also very robust and flexible. During the last two decades, an explosive growth of experimental and theoretical efforts toward a better understanding of the behavior of such a complex system as the human brain has been made. One of the major impulses was the recent conception of the memory-content addressable (or associative) memory proposed by Hopfield and others, which accounted for the collective behavior of large interconnected neural networks. In these models, the information is stored in a system by changing the interaction between neurons such that the system has local stable points or attractors which represent the information stored. The retrieving process starts from a point of the state space of the system, which represents a partial information of the memory; then the time evolution of the system brings it to the attractor representing the complete information. This totally different idea from types of memories such as notebooks, disks, etc., which can only be retrieved under the supervision of a list, index, or by specific instructions, seems much closer to human brains, and provides a general model for pattern recognitions, control problems, and optimizations. Many authors have studied, analyzed, and generalized this model, and for more details the reader is referred to [1–6] and the references given therein.

As many authors have pointed out, noise is inevitable and even necessary as a random driving force [2–3,5–8]. However, as a general result in mathematics we know that under the perturbation of any slightly time invariant noise, the system goes to states with the same distribution, no matter where it started from; i.e., the system is ergodic. Therefore, it seems that the model is very unrobust under random perturbations, and this is just opposite to the way the brain works. A natural question, then, is how *the ergodicity breaks down.* Amit [2] considered a system with infinitely many neurons, and took different phases instead of stable attractors. Therefore, the infinite system could be nonergodic under the assumption of the symmetry of interaction between neurons. However, in this way, the memory capacity of the system seems relatively too small with respect to the size of the system, and the assumption above seems to be too restrictive.

Since noise provides a chance for the system to be able to "jump out" of local minima traps, decreasing of noise is used in the optimization program, i.e., by simulated annealing techniques. A more striking innovation by Lewenstein and Nowak [9,10]—that for allowing the model neural networks systematically to *adjust their own precision and noise level controlling*—is successfully used during the dynamical evolution of the system.

In this paper, we aim to account for the role of noises in general network systems. In fact, the noise is not only a perturbation but also a *controllable driver* for the system *breaking down the ergodicity* and *evolving at a different precision.*

We are naturally curious about why and how this mechanism works so well. Is there any basic general rule behind these phenomena? What parameters of the system would determine its different limit behavior? Let us consider two models.

(I) A system of N two-state neurons, the state of the ith neuron being denoted by a variable $x(i)$ which takes the value 1 when i is firing and -1 when it is inhibited. A state x of the system is a variable, taking values in the set

*On leave from the Department of Probability and Statistics, Peking University.

of vertices of an N-dimensional hypercube $X=\{-1,1\}^N$. In the deterministic, noiseless case, the evolution of the system starting in a state $x=(x(1),\ldots,x(N))\in X$ is governed by a parallel discrete time threshold dynamics ϕ on X, which is generated by neuronal interactions through a given firing function $H(x)=(h(x)(i);i=1,\ldots,N)$, in accordance with the threshold conditions

$$\phi x(i)=\theta(h(x)(i)),\quad i=1,\ldots,N\ , \tag{1.1}$$

where $\theta(u)$ is the step function (taking the value 1 if $u\geq 0$, and -1 otherwise).

For example, when the interaction acts pairwise between neurons, we have

$$h(x)(i)=\tfrac{1}{2}\sum_j w_{ij}x(j)-\theta(i)\ , \tag{1.2}$$

where $w_{ij},\theta(i)$ are given real values $i,j=1,\ldots,N$, cf., e.g., [1–3,9,11].

The stochastic dynamics of the above system is defined by the following transition probability for a flip of the ith neuron from $x(i)$ to $y(i)$:

$$P[y(i)=1]=\frac{1}{1+\exp[-2\beta h(x)(i)]}\ , \tag{1.3}$$

and the transition probability from x to $y=(y(1),\ldots,y(N))$ is

$$P_{xy}(\beta)=\prod_{i=1}^{N}\{1+\exp[-2\beta y(i)h(x)(i)]\}^{-1}\ , \tag{1.4}$$

where $\beta\geq 0$ is the inverse temperature which can be thought of as *a quantity related to the concentrating level of thinking or attention* [12].

(II) For the case of a continuous model, we consider an ordinally differential equation (ODE)

$$dx_t/dt=f(x_t),\quad x_t=(x_t(1),\ldots,x_t(N))\ , \tag{1.5}$$

or a difference equation

$$x_{n+1}=x_n+F(x_n)\ , \tag{1.6}$$

as, e.g., in [13,14,8], where f and F are two given $\mathbb{R}^N$-valued measurable functions.

The random perturbation of (1.5) by a white noise can be the stochastic differential equation (SDE) (called the Langevin dynamics in physics)

$$dx_t=f(x_t)dt+\epsilon dw_t\ , \tag{1.7}$$

(more generally in the case of a colored noise)

$$dx_t=f(x_t)dt+\epsilon\sigma(x_t)dw_t\ , \tag{1.8}$$

or

$$x_{n+1}=x_n+F(x_n)+\epsilon g(\xi_n)\ , \tag{1.9}$$

where w_t is a Brownian motion on $\mathbb{R}^N$, ξ_n is a $\mathbb{R}^N$-valued noise, ϵ is a non-negative constant, and σ and g are two real-valued measurable functions on $\mathbb{R}^N$.

In Sec. II, examples of form (I) are discussed. By a suitable n-dependent choice of β the detailed limit behavior of the system is analyzed.

In Sec. III, a two-stage annealing model is discussed. It is shown that by a suitable n-dependent choice of the temperature, $\beta(n)=\gamma\ln(n+n_0)$, with γ the cooling rate, $\beta(0)=\gamma\ln(n_0), n_0>1$ the initial inverse temperature: there exist a series of critical positive values

$$\gamma_1<\gamma_2<\cdots<\gamma_q\ ,$$

and the asymptotes of the system are described by different attractors. In fact, the γ_i's characterize the costs for transitions among configurations and attractors if the system is reversible. In general, they are determined by the action functional (i.e., the Lagrangian) among configurations of the system, and essentially by the firing function and the interactions. In fact, attractors associate with each other differently at different increasing rate γ's of the inverse temperature. When $\gamma>\gamma_q$, all basins of attractors become separate (distinguishable); i.e., they are not reachable from each other with nearly probability 1 as the initial inverse temperature $\beta(0)$ is large enough (see examples in Sec. II). This behavior is very close to that in the deterministic dynamics. When γ becomes smaller and the initial inverse temperature is large enough, some of the transitions become active in the sense that starting in the basin of the ith attractor the process will almost surely reach the jth attractor at some later time. Finally, when γ is small enough, all transitions are active and the system becomes an ergodic one; that is, the limit behavior of the system is independent of its initial condition. This corresponds to the case of simulated annealing. Hence we can put attractors and their basins on a net tree: $\gamma<\gamma_1$ corresponds to its root and $\gamma>\gamma_q$ to its top. $0<\gamma_1<\gamma_2<\cdots<\gamma_q$ are its branching nodes (see also examples in Sec. II). The precision of memory retrieving or pattern recognition can be controlled by γ. This provides a model and an explanation of the association of a content addressable memory. *In fact, the interaction of the network contains not only the simple association of attractors and their basins, but also the association among different attractors* in different γ levels.

In Sec. IV data of computer simulations are presented as a support to our theory. In Sec. V, we will give some discussions on this theory. Some possible applications are discussed.

II. EXAMPLES

The dynamical system perturbed by a noise can be represented as a finite state Markov chain. For simplicity, let us consider only attractors and neglect their basin in the simple examples of this section (see remark 1 of Sec. III for how to obtain examples in this section from the general models defined in Sec. I). Assume the transition matrix at time n is

$$P(\beta(n))=P^{(n,n+1)}(\beta(n))$$

$$=[p_{ij}(\beta(n))]_{i,j=1,\ldots,S}\ ,$$

where S is the number of states (attractors). We assume that the noise level decreases as time goes to infinity, i.e., $\beta(n)=\gamma \ln(n+n_0)$ for the nth step transition, where $n_0>1$.

