Cooling schedules for learning in neural networks

Tom M. Heskes, Eddy T. P. Slijpen, and Bert Kappen
Department of Medical Physics and Biophysics, University of Nijmegen, Geert Grooteplein Noord 21, 6525 EZ Nijmegen, The Netherlands
(Received 2 November 1992)

We derive cooling schedules for the global optimization of learning in neural networks. We discuss a two-level system with one global and one local minimum. The analysis is extended to systems with many minima. The optimal cooling schedule is (asymptotically) of the form $\eta(t)=\eta^*/\ln t$, with $\eta(t)$ the learning parameter at time t and η^* a constant, dependent on the reference learning parameters for the various transitions. In some simple cases, η^* can be calculated. Simulations confirm the theoretical results.

PACS number(s): 87.10.+e

I. INTRODUCTION

Global optimization of learning in neural networks is currently an important subject. How can one be sure that the learning network reaches the optimal state, i.e., the global minimum of some error criterion, and does not get stuck in a local minimum? A well-known strategy to find the global minimum and not just a local minimum is simulated annealing [1]: a noise parameter, say temperature, is cooled down slowly. In the beginning of the search for the optimal solution, temperature is relatively high and large steps are possible. At the end, when the system is likely to be in the vicinity of the optimal state, temperature is low and only small steps are made.

On-line learning in neural networks is also a stochastic process. At each learning step, a training pattern is drawn *at random* from the environment (the total set of training patterns) and presented to the network. The learning parameter sets the typical scale of the weight change at each update. A large learning parameter leads to large fluctuations in the network's representation [2]. So, in a way, the learning parameter can be viewed as a noise parameter akin to the temperature in simulated annealing. Therefore it seems worthwhile to search for cooling schedules for the learning parameter that guarantee convergence to the optimal network state.

Usually, simulated annealing techniques are applied to stochastic processes for which the stationary probability distribution for a fixed value of the noise parameter is a Gibbs distribution. Well-known examples are Langevin algorithms for diffusion-type processes [3] and annealing algorithms for combinatorial optimization [1]. However, for stochastic learning processes, the stationary distribution is in general unknown and is not a simple Gibbs distribution [4,5]. This makes it more difficult to find a cooling schedule for the learning parameter.

Roughly speaking, there are two different approaches to study the consequences of the noise introduced by the random presentation of patterns. The "mathematical" approach describes learning in the context of stochastic approximation theory and has led to many important, rigorously proven theorems (see, e.g., [6,7]). More specifically, Kushner [8] describes a cooling schedule of the type we will derive, and shows that it leads to global optimization if one of the parameters in this schedule is chosen large enough. We will try to derive the optimal cooling schedule, i.e., the cooling schedule that leads to the optimal network state, not only with probability 1, but also as fast as possible. To this end, we will follow the "physical" approach which treats learning as a stochastic process governed by a master equation. The main benefit of this approach is its applicability if one aims at (approximate) quantitative results (see, e.g., [9,2]).

In Sec. II we will briefly summarize the results of a previous study [5] that are essential for the rest of this paper. These results will be used in Sec. III to derive a cooling schedule for a two-level system with one global and one local minimum. The two-level case is generalized to various minima in Sec. IV. The simulations in Sec. V will be used to test the derived cooling schedules. In Sec. VI we will discuss the possible applications and the limitations of the results.

II. LEARNING WITH LOCAL MINIMA

At every learning step, a training pattern, denoted by an n-dimensional vector $\vec{x}$, is drawn at random from the environment Ω and the N-dimensional weight vector $\mathbf{w}$, containing the strength of all synapses and thresholds, changes its state from $\mathbf{w}$ to $\mathbf{w}+\Delta\mathbf{w}$, obeying

$$\Delta\mathbf{w}=\eta\mathbf{f}(\mathbf{w},\vec{x}) , \tag{1}$$

with the learning parameter η and the learning rule $\mathbf{f}(\mathbf{w},\vec{x})$. We will restrict ourselves to learning rules that perform stochastic gradient descent on some error function $E(\mathbf{w})$, i.e., that obey

$$\langle \mathbf{f}(\mathbf{w},\vec{x})\rangle_\Omega=-\nabla E(\mathbf{w}) ,$$

where $\langle\ \rangle_\Omega$ stands for the average over all training patterns and ∇ for the derivative with respect to the network state $\mathbf{w}$. The existence of such an error potential $E(\mathbf{w})$ facilitates a global description of the learning process: the lower the error potential $E(\mathbf{w})$, the "better" the network state $\mathbf{w}$. Well-known examples are backpropagation [10], Hebbian learning [11], and Kohonen-type

 ©1993 The American Physical Society

learning [12].