Example 1. Assume that the transition probability matrix at time n is of the form

$$p_{ij}(\beta(n))=\begin{cases} e^{-f_{ij}\beta(n)} & \text{if } i\neq j,\quad 1\leq i,\quad j\leq S-1 \\ e^{-F_i\beta(n)} & \text{if } j=S,\quad i\neq j \\ 1-\sum\limits_{j\leq S-1} e^{-f_{ij}\beta(n)}-e^{-F_i\beta(n)} & \text{if } i=j\ , \end{cases}$$

where f_{ij} and F_i are positive numbers such that $F_{S-1}>\cdots>F_1>0$. For f_{ij} sufficiently large with respect to F_{S-1} (for all $i,j=1,\ldots,S-1$), we have

$$P(\beta(n))=\begin{bmatrix} 1-e^{-F_1\beta(n)}+* & * & * & \cdots & * & e^{-F_1\beta(n)} \\ * & 1-e^{-F_2\beta(n)}+* & * & \cdots & * & e^{-F_2\beta(n)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ * & * & * & \cdots & 1-e^{-F_{S-1}\beta(n)}+* & e^{-F_{S-1}\beta(n)} \\ * & * & * & \cdots & * & 1+* \end{bmatrix}$$

$$=U\begin{bmatrix} 1+* & * & \cdots & * \\ * & 1-e^{-\beta(n)F_1}+* & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ * & * & \cdots & * \\ * & * & \cdots & 1-e^{-\beta(n)F_{S-1}}+* \end{bmatrix}U^{-1}\ ,$$

where $*$ represents $o(\exp[-\beta(n)F_{S-1}])$, and

$$U=\begin{bmatrix} 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & 1 \\ 1 & 0 & 0 & \cdots & 0 \end{bmatrix}$$

and

$$U^{-1}=\begin{bmatrix} 0 & \cdots & 0 & 1 \\ 1 & \cdots & 0 & -1 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 1 & -1 \end{bmatrix}.$$

Therefore, the m-step transition matrix is

$$\prod_{n=1}^{m} P(\beta(n))=U\begin{bmatrix} 1+* & * & \cdots & * \\ * & \prod\limits_{n=1}^{m}(1-e^{-\beta(n)F_1})+* & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ * & * & \cdots & \prod\limits_{n=1}^{m}(1-e^{-\beta(n)F_{S-1}})+* \end{bmatrix}U^{-1}\ .$$

Set

$$\gamma_{S-i}=\frac{1}{F_i},\quad i=1,\ldots,S-1\ .$$

(1) When $\gamma<\gamma_1=1/F_{S-1}$, we have $\gamma F_i\leq\gamma F_{S-1}<1$, $i=1,\ldots,S-1$, and

$$\prod_{n=1}^{\infty}(1-e^{-F_i\beta(n)})=0\ .$$

Then the m-step transition matrix has the limit

$$\lim_{m\to\infty}\prod_{n=1}^{m}P(\beta(n))=U\begin{bmatrix}1&0&\cdots\\0&0&\cdots\\\vdots&\vdots&\ddots\\0&0&\cdots\end{bmatrix}U^{-1}$$

$$=\begin{bmatrix}0&\cdots&0&1\\0&\cdots&0&1\\\cdot&\cdots&\cdot&\cdot\\0&\cdots&0&1\end{bmatrix}$$

and this means the system always goes to the last attractor no matter where it starts from. *The system is ergodic.* This is actually the case of simulated annealing for the optimization procedure.

(2) When $\gamma_k:=F_{S-k}^{-1}<\gamma<\gamma_{k+1}:=F_{S-k-1}^{-1}$, $\gamma_S=\infty$, and $k=1,\ldots,S-1$ we see that

$$1>\gamma F_{S-k-1}$$
$$>\gamma F_i,(\forall i<S-k)\Longrightarrow\prod_{n=1}^{\infty}(1-e^{-F_i\beta(n)})=0\,,$$

$$1<\gamma F_{S-k}$$
$$\le\gamma F_i,(\forall i>S-k)$$
$$\Longrightarrow\prod_{n=1}^{\infty}(1-e^{-F_i\beta(n)})=a_i(\beta(0))>0\,.$$

Therefore we have

$$\lim_{m\to\infty}\prod_{n=1}^{m}P(\beta(n))=U\begin{bmatrix}1&&&&&&\\&0&&&&&\\&&\ddots&&&&\\&&&0&&&\\&&&&a_{S-k}(\beta(0))&&\\&&&&&\ddots&\\&&&&&&a_{S-1}(\beta(0))\end{bmatrix}U^{-1}$$

$$=\begin{bmatrix}0&&&&&&1\\&\ddots&&&&&\vdots\\&&0&&&&1\\&&&a_{S-k}(\beta(0))&\cdots&0&1-a_{S-k}(\beta(0))\\&&&\vdots&\ddots&\vdots&\vdots\\&&&0&\cdots&a_{S-1}(\beta(0))&1-a_{S-1}(\beta(0))\\0&\cdot&\cdot&0&\cdots&0&1\end{bmatrix}.$$

In this case, only when the system starts from the first and $(n-k-1)$th attractor will it end up at the last attractor. With positive probability $a_{S-K}(\beta(0)),\ldots,a_{S-1}(\beta(0))$, the system remains in the same attractor it started from. Moreover, the probability $a_i(\beta(0)), i\ge S-k$ will be nearly 1 as $\beta(0)$ is large enough.

(3) We put the attractors together, according to the limit behavior, as γ goes to infinity, and we obtain a net tree as in Fig. 1. Attractor S is the strongest one; when $\gamma_1<\gamma<\gamma_2$, only attractor $S-1$ is distinguishable from attractor S and all the others will join attractor S. As γ becomes large, more and more attractors are distinguishable. Only when γ becomes very large, as $\gamma>\gamma_{S-1}$, does every attractor become distinguishable with positive probability.

Example 2. Assume that, as $n\to\infty$,

$$P(\beta(n))=\begin{bmatrix}1-e^{-F_1\beta(n)}+*&e^{-F_1\beta(n)}+*&*&\cdots&*\\*&1-e^{-F_2\beta(n)}+*&e^{-F_2\beta(n)}l+*&\cdots&*\\\vdots&\vdots&\vdots&\ddots&\vdots\\*&*&*&\cdots&e^{-F_{S-1}\beta(n)}+*\\e^{-F_S\beta(n)}+*&*&*&\cdots&1-e^{-F_S\beta(n)}+*\end{bmatrix},$$

where

$$0<F_1<F_2<\cdots<F_S<\infty\,,$$

the transition probabilities at $*$ being $o(\exp[-\beta(n)F_S])$.

Set

$$\gamma_k=\frac{1}{F_{S-k}},\quad k=1,\ldots,S-1\,.$$

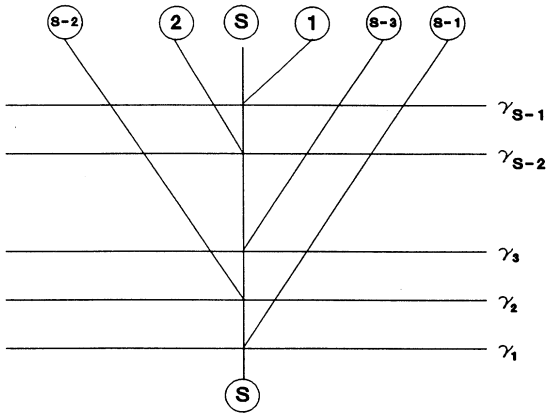

FIG. 1. i inside a circle is the ith attractor.

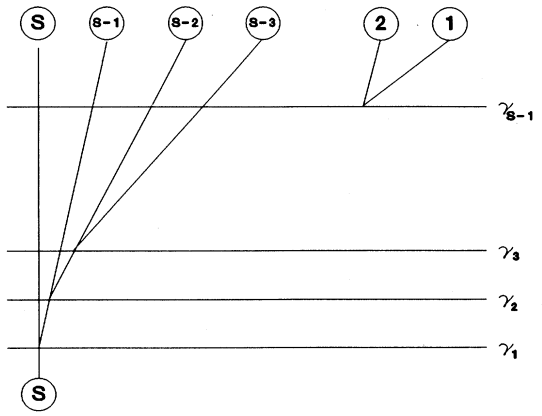

FIG. 2. i inside a circle is the ith attractor.

When $\gamma < \gamma_1$,

$$\prod_{n=1}^{m} P(\beta(n)) \to \begin{bmatrix} 0 & \cdots & 0 & 1 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & 1 \end{bmatrix}$$

as $m \to \infty$ and $\beta(0) \to \infty$; when $\gamma_k < \gamma < \gamma_{k+1}$,

$$\prod_{n=1}^{m} P(\beta(n)) \to \begin{bmatrix} & S-k & & & & & \\ 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ \cdot & \cdots & \cdot & \cdot & 1 & & \\ \cdot & \cdots & \cdot & \cdot & & \ddots & \\ \cdot & \cdots & \cdot & \cdot & & & 1 \end{bmatrix}$$

as $m \to \infty$ and then $\beta(0) \to \infty$; when $\gamma > \gamma_{S-1}$, $m \to \infty$ and then $\beta(0) \to \infty$,

$$\prod_{n=1}^{m} P(\beta(n)) \to \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & 1 \end{bmatrix}.$$

This limit behavior is the same as for deterministic dynamics. The net tree is shown in Fig. 2.

Example 3. If $F_1 = F_2 = \cdots = F_S =: F$, and the transition probability matrix is the same as in example 2,

$$P(\beta(n)) = \begin{bmatrix} 1-e^{-F\beta(n)}+* & e^{-F\beta(n)}+* & * & \cdots & * \\ * & 1-e^{-F\beta(n)}+* & e^{-F\beta(n)}+* & \cdots & * \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ * & * & * & \cdots & e^{-F\beta(n)}+* \\ e^{-F\beta(n)}+* & * & * & \cdots & 1-e^{-F\beta(n)}+* \end{bmatrix}$$

$$= U \begin{bmatrix} 1+* & * & * & \cdots & * \\ * & 1-e^{-F\beta(n)+i(2\pi/S)}+* & * & \cdots & * \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ * & * & * & \cdots & e^{-F\beta(n)+i[2(S-1)\pi/S]}+* \end{bmatrix} U^{-1},$$

where U is a Vandermonde matrix independent of $\beta(n)$, and $*$ is $o(\exp[-\beta(n)F])$. In this case, we have

$$\lim_{m\to\infty} \prod_{n=1}^{m} P(\beta(n)) = \lim_{m\to\infty} \times U \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & \prod_{n=1}^{m} (1-e^{-F\beta(n)+i(2\pi/S)}) & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \prod_{n=1}^{m} (1-e^{-F\beta(n)+i[2(S-1)\pi/S]}) \end{bmatrix} U^{-1}.$$

When $\gamma < 1/F$, the limit matrix is

$$\begin{bmatrix} \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\ \cdot & \cdot & \cdots & \cdot \\ \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \end{bmatrix}$$

and this means that the limit behavior is independent of the initial values.

When $\gamma > 1/F$, the limit behavior is the same as for the deterministic dynamics, that is, the limit matrix nearly approaches the identity matrix when $\beta(0)$ is large enough.

Example 4.

$$P(\beta(n)) = (AB),$$

where A equals

$$\begin{bmatrix} 1-e^{-F_3\beta(n)}+* & e^{-F_3\beta(n)}+* & * & * \\ e^{-F_3\beta(n)}+* & 1-e^{F_3\beta(n)}-e^{-F_2\beta(n)}+* & e^{-F_2\beta(n)}+* & * \\ * & * & 1-e^{-F_3\beta(n)}+* & e^{-F_3\beta(n)}+* \\ e^{-F_2\beta(n)}+* & * & e^{-F_3\beta(n)}+* & C+* \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & e^{-\beta(n)F_1}+* & * & * \end{bmatrix},$$

and B is

$$\begin{bmatrix} * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & e^{-F_1\beta(n)}+* & * & * \\ 1-e^{-\beta(n)F_3}+* & e^{-\beta(n)F_3}+* & * & * \\ e^{-F_3\beta(n)}+* & 1-e^{-\beta(n)F_3}-e^{-\beta(n)F_2}+* & e^{-\beta(n)F_2}+* & * \\ * & * & 1-e^{-F_3\beta(n)}+* & e^{-\beta(n)F_3}+* \\ e^{-F_2\beta(n)}+* & * & e^{-\beta(n)F_3}+* & 1-e^{-F_3\beta(n)}-e^{-\beta(n)F_2}+* \end{bmatrix}$$

for $C = 1-e^{-F_3\beta(n)}-e^{-\beta(n)F_2}-e^{-\beta(n)F_1}$, $*$ being $o[\exp(-\beta F_1)]$, and $F_1 > F_2 > F_3$. Let $\gamma_i = 1/F_i$, $i = 1, 2,$ and 3.

When $\gamma > \gamma_3$, the limit matrix [in the limit that $n \to \infty$ first and then $\beta(0) \to \infty$] is the identity matrix.

When $\gamma_2 < \gamma < \gamma_3$,

$$\lim_{\beta(0)\to\infty} \lim_{m\to\infty} \prod_{n=1}^{m} P(\beta(n)) = \begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ A_{21} & A_{22} & A_{23} & A_{24} \\ A_{31} & A_{32} & A_{33} & A_{34} \\ A_{41} & A_{42} & A_{43} & A_{44} \end{bmatrix},$$

with $A_{11} = A_{22} = A_{33} = A_{44} = \frac{1}{2}\underline{1}$, $\underline{1}$ being the 2×2 matrix with all elements equal to 1, and A_{ij} and $i \neq j$ being identical to zero 2×2 matrices. When $\gamma_1 < \gamma < \gamma_2$,

$$\lim_{\beta(0)\to\infty} \lim_{m\to\infty} \prod_{n=1}^{m} P(\beta(n)) = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix},$$

with $B_{11} = B_{22} = \frac{1}{4}\underline{1}$, $\underline{1}$ being the 4×4 matrix with all elements equal to 1, and $B_{12} = B_{21} = 0$. When $\gamma < \gamma_1$,

$$\lim_{\beta(0)\to\infty} \lim_{m\to\infty} \prod_{n=1}^{m} P(\beta(n)) = \frac{1}{8}\underline{1},$$

where $\underline{1}$ is the 8×8 matrix with all elements equal to 1. The net tree is shown in Fig. 3. Only when $\gamma < \gamma_i$ does the transition path under line γ_i become active.

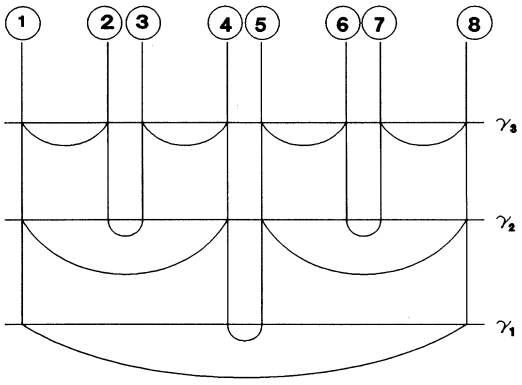

FIG. 3. i inside a circle is the ith attractor.

III. TWO-STAGE ANNEALING

Let $\{\xi_n\}$ be the Markov chain with the transition probability from configuration x-y defined as in Sec. I,

$$P_{xy}(\beta(n))=\prod_{i=1}^{N}\{1+\exp[-\beta(n)F(x,y,i)]\}^{-1} \tag{3.1}$$

at time n, where

$$F(x,y,i)=2y(i)h(x)(i)\neq 0, \quad \forall x,y\in X\ , \quad i=1,\ldots,N\ .$$

For $\beta(n)=\infty$, $n\geq 1$, we define

$$p_{xy}(\infty)=\begin{cases}1 & \text{if } y=\phi x\\ 0 & \text{otherwise}\ ,\end{cases}$$

where ϕ is given by (1.1). Therefore the Markov chain ξ_n is a random perturbation of the deterministic dynamical system ϕ.

Definition 1. A set $A=\{x_1,\ldots,x_p\}\subset X$ is called an attractor of ξ_n if

$$\phi x_i=x_{i+1}, \quad i=1,\ldots,p\ ,$$

where $x_{p+1}:=x_1$. *The attractive basin of an attractor A of ξ_n is defined by*

$$B(A)=\{y\in X,\exists n_0,\ y_0=y,y_1\in X,\ldots,y_{n_0-1}\in X,y_{n_0}\in A \ \text{ satisfying } \phi y_i=y_{i+1},i=0,\ldots,n_0-1\}\ .$$

The definition of the attractor and attractive basin is coincident with the usual definition of the attractor and attractive basin for a deterministic system [6,3]. Let $A_1,\ldots,A_S$ be the set of all attractors of ξ_n.