Because of the random presentation of the training patterns, the learning procedure as defined in Eq. (1) is a stochastic process. The probability $P(\mathbf{w},t)$ that the network is in state $\mathbf{w}$ at time t obeys the master equation [2]

$$\frac{\partial P(\mathbf{w}',t)}{\partial t}=\int d^N w[T(\mathbf{w}'|\mathbf{w})P(\mathbf{w},t)-T(\mathbf{w}|\mathbf{w}')P(\mathbf{w}',t)] , \tag{2}$$

with transition probability

$$T(\mathbf{w}'|\mathbf{w})=\int d^n x\, \rho(\mathbf{w},\vec{x})\delta^N(\mathbf{w}'-\mathbf{w}-\eta\mathbf{f}(\mathbf{w},\vec{x})) .$$

It is impossible to solve this master equation in general. However, using standard arguments from the theory of unstable stochastic processes [13], it can be shown that after an initial time of order $1/\eta$ the probability distribution $P(\mathbf{w},t)$ obeys to a very good approximation [5]

$$P(\mathbf{w},t)=\sum_\alpha n_\alpha(t)p_\alpha(\mathbf{w})+P_{\rm rest}(\mathbf{w},t) .$$

$n_\alpha(t)$ is the occupation number at minimum α and $p_\alpha(\mathbf{w})$ a local normalized probability distribution with its average at $\mathbf{w}_\alpha^*$, the position of minimum α of the error potential $E(\mathbf{w})$. $P_{\rm rest}(\mathbf{w},t)$ stands for the probability to find a weight vector $\mathbf{w}$ outside the direct vicinity of the minima. For small learning parameters, its probability mass is negligible in comparison with the probability mass in the neighborhood of the minima.

Transitions between different minima are rare. The exchange of probability mass between the various minima is governed by

$$\frac{dn_\alpha(t)}{dt}=\sum_\beta\left[\frac{n_\beta(t)}{\tau_{\alpha\beta}}-\frac{n_\alpha(t)}{\tau_{\beta\alpha}}\right] , \tag{3}$$

where $\tau_{\alpha\beta}$ is the transition time from minimum β to α. From theory and simulations we deduce that these transition times are of the form [5]

$$\tau_{\alpha\beta}=\frac{1}{A_{\alpha\beta}\eta^{d_{\alpha\beta}}}\exp\left[\frac{\tilde{\eta}_{\alpha\beta}}{\eta}\right] . \tag{4}$$

$\tilde{\eta}_{\alpha\beta}$ is called the reference learning parameter for the transition from β to α. In [5] we have described a method to calculate, or at least estimate, this parameter. We do not know how to calculate $A_{\alpha\beta}$ and $d_{\alpha\beta}$, but these parameters become less and less important for smaller learning parameters. In the next section we will investigate whether an efficient cooling schedule really depends on these parameters.

III. A TWO-LEVEL SYSTEM

We consider a system with one global minimum $E_1\equiv E(\mathbf{w}_1^*)$ and one local minimum $E_2\equiv E(\mathbf{w}_2^*)$. We will assume that $\tilde{\eta}_{12}<\tilde{\eta}_{21}$, i.e., that for small learning parameters the transition from the local minimum to the global minimum is easier than vice versa. In Sec. VI we will argue that this is true in most practical situations. Furthermore, we will treat the system as a true two-level system, i.e., we define the average error $E(t)$ by

$$E(t)\equiv n_1(t)E_1+n_2(t)E_2 . \tag{5}$$

Combining this with $n_1(t)+n_2(t)=1$, the occupation numbers are uniquely determined once $E(t)$ is given:

$$n_1(t)=\frac{E_2-E(t)}{E_2-E_1} , \quad n_2(t)=\frac{E(t)-E_1}{E_2-E_1} . \tag{6}$$

Note that the possibility to express the occupation numbers in terms of the average error potential is particular for a two-level system: it fails for three or more levels. Systems with more minima will be studied in Sec. IV.