Let γ be the cooling rate such that $\beta(n)=\gamma\ln(n+n_0)$ for $n_0>1$, and set

$$\begin{aligned}\underline{F}&=\min_{\substack{x,y\in X,x\neq y\\ F(x,y)\neq 0}}F(x,y)\ ,\\ \bar{F}&=\max_{\substack{x,y\in X,x\neq y\\ F(x,y)\neq 0}}F(x,y)\ ,\end{aligned} \tag{3.2}$$

where $F(x,y)=\sum_i[-F(x,y,i)]\vee 0$. For $\gamma>1/\underline{F}$, $\beta(0)\to\infty$, and $n\to\infty$, it follows easily from the definition that the system behaves nearly as in the deterministic case:

$$\lim_{\beta(0)\to\infty}\lim_{n\to\infty}P(\xi_n\in A_i|\xi_0=x)=\begin{cases}1 & \text{if } x\in B(A_i)\\ 0 & \text{if } x\notin B(A_i)\end{cases}$$

for $i=1,\ldots,S$. We call the limit procedure

$$\lim_{\beta(0)\to\infty}\lim_{n\to\infty}$$

a *two-stage annealing,* similar to the concept that first appeared in the paper by van Hemmen *et al.* [15]. For $\gamma<1/\bar{F}$, there exists a subset A_0 of $\cup_{i=1}^{S}A_i$ such that

$$\lim_{\beta(0)\to\infty}\lim_{n\to\infty}P(\xi_n\in A_0|\xi_0=x)=1$$

for all $x\in X$. The limit behavior in this latter case is thus independent of the initial states and we have ergodicity. This property has been successfully used in the simulated annealing. More interesting is what happens when $1/\bar{F}<\gamma<1/\underline{F}$, as we discussed in Sec. II.

Let $\tau(A)$ be the first hitting time of a set A. We have the following bifurcation phenomena; see [7,16–18] for proofs and discussions.

Theorem 1. There exists a sequence of positive numbers

$$\gamma_0=0<\bar{F}^{-1}\leq\gamma_1<\gamma_2<\cdots<\gamma_q\leq\underline{F}^{-1}<\infty=\gamma_{q+1}$$

such that for any $\gamma_{k-1}<\gamma<\gamma_k$, $k=1,\ldots,q+1$ all attractors are divided into groups:

$$R_{1,k},\ldots,R_{s_k,k}\ ,$$

$R_{1,k'},\ldots,R_{s_{k'},k'}$ *are finer than* $R_{1,k},\ldots,R_{s_k,k}$ if $k'>k$. *Furthermore, for $x\in X$,*

(i) $\lim_{\beta(0)\to\infty}\lim_{n\to\infty}P[\tau(A)<\infty|\xi_0=x]=1$ if $A\in R_{i,k}$ and $x\in B(A')$, $A'\in R_{i,k}$;

(ii) $\lim_{\beta(0)\to\infty}\lim_{n\to\infty} P[\tau(A'')=\infty\,|\xi_0=x]=1$ if $A''\notin R_{i,k}$ and $x\in B(A'), A'\in R_{i,k}$;

(iii) $\lim_{\beta(0)\to\infty}\lim_{n\to\infty} P(\xi_n\in \mathcal{S}^k R_{i,k}\,|\xi_0=x)=\begin{cases}1 & \text{if } x\in B(A), A\in R_{i,k}\\ 0 & \text{otherwise},\end{cases}$

where $\mathcal{S}^k R_{i,j}$ is a subset of $R_{i,k}$ and defined in theorem 2 below, and A, A' and A'' are attractors.

Theorem 1 shows that the system is not ergodic as a whole [(ii) of theorem 1], but that there is a kind of locally ergodic phenomena: inside each $R_{i,k}$ the system is ergodic [(i) of theorem 1]. Although it will finally stop in a subset of $R_{i,k}$, i.e., $\mathcal{S}^k R_{i,k}$, the process visits each attractor in $R_{i,k}$ [(i) of theorem 1]. When the process is reversible, the subset $\mathcal{S}^k R_{i,k}$ is just the set of all attractors which attains the global minimum of the energy inside $\cup_{A\in R_{i,k}} B(A)$. Therefore, theorem 1 describes the phenomena of local annealing.

Since $R_{1,k'},\ldots,R_{s_{k'},k'}$ are finer than $R_{1,k},\ldots,R_{s_k,k}$ if $k'>k$, each $R_{i,k}$ is a union of some $R_{j_1,k'},\ldots,R_{j_i,k'}$, and we can think of $R_{i,k}$ has having j_i branches. We can put these branches on a tree such that $R_{i,k}, i=1,\ldots,s_k, k=1,\ldots,q$ are its branching nodes and the root is X, and its top leaves are all basins of attraction, the kth layer of nodes being $R_{1,k},\ldots,R_{s_k,k}$. This tree shows that at different levels of cooling rate γ the system cooperates with different scales. As the cooling rate γ changes, the noise drives the dynamics to visit different groups of attractors.

In examples 1, 2, and 3 of Sec. II we showed that $\mathcal{S}^k R_{i,k}$ is a unique attractor of the process for all k. So we are able to take the attractor to represent $R_{i,k}$ [see (iii) of theorem 1]. Also, trees among $R_{i,k}$ are trees among the attractors which are given in Figs. 1 and 2. In example 4, $\mathcal{S}^k R_{i,k}=R_{i,k},\forall k$. The tree of example 4 (Fig. 3) is among $R_{i,k}$, which is different from that of examples 1, 2, and 3.

$R_{i,k}$, $\mathcal{S}^k R_{i,k}$, γ_k, $i=1,\ldots,s_k$, and $k=1,\ldots,q$ in theorem 1 are determined by $F(x,y,i)$ as follows: Suppose that a function $H(\cdot,\cdot)$ is defined on a finite set $A\times A$. For $a,b\in A$ we say that b is reachable from a at height h [with respect to $H(\cdot,\cdot)$], writing $a\xrightarrow{h} b$, if there exists a sequence $a_0=a,a_1\in A,\ldots,a_{n-1}\in A,a_n=b$ such that

$$H(a_i,a_{i+1})=H_{a_i}=\min_{\substack{c\in A\\ c\neq a_i}} H(a_i,c)\le h\ ,\qquad i=0,\ldots,n-1\ . \tag{3.3}$$

Assume that all attractors of ξ_n are $A_1,\ldots,A_S$,

$$A_i=\{x_1^{(i)},\ldots,x_{k(i)}^{(i)}\},\quad i=1,\ldots,S\ .$$

Define

$$T^{(1)}(i,j)=\min_{\substack{1\le m\le k(i)\\ n\le |X|}}\ \min_{\substack{x_1\in B(A_i),\ldots,x_{n-1}\in B(A_i)\\ x_n\in B(A_j)}}\left[F(x_m^{(i)},x_1)+\sum_{t=1}^{n-1}F(x_t,x_{t+1})\right]$$

if $i\neq j$ and $T^{(1)}(i,j)=0$, and if $i=j$ where $i,j\in\{1,\ldots,S\}$, $|\cdot|$ is the cardinality of a set.

Let

$$\begin{aligned}&R_{i,1}=A_i, i\in\{1,\ldots,S\}\ ,\\ &T^{(1)}=\min_{i=1,\ldots,S} T_i^{(1)}=\gamma_q^{-1}\ ,\\ &\mathcal{S}_1(\cup_{i\in\mathcal{J}}R_{i,1})=\cup_{i\in\mathcal{J}}\mathcal{S}_1R_{i,1}=\cup_{i\in\mathcal{J}}A_i\ \text{ for any set } \mathcal{J}\subset\{1,\ldots,S\}\ .\end{aligned} \tag{3.4}$$

Note that where $T_i^{(1)}$ and $T_i^{(k)}$, $k=2,\ldots,S$ below are defined according to (3.3) with $h=\infty$.

Assume that we have defined $R_{i,k}$, $\gamma_{q-k+1}=(T^{(k)})^{-1}$, $T^{(k)}(i,j)$, $i,j=1,\ldots,s_k$; then we define $R_{i,k+1}$, γ_{q-k}, $T^{(k+1)}(i,j)$, $i,j=1,\ldots,s_{k+1}$ by induction as follows:

$$R_{i,k+1}=\{R_{i,k},T_i^{(k)}>T^{(k)}\}\cup\{R_{j,k},j\xrightarrow{T^{(k)}}i\}\ , \tag{3.5}$$

and $\mathcal{S}_{k+1}$ is a mapping term from $R_{i,k+1}$, $i=1,\ldots,s_{k+1}$ to a subset of $\{R_{1,k},\ldots,R_{s_k,k}\}$,

$$\begin{aligned}&\mathcal{S}_{k+1}(\cup_{i\in\mathcal{J}}R_{i,k+1})=\cup_{i\in\mathcal{J}}\mathcal{S}_{k+1}R_{i,k+1}\ ,\\ &\mathcal{S}_{k+1}R_{i,k+1}=\begin{cases}\{R_{i,k}\} & \text{if } \exists T_i^{(k)}>T^{(k)}, R_{i,k}\in R_{i,k+1}\\ \cup\{R_{j,k}\}=R_{i,k+1} & \text{otherwise},\end{cases}\end{aligned} \tag{3.6}$$

where $\mathcal{J}$ is a subset of $\{1,\ldots,s_{k+1}\}$.

Without loss of generality, we can assume that all effectively distinct sets defined above are $R_{1,k+1},\ldots,R_{s_{k+1},k+1}$. Define

$$T^{(k+1)}(i,j)=\begin{cases}\max\limits_{R_{j,k}\subset R_{i,k+1}} T_j^{(k)}+\min\limits_{\substack{R_{j',k}\subset R_{i,k+1}\\ R_{j'',k}\subset R_{j,k+1}}}[T^{(k)}(j',j'')-T_{j'}^{(k)}] & \text{if } i\neq j\\ 0 \text{ if } i=j\end{cases} \tag{3.7}$$

where $i,j=1,\ldots,s_{k+1}$. The physical meaning of $T^{(k+1)}(i,j)$ is that it measures the lowest cost for the process ξ_n to go from attractor A_i to A_j. When the process ξ_n is reversible, $T^{(n+1)}(i,j)$ is the lowest energy barrier for the process ξ_n to overcome to go from the attractor i to the attractor j. Set

$$\gamma_{q-k}^{-1}=T^{(k+1)}=\min_{i,j=1,\ldots,s_{k+1},i\neq j} T^{(k+1)}(i,j)>T^{(k)}\ . \tag{3.8}$$

The following theorem is proved in [7,16–18] using the large deviation results of Freidlin and Wentzell [19]. An essential idea is contained in examples 1–4 of Sec. II.

Theorem 2. $R_{i,k}$, $\mathcal{S}^kR_{i,k}=\mathcal{S}_1\cdots\mathcal{S}_kR_{i,k}$, $i=1,\ldots,s_k$, $k=1,\ldots,q$, and γ_k, $k=1,\ldots,q$ are determined by the firing function in terms of formulas (3.4)–(3.8), respectively.

Aiming at a clearer understanding of theorems 1 and 2, we turn to example 1 of Sec. II. Under the circumstances of example 1 of Sec. II, by the definition of $T^{(1)}(i,j)$, we have

$$T^{(1)}(i,j)=\begin{cases}F_i & \text{if } j=S,\quad i\neq S\\ f_{ij} & \text{if } j\neq S\ ,\end{cases}$$

for $i,j=1,\ldots,S,i\neq j$ and

$$R_{i,1}=i,\quad i=1,\ldots,S\ .$$

Therefore

$$T^{(1)}=\min_{i=1,\ldots,S} T_i^{(1)}=\frac{1}{\gamma_{S-1}}=F_1\ .$$

Theorem 1 now tells us that

$$\lim_{\beta(0)\to\infty}\lim_{n\to\infty} P(\xi_n=i|\xi_0=i)=1$$

[see (3) of example 1] provided that

$$\gamma_{S-1}<\gamma<\infty\ .$$

This is exactly the same behavior as in the deterministic case ($\beta=\infty$).

Now we are going to define $T^{(2)}(i,j)$, $R_{i,2}$, and $T^{(2)}$. From (3.4) we see that

$$R_{i,2}=\begin{cases}i & \text{if } i=2,\ldots,S-1\\ \{S,1\} & \text{if } i=S\ .\end{cases}$$

For $i,j=2,\ldots,S$, (3.7) now reads

$$T^{(k)}(i,j)=\begin{cases}F_i & \text{if } j=S,\quad i\neq S\\ f_{ij} & \text{if } j\neq S\end{cases}$$

for $i,j=2,\ldots,S,i\neq j$ and

$$S_2\{1,S\}=S,\quad T^{(2)}=F_2=\frac{1}{\gamma_{S-2}}\ .$$

So if $\gamma_{S-2}<\gamma<\gamma_{S-1}$, theorem 1 turns out to be

$$\lim_{\beta(0)\to\infty}\lim_{n\to\infty} P(\xi_n=i|\xi_0=i)=1$$

if $i=2,\ldots,S-1$ and

$$\lim_{\beta(0)\to\infty}\lim_{n\to\infty} P(\xi_n=S|\xi_0\in\{1,S\})=1\ .$$

The above conclusions imply that in this case attractor 1 and attractor S merge together; after taking two-stage annealing, process ξ_n will always go to attractor S, independent of whether process ξ_n starts from attractor S or attractor 1. Also, process ξ_n will remain in the state where it started when $\xi_0=i$, $i\neq 1,S$.

For the general case $k=2,\ldots,S-1$, we have that [see (2) of example 1]

$$R_{i,k}=\begin{cases}i & \text{if } i=k,\ldots,S-1\\ \{S,1,\ldots,k-1\} & \text{if } i=S\ .\end{cases}$$

For $i,j=k,\ldots,S$, after a simple calculation according to (3.7), we obtain that

$$T^{(k)}(i,j)=\begin{cases}F_i & \text{if } j=S,\quad i\neq S\\ f_{ij} & \text{if } j\neq S\end{cases}$$

for $i,j=k,\ldots,S,i\neq j$ and

$$S^k\{1,\ldots,k-1,S\}=S_1\ldots S_k\{1,\ldots,S\}=S\ ,$$

$$T^{(k)}=F_k=\frac{1}{\gamma_{S-k}}\ .$$

Hence from theorem 1, we deduce that when $\gamma_{S-k}<\gamma<\gamma_{S-k+1}$, $\gamma_0=0$, and $k=1,\ldots,S$, theorem 1 turns out to be

$$\lim_{\beta(0)\to\infty}\lim_{n\to\infty} P(\xi_n=i|\xi_0=i)=1$$

if $i=k,\ldots,S-1$ and

$$\lim_{\beta(0)\to\infty}\lim_{n\to\infty} P(\xi_n=S|\xi_0\in\{1,\ldots,k-1,S\})=1\ .$$

The above results tell that in this case the attractor S and

attractors $1,\ldots,k-1$ merge together. Whenever the process ξ_n starts from the attractor $i, i=1,\ldots,k-1$, S, it will finally go to the attractor S (see Fig. 1).

For the continuous model (II) in Sec. I, we have similar results concerning its limit behavior and the tree structure of its attractor: taking the cooling rate such that

$$\gamma^{-1}=\lim_{t\to\infty}\frac{\epsilon^2(t)\ln(t_0+t)}{2},$$

we have a bifurcation as in theorem 1.

Remark 1. It is easy to see that the bifurcation in theorem 1 can be related to the bifurcations for the following singularly nonautonomous ordinally differential equations and system of difference equations:

$$\frac{dP(t,t_0)}{dt}=Q(\beta(t))P(t,t_0),$$

$$P(n+1,n_0)=P(n,n_0)P(\beta(n)),$$

where $P(t,t_0)$ $[P(n,n_0)]$ is the probability matrix at time t $[n]$ starting from time t_0 or n_0, and $Q(\beta(n))$ and $P(\beta(n))$ are the Q matrix [4] at time t and transition probability at time n, respectively. The bifurcation happens for the limit of these equations.