Using Eqs. (3), (5), and (6), we can write a differential equation for the average error potential (for notational convenience we will drop the explicit time dependency):

$$\frac{dE}{dt}=(E_2-E_1)\frac{dn_2}{dt}=-\left[\frac{E-E_1}{\tau_{12}}-\frac{E_2-E}{\tau_{21}}\right] . \tag{7}$$

The optimal cooling schedule is found by choosing η such that the term between large parentheses is as large as possible [14], i.e., such that

$$\frac{\tilde{\eta}_{12}+d_{12}\eta}{\tau_{12}}(E-E_1)=\frac{\tilde{\eta}_{21}+d_{21}\eta}{\tau_{21}}(E_2-E) ,$$

or, writing the average error E in terms of the learning parameter η,

$$E=\frac{\tau_{21}(\tilde{\eta}_{12}+d_{12}\eta)E_1+\tau_{12}(\tilde{\eta}_{21}+d_{21}\eta)E_2}{\tau_{21}(\tilde{\eta}_{12}+d_{12}\eta)+\tau_{12}(\tilde{\eta}_{21}+d_{21}\eta)} . \tag{8}$$

The fastest path specifies η as a function of E and vice versa. The time trajectory of the optimal η can be calculated from

$$\frac{d\eta}{dt}=\left[\frac{dE}{d\eta}\right]^{-1}\frac{dE}{dt} .$$

Using Eqs. (7) and (8), we obtain

$$\frac{d\eta}{dt}=-\eta^2\left[(\tilde{\eta}_{21}+d_{21}\eta)(\tilde{\eta}_{12}+d_{12}\eta)+\frac{(d_{21}\tilde{\eta}_{12}-d_{12}\tilde{\eta}_{21})\eta^2}{\tilde{\eta}_{21}-\tilde{\eta}_{12}+(d_{21}-d_{12})\eta}\right]^{-1}\times\left[\frac{\tilde{\eta}_{21}+d_{21}\eta}{\tau_{21}}+\frac{\tilde{\eta}_{12}+d_{12}\eta}{\tau_{12}}\right] . \tag{9}$$

It is not possible to solve this differential equation explicitly. For large t, we expect $\eta\to 0$. Keeping only the lowest orders in η and noting that in this limit $\tau_{12}(\eta)\ll\tau_{21}(\eta)$, we obtain

$$\frac{d\eta}{dt}=-\frac{\eta^2}{\tilde{\eta}_{21}\tau_{12}(\eta)} .$$

For large t the approximate solution of this differential equation is

$$\eta(t)=\frac{\tilde{\eta}_{12}}{\ln\{(\tilde{\eta}_{12}A_{12}/\tilde{\eta}_{21})[t/(\ln t)^{d_{12}}]\}}+\mathcal{O}\left[\frac{(\ln\ln t)^2}{(\ln t)^3}\right] . \tag{10}$$

Backsubstitution in Eq. (9) confirms that this is really a consistent approximation for $\eta(t)$. The lowest-order approximation of Eq. (10) yields

$$\eta(t)=\frac{\tilde{\eta}_{12}}{\ln t}+\mathcal{O}\left[\frac{\ln\ln t}{(\ln t)^2}\right] . \tag{11}$$

This constitutes our final cooling schedule. It does not depend on the parameters $A_{\alpha\beta}$ and $d_{\alpha\beta}$ in Eq. (4). We only have to compute the reference learning parameter $\tilde{\eta}_{12}$ for the transition from the local to the global minimum.

In a sense, the derived cooling schedule is indeed optimal. A "faster" cooling schedule, e.g., $\eta(t)=\tilde{\eta}_{12}/5\ln t$, cannot guarantee that a network starting at the local minimum will indeed reach the global minimum. We could say that the transition from the local to the global minimum is "closed." The optimal cooling schedule keeps this transition just "open." A "slower" cooling schedule, e.g., $\eta(t)=5\tilde{\eta}_{12}/\ln t$, gives also an open transition, but convergence might take much longer than with the optimal cooling schedule. By looking at the transition times we can easily check whether a particular transition is open or closed. If the transition time grows at most linearly with time t the transition is open, if it grows faster than linearly with time t the transition is closed. For the optimal cooling schedule (11) the transition time τ_{12} from the local to the global minimum grows linearly with time t.