Remark 2. The bifurcation parameters $\gamma_1,\ldots,\gamma_q$ can also be defined by the large deviation rate of the corresponding Markov chain or stochastic differential equation for the stochastic dynamics of the system. This is in accord with the fact that, as already mentioned, the proofs for the theorems in this section are mostly based on the large deviation theory of Freidlin and Wentzell (see [19,7,16,17]).

Remark 3. Models of form (I) are usually called synchronous dynamics (Little-type model). Although we stated our results only for synchronous dynamics, theorems 1 and 2 are true for asynchronous dynamics (Hopfield models) [17,18].

IV. NUMERICAL EXAMPLES

In this section, we follow the notation of Sec. I and take $\beta(n)=\gamma\ln(n+n_0)$, and $\theta_i=0$, $i=1,\ldots,|X|$; here $||$ represents the number of elements of a set. First, we

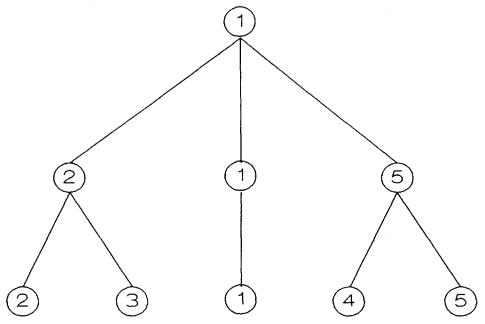

FIG. 4. The tree structure among five attractors in example 1 of Sec. IV.

TABLE I. $T^{(1)}(i,j)$, $i,j=1,\ldots,5$.

	1	2	3	4	5
1	0	17.6	19.5	19.5	17.6
2	6.0	0	7.9	7.5	8.0
3	6.0	9.1	0	11.3	7.0
4	0.5	5.7	11.3	0	1.1
5	0.5	8.0	7.5	4.9	0

TABLE II. $M=N_{22}$, $N_{21}+N_{22}=500$, $\mathcal{P}=M/500$, and $n_0=1500$.

γ	4.5	5.0	5.2	5.8
M	0	18	45	172
$\mathcal{P}$	0.0	3.6	9.0	34.4
γ	6.0	6.05	6.1	
M	233	262	263	
$\mathcal{P}$	46.6	52.4	52.6	
γ	6.2	6.3	6.4	7.0
M	299	329	345	432
$\mathcal{P}$	59.8	65.8	69.0	86.4
γ	8.0	9.0	10.0	
M	481	495	500	
$\mathcal{P}$	96.2	99.0	100	

TABLE III. $M=N_{33}$, $N_{33}+N_{31}=500$, $\mathcal{P}=M/500$, and $n_0=2000$.

γ	4.5	5.0	5.2	5.5
M	1	17	57	108
$\mathcal{P}$	0.2	3.4	11.4	21.6
γ	5.8	5.9	6.0	
M	193	232	246	
$\mathcal{P}$	38.6	46.4	49.2	
γ	6.1	6.2	6.3	6.4
M	266	289	299	316
$\mathcal{P}$	53.2	57.8	59.8	65.2
γ	7.0	8.0	10.0	
M	417	478	500	
$\mathcal{P}$	83.4	95.6	100	

TABLE IV. $M=N_{44}$, $N_{44}+N_{41}=500$, $\mathcal{P}=M/500$, and $n_0=2000$.

γ	0.4	0.45	0.48	0.49	0.5
M	0	48	172	217	253
$\mathcal{P}$	0.0	9.6	34.4	43.3	50.6
γ	0.52	0.54	0.6	0.7	0.8
M	322	373	452	496	500
$\mathcal{P}$	64.6	74.6	90.4	99.2	100

TABLE V. $M=N_{55}$, $N_{55}+N_{51}=500$, $\mathcal{P}=M/500$, and $n_0=2000$.

γ	0.3	0.4	0.46	0.48	0.5
M	0	3	92	164	231
$\mathcal{P}$	0.0	0.6	18.4	32.8	46.2
γ	0.52	0.54	0.6	0.7	0.8
M	304	363	455	496	500
$\mathcal{P}$	60.8	72.6	91	99.2	100

TABLE VI. Energies of the eight attractors.

Att.	1	2	3	4
Energy	0.5091	0.4989	0.5130	0.502 24
Att.	5	6	7	8
Energy	0.4952	0.5033	0.3875	0.5134

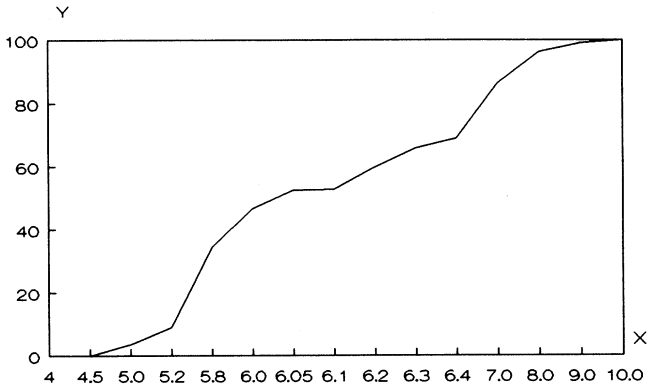

FIG. 5. $Y=\mathcal{P}$ in Table II. $X=\gamma$.

TABLE VII. Overlap of the eight attractors with patterns.

	att. 1	att. 2	att. 3	att. 4
Pat. 1	1.000	0.000	−0.133	−0.026
Pat. 2	−0.013	0.986	0.080	0.026
Pat. 3	−0.120	0.800	0.986	0.026
Pat. 4	−0.026	0.040	0.040	1.000
Pat. 5	0.026	−0.066	−0.040	−0.000
	att. 5	att. 6	att. 7	att. 8
Pat. 1	−0.013	0.026	0.466	0.120
Pat. 2	0.066	−0.053	0.570	−0.066
Pat. 3	0.040	−0.026	0.040	−1.000
Pat. 4	0.053	0.040	0.000	−0.026
Pat. 5	−0.986	1.000	−0.506	0.026

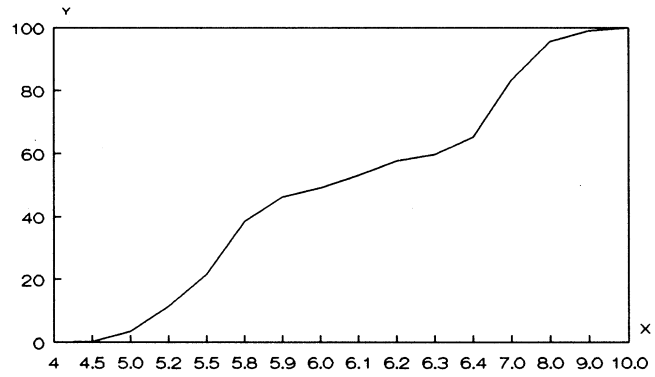

FIG. 6. $Y=\mathcal{P}$ in Table III. $X=\gamma$.

TABLE VIII. Numerical results for $n(i)=|\{n_1:\xi_{n_1}\in A_i;|\xi_0\in B(A_k)\}|$, $i,k=1,\ldots,8$.

	$n(1)$	$n(2)$	$n(3)$	$n(4)$
(0,1)	1033	0	0	0
(0,2)	0	1012	0	0
(0,3)	0	15	1023	0
(0,4)	0	0	0	1046
(0,5)	0	0	0	0
(0,6)	0	0	0	0
(0,7)	1040	10	0	0
(0,8)	0	0	0	0
	$n(5)$	$n(6)$	$n(7)$	$n(8)$
(0,1)	0	0	0	2
(0,2)	0	0	0	0
(0,3)	0	0	0	2
(0,4)	0	0	0	0
(0,5)	1008	0	0	0
(0,6)	0	1027	0	0
(0,7)	0	0	20	0
(0,8)	2	0	0	1019

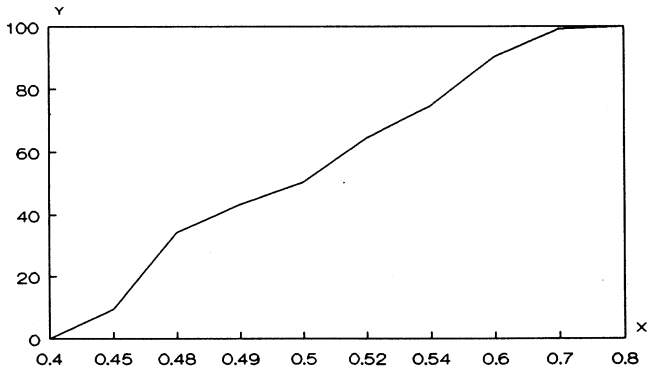

FIG. 7. $Y=\mathcal{P}$ in Table IV. $X=\gamma$.

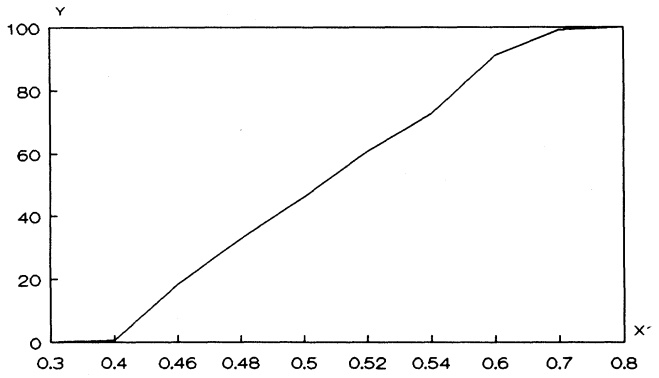

FIG. 8. $Y=\mathcal{P}$ in Table V. $X=\gamma$.

consider a toy model, in which the number of neurons is 10. The reason for considering this model first is that in this extremely simple case, we can calculate γ_i explicitly.

Example 1. Set $N=10$, then $X=\{-1,1\}^{10}$. We assume that $F(x,y,i)$ has the form

$$F(x,y,i)=y(i)\left[\sum T_{ij}x(j)\right] .$$

The connection matrix ($T_{ij}=T_{ji}$, $T_{ii}=0, i,j=1,\ldots,10$) is randomly generated. As a dynamics we take the synchronous dynamics defined by (1.4) (Little network) [2]. There are five attractors and the tree structure among them is shown in Fig. 4. $T^{(1)}(i,j), i,j=1,\ldots,5$ are displayed in Table I (see Sec. III). Here we have $\gamma_1=0.5$ and $\gamma_2=6.0$. As $\gamma<(T^{(1)})^{(-1)}=\gamma_2\approx 6.0$, after performing the two-stage annealing we have

$$N_{21}=|\{n_1:\xi_{n_1}\in A_1|\xi_0\in B(A_2)\}|$$
$$>N_{22}=|\{n_1:\xi_{n_1}\in A_2|\xi_0\in B(A_2)\}|$$

and

$$N_{31}=|\{n_1:\xi_{n_1}\in A_1|\xi_0\in B(A_3)\}|$$
$$>N_{33}=|\{n_1:\xi_{n_1}\in A_3|\xi_0\in B(A_3)\}| ,$$

where $n_1=5000$. Numerical results are given in Tables II and III and Figs. 5 and 6 after simulation 500 times.

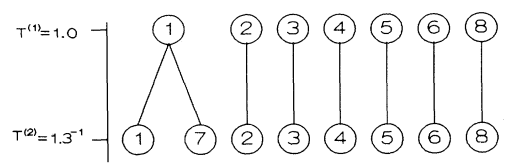

FIG. 9. An explanation of the structure among attractors in example 2 of Sec. IV.

Similarly, as $\gamma<(T^{(2)})^{-1}=\gamma_1\approx 0.5$, after performing the two-stage annealing, we have

$$N_{41}=|\{n_1:\xi_{n_1}\in A_1|\xi_0\in B(A_4)\}|$$
$$>N_{44}=|\{n_1:\xi_{n_1}\in A_4|\xi_0\in B(A_4)\}|$$

and

$$N_{51}=|\{n_1:\xi_{n_1}\in A_1|\xi_0\in B(A_5)\}|$$
$$>N_{55}=|\{n_1:\xi_{n_1}\in A_5|\xi_0\in B(A_5)\}| ,$$

where $n_1=5000$. Tables IV and V and Figs. 7 and 8 are numerical results after simulation 500 times.

Example 2. $N=150$, $X=\{-1,1\}^{150}$, and $F(x,y,i)$ is the same as in example 1 but with asynchronous dynamics (Hopfield model; see remark 4) [2,3]. T_{ij} is given by

$$T_{ij}=\begin{cases}\dfrac{1}{N}\displaystyle\sum_{\mu=1}^{5}\xi^{(\mu)}(i)\xi^{(\mu)}(j) & \text{if } i\neq j\\ 0 & \text{if } i=j ,\end{cases}$$

where $\xi^{(\mu)}=\{\xi^{(\mu)}(i), i=1,\ldots,N\}$, $\mu=1,\ldots,5$ are randomly generated patterns, $n_0=1500$. In our simulation, eight attractors are involved. The seventh attractor is a spurious state [2]. The energy of the eight attractors and their overlap with the patterns are given in Tables VI and VII, respectively. As in example 1, we also let $n_1=5000$. Figure 9 is the tree structure among the eight attractors. As $(1.3)^{-1}<\gamma<1.0$, after performing the two-stage annealing we have

$$|\{n_1:\xi_{n_1}\in A_1|\xi_0\in B(A_7)\}|$$
$$>|\{n_1:\xi_{n_1}\in A_7|\xi_0\in B(A_7)\}| .$$

In Table VIII, we present the numerical results for $n(i)=|\{n_1:\xi_{n_1}\in A_i|\xi_0\in B(A_k)\}|$, $i,k=1,\ldots,8$, and $(0,i)$ representing $\xi_0\in B(A_i)$, $i=1,\ldots,8$ for $\gamma=1$.

V. DISCUSSION

There are several possible applications for the use of a controllable cooling rate as discussed in Sec. I.

(1) Breakdown of ergodicity and kick out of spurious minima. As many authors pointed out, after learning (for instance, by Hebb's or Peretto's law [20]) there are spurious minima in addition to the retrieval patterns which we sought after. Even though their basins are usually smaller than those of the retrieval ones, a number of them could be large. The application of a suitable cooling rate provides a way to avoid the trap by spurious minima, ending in the expected retrieval patterns depending on initial conditions. The ergodicity is then broken as expected. In Fig. 10 the energy landscape is visualized for a system where 1, 2, and 3 are retrieval states, and the other local minima are spurious. Choosing $\beta(n)=\gamma\ln(n+n_0)$ for some $\gamma\geq 0$, $n_0>0$, we can show that there exist $\gamma_1<\gamma_2$ such that if we choose $\gamma\in(\gamma_1,\gamma_2)$ the time-dependent stochastic dynamics will have retriev-

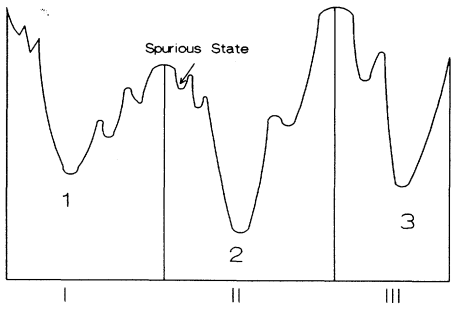

FIG. 10. The time-dependent stochastic dynamics will only stop at states 1, 2, or 3 depending on whether the initial condition belongs to I, II, or III, respectively.

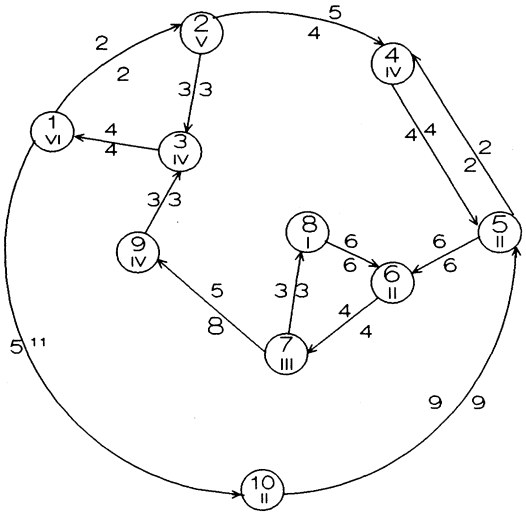

FIG. 11. An example of the architecture of attractors (memories): $1,\ldots,10$ inside a circle represent attractors; I, . . . , VI inside a circle are their levels.

ing only to retrieval states 1, 2, and 3 depending on whether the initial condition belongs to I, II, and III, respectively, without being trapped by the spurious local minima (cf. Fig. 10 and theorem 1 in Sec. III). Many authors worry about spurious minima in systems obtained by successively following laws such as those of Hebb and Peretto, and put great effort into eliminating them. In the case of the Hopfield network, in [21] it is shown how spurious states can be avoided by using a noise of fixed, time-independent strength. Here we describe another way by which one can get rid of spurious minima, even though they exist together with expected retrieval memories. The control is quite robust, since for all γ in the interval $[\gamma_k,\gamma_{k+1}]$ the system has the same behavior. This also seems to agree with the statement of Skarda and Freeman [8] that complex behavior seems random but actually has some hidden order to shift from one complex pattern to another.