IV. VARIOUS LOCAL MINIMA

We will try to find a cooling schedule in the case of $M-1$ local minima and one global minimum at $\mathbf{w}_1^*$, $M>2$. Generalization to more global minima is straightforward. We denote the stationary distribution of the master equation (2) for constant learning parameter η by $P_\eta(\mathbf{w},\infty)$. In the limit $\eta\to 0$, this stationary distribution concentrates at the (local) minima of the error potential [7], i.e.,

$$\lim_{\eta\to 0}P_\eta(\mathbf{w},\infty)=\sum_{\alpha=1}^{M}p_\alpha\delta^N(\mathbf{w}-\mathbf{w}_\alpha^*) . \tag{12}$$

Since the stationary distribution does not depend explicitly on the error potential, there is no guarantee that it will concentrate near the global minimum, i.e., that $p_\alpha=\delta_{\alpha 1}$. Nevertheless, in order to make some progress, we will postulate that

$$\lim_{\eta\to 0}P_\eta(\mathbf{w},\infty)=\delta^N(\mathbf{w}-\mathbf{w}_1^*) , \tag{13}$$

i.e., that for small learning parameters the stationary probability distribution will concentrate on the global minimum. In Sec. VI we will argue why this postulate is reasonable in most practical situations. However, if another minimum is more "attractive," e.g., minimum 2 if $p_\alpha=\delta_{\alpha 2}$ in Eq. (12), then a cooling schedule, at least one of the type we will derive, will drive all learning networks to this minimum.

Instead of trying to solve the master equation in weight space, we will study the dynamics of the occupation numbers at the various minima given in Eq. (3). If we define the transition matrix $\Gamma(\eta)$ by

$$\Gamma_{\alpha\beta}(\eta)=-\frac{1}{\tau_{\alpha\beta}}(\eta) \quad \text{for } \alpha\neq\beta ,$$

$$\Gamma_{\alpha\alpha}(\eta)=\sum_{\beta\,(\neq\alpha)}\frac{1}{\tau_{\beta\alpha}(\eta)} ,$$

then the dynamics of the occupation numbers for time-dependent $\eta(t)$ is written

$$\frac{d\vec{n}(t)}{dt}=-\Gamma(\eta(t))\vec{n}(t) . \tag{14}$$

Our goal is now to find a cooling schedule $\eta(t)$ such that the solution $\vec{n}(t)$ of this differential equation obeys

$$\lim_{t\to\infty}\vec{n}(t)=(1,0,\ldots,0,0)^T ,$$

i.e., such that in the end all the probability mass is concentrated at the global minimum.

We denote the left and right eigenvectors of $\Gamma(\eta)$ by $\vec{a}_i(\eta)$ and $\vec{b}_i(\eta)$, respectively. $\lambda_i(\eta)$ stands for the corresponding eigenvalue, $\kappa_i(\eta)$ for the real part of this eigenvalue, and $\Delta_i(t)$ for the projection of $\vec{n}(t)$ on the left eigenvector $\vec{a}_i(\eta)$:

$$\Delta_i(t)\equiv\vec{a}_i(\eta)\cdot\vec{n}(t) .$$

Now $\vec{n}(t)$ can be written

$$\vec{n}(t)=\sum_{i=0}^{M-1}\Delta_i(t)\vec{b}_i(\eta) .$$

Note that $1-\Gamma(\eta)$ is a stochastic matrix, i.e., all elements of $1-\Gamma(\eta)$ are non-negative and the elements in each row add up to 1 (see, e.g., [15] for some general properties of stochastic matrices). So, $\Gamma(\eta)$ has one zero eigenvalue with corresponding left eigenvector $\vec{a}_0(\eta)=(1,1,\ldots,1,1)$. All other eigenvalues have positive real parts. If we order the eigenvalues such that

$$0=\kappa_0(\eta)<\kappa_1(\eta)\leq\cdots\leq\kappa_{M-2}(\eta)\leq\kappa_{M-1}(\eta)\leq 2 ,$$

then $\Delta_1(t)$ gives the slowest convergence to the stationary solution $\vec{b}_0(\eta)$. In these terms, postulate (13) reads

$$\lim_{\eta\to 0}\vec{b}_0(\eta)=(1,0,\ldots,0,0)^T . \tag{15}$$

From Eq. (14) we derive the following differential equation for the projections:

$$\frac{d\Delta_i(t)}{dt}=-\lambda_i(\eta(t))\Delta_i(t)+R_i(t) , \tag{16}$$

with

$$R_i(t)\equiv\frac{d\eta(t)}{dt}\left[\frac{d\vec{a}_i(\eta(t))}{d\eta(t)}\right]\cdot\vec{n}(t) ,$$