(2) An architecture of associative memory. The previous net-tree structure for attractors of the neural network provides an architecture of associative memory. This means that stored patterns are hierarchically distinguished as strongly and weakly memorized states [22–29]. The former corresponds to deeper minima of the energy (or quasipotential in the asymmetric case).

Let us consider a system as in Fig. 11 with ten attractors. $F(x,y)$ is given by the number at the right-hand side of the arrow which starts from x to y; the number at the left-hand side if the arrow starts from x to y is the action functional calculated according to Sec. III, and the others are negligible.

When γ is greater than the number on the left-hand side of the arrow from x to y given by theorem 2, initiating at x, eventually it will reach y (we then write by $x\rightarrow y$) and finally end in the state with lowest energy among the states it can reach.

For instance, when $6>\gamma>4$, with $\beta(0)$ large enough, almost surely $9\rightarrow3\rightarrow1\rightarrow2\rightarrow3$, $4\rightarrow5\rightarrow4$, $6\rightarrow7\rightarrow8$ and the other paths are not possible. This means that attractors are connected in four groups: starting from basins of attractors in $\{9,3,2,1\}$, it will finally home onto state 3; from basins of attractors 4 and 5 onto 5; from basins of $\{6,7,8\}$ onto state 8; and from the basin of attractor 10 it will always be trapped in this basin and finally home onto 10. When $9>\gamma>8$, all paths with arrows can be passed through but for $1\rightarrow10$ and $10\rightarrow5$. In this case, attractors are divided into two groups: state 10 and the group of all others, which, starting in any state except the basin of 10, finally end up at state 8; and which, starting in the basin of 10, always stay in it and finally home onto attractor 10. When $\gamma>9$, it always ends up at attractor 8, the deepest energy well, and this is exactly the case as in the simulated annealing for solving hard optimization problems [30,31].

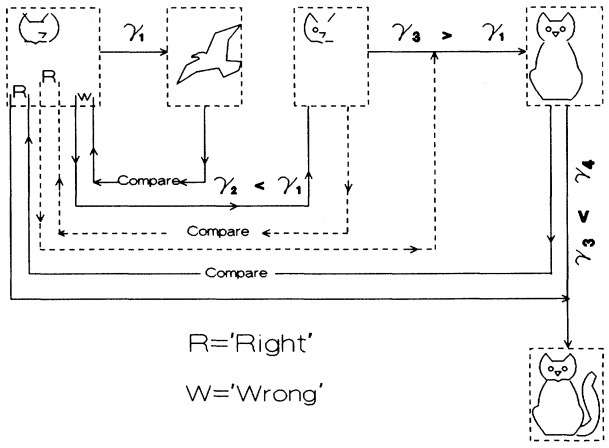

FIG. 12. An explanation for the retrieving of memories.

Thus states associate with each other in different levels. The faster the cooling rate required, the stronger control the network has and more precise recognition (or retrieving) it obtains. For smaller values of the cooling rate γ, the area covered by the process ξ_n increases, and the dependence on the initial condition decreases. That is, the association is hierarchically built up in the network. If the cooling rate is in suitable control, one can control the precision of the recognition as desired.

In the pattern recognition network, one can design the cooling rate differently in different circumstances as in [9,10]. Moreover, for the complex information, i.e., time sequence or cycles of patterns (consider example 4 of Sec. II), our results provide an approach to control the precision, and therefore one can obtain detail as expected by raising or lowering the cooling rate even several times under some comparison of the initial information. Figure 12 roughly illustrates this idea. Moreover, the "general arouse" of Skarda and Freeman plays a role similar to that of the "attention" [described here by $\beta(n)$], cooperating with the initial condition; see [8].

ACKNOWLEDGMENTS

This paper was partially supported by the A. v. Humboldt Foundation of Germany, and the NSF of China (J.F.).

[1] S. I. Amari, Proc. IEEE **78**, 1443 (1990).
[2] D. Amit, *Modeling Brain Function* (Cambridge University Press, Cambridge, UK, 1989).
[3] J. J. Hopfield, Proc. Natl. Acad. Sci. U.S.A. **79**, 2554 (1982).
[4] T. Liggett, *Interacting Particle System* (Springer-Verlag, Berlin, 1986).
[5] P. Peretto, J. Phys. (Paris) **48**, 711 (1988).
[6] J. Hertz, A. Krogh, and R. Palmer, *Introduction to the Theory of Neural Computation* (Addison-Wesley, Reading, MA, 1991).
[7] J. F. Feng, Ph.D. thesis, Peking University, 1991.
[8] C. Skarda and W. Freeman, Behavioral Brain Sci. **6**, 161 (1987).
[9] M. Lewenstein and A. Nowak, Phys. Rev. Lett. **62**, 225 (1989).
[10] M. Lewenstein and A. Nowak, Phys. Rev. A **40**, 4652 (1989).
[11] W. A. Little and G. L. Shaw, Behavioral Biol. **14**, 115 (1975).
[12] *Neural Network for Computing,* edited by J. S. Denker (AIP, New York, 1986).
[13] S. Grossberg, *The Adaptive Brain* (Elsevier, Amsterdam, 1987), Vol. 2.
[14] S. Grossberg, Cognitive Sci. **11**, 23 (1987).
[15] J. L. van Hemmen, D. Grensing, A. Huber, and R. Kühn, J. Stat. Phys. **50**, 231 (1988); **50**, 259 (1988).
[16] J. F. Feng and M. P. Qian, in *Proceeding of the Wuhan Meeting on Probability and Statistics,* edited by A. Badrikian, P.-A. Meyer, and J.-A. Yan (World Scientific, Singapore, 1993), p. 149.
[17] J. F. Feng and M. P. Qian, Acta Appl. Math. (to be published).
[18] J. F. Feng and M. P. Qian, Adv. Math. **23**, 50 (1994).
[19] M. I. Freidlin and A. D. Wentzell, *Random Perturbations of Dynamical Systems* (Springer-Verlag, Berlin, 1984).
[20] M. P. Qian, G. L. Gong, and J. W. Clark, Phys. Rev. A **43**, 1061 (1993).
[21] D. Amit, H. Gutfreund, and H. Sompolinsky, Phys. Rev. A **32**, 1007 (1985).
[22] *Neurocomputing: Foundations of Research,* edited by J. A. Anderson and E. Rosenfeld (MIT Press, Cambridge, MA, 1988).
[23] K. Binder and D. W. Heerman, *Monte Carlo Simulation in Statistical Mechanics* (Springer-Verlag, Berlin, 1988).
[24] J. W. Clark, *Nonlinear Phenomena in Complex System* (North-Holland, Amsterdam, 1989).
[25] J. D. Cowan and D. H. Sharp, Proc. Am. Acad. Arts Sci. **117**, 85 (1988).
[26] S. Geman and D. Geman, IEEE Trans. Pattern Anal. Machine Intelligence **6**, 721 (1984).
[27] D. Ackley, G. Hinton, and T. Sejnowski, Cognitive Sci. **9**, 147 (1985).
[28] J. J. Hopfield and D. W. Tank, Biol. Cybernetics **52**, 141 (1985).
[29] T. Kohonen, *Self-Organization and Associative Memory,* 3rd ed. (Springer-Verlag, Berlin, 1989).
[30] S. Kirkpatrick, C. Gelatt, and M. P. Vecchi, Science **220**, 671 (1983).
[31] P. J. M. van Laarhoven and E. H. L. Aarts, *Simulated Annealing: Theory and Applications* (Reidel, Dordrecht, 1987).