an extra term due to the time dependency of the learning parameter. We are interested in the conditions under which the projection $\Delta_i(t)$ vanishes in the limit $t \to \infty$. In these considerations the term $R_i(t)$ can be neglected if the integral over $R_i(t)$ is bounded, i.e., if for some t_0

$$\int_{t_0}^{\infty} dt\,|R_i(t)| < \infty \ .$$

The proof is straightforward. Rewriting the integral over t in an integral over η and using $\|\vec{n}(t)\| \leq 1$, we obtain

$$\int_{t_0}^{\infty} dt\,|R_i(t)| \leq \int_{\eta(\infty)}^{\eta(t_0)} d\eta \left\| \frac{d\vec{a}_i(\eta)}{d\eta} \right\| < \infty \ ,$$

since the second integral is over a bounded interval.

The cooling schedule $\eta(t)$ has to guarantee that all projections $\Delta_i(t)$ vanish, except $\Delta_0(t)$, the projection on the eigenvector with zero eigenvalue. In that case the only remaining component is in the direction of $\vec{b}_0(\eta(t))$. This eigenvector must converge to $(1,0,\ldots,0,0)^T$ for $t \to \infty$. Comparison with Eq. (15) yields

$$\lim_{t\to\infty} \eta(t) = 0 \ ,$$

i.e., in the end the learning parameter should go to zero. The slowest convergence is determined by the eigenvalue $\lambda_1(\eta(t))$. From Eq. (16) we deduce the requirement

$$\int^{\infty} dt\, \kappa_1(\eta(t)) = \infty \ .$$

The optimal cooling schedule is found if this condition is just fulfilled, i.e., if

$$\kappa_1(\eta(t)) \propto \frac{1}{t} \quad \text{for } t \to \infty \ . \tag{17}$$

In the Appendix we derive

$$\kappa_1(\eta) \sim \exp\left[-\frac{\eta^*}{\eta}\right] \quad \text{for } \eta \to 0 \ ,$$

with

$$\eta^* = -\lim_{\eta\to 0} \eta \ln \left| \frac{\frac{\partial}{\partial\lambda}\det[\Gamma(\eta)-\lambda]|_{\lambda=0}}{\frac{1}{2}\frac{\partial^2}{\partial\lambda^2}\det[\Gamma(\eta)-\lambda]|_{\lambda=0}} \right| \ .$$

Comparing with Eq. (17), we conclude that the optimal cooling schedule is of the form

$$\eta(t) = \frac{\eta^*}{\ln t} \quad \text{for } t \to \infty \ . \tag{18}$$

This kind of "exponentially slow" cooling schedule is common ground in the theory of stochastic processes for global optimization [1,3]. Kushner [8] already showed that this schedule works for large enough η^*. Knowledge about the optimal η^* can be very useful since it prevents the cooling schedule from being slower than strictly necessary. In cooling schedules for simulated annealing the optimal η^* is called "the critical depth" [16]. It is the depth (suitably defined) of the deepest local minimum which is not a global minimum state [17]. In this context, the approach taken in [18,19] is most similar to ours: the critical depth is computed from the structure of a Markov chain, i.e., from the transition probabilities between different states.

In the Appendix we derive the following bounds for η^*,

$$\tilde{\eta}_{\min} \leq \eta^* \leq \tilde{\eta}_{\min} + (M-1)(\tilde{\eta}_{\max} - \tilde{\eta}_{\min}) \ ,$$

with $\tilde{\eta}_{\min}$ and $\tilde{\eta}_{\max}$ the smallest and the largest finite reference learning parameter, respectively. The lower bound can be explained from the considerations at the end of Sec. III. A choice $\eta^* < \tilde{\eta}_{\min}$ is definitely wrong since then *all* transition times grow faster than linearly with time t and thus all transitions are "closed." The eigenvalue $\lambda_1(\eta)$ that gives the slowest convergence to the stationary solution is related to the transition time for the most difficult transition indispensable to reach the global minimum from any arbitrary initial weight configuration. The optimal cooling schedule keeps this transition open but may close transitions that are not needed.

V. SIMULATIONS

To illustrate the performance of the derived cooling schedules we will use the same toy problems as in [5]. There it is shown that, if $\mathbf{x}$ is drawn according to a suitable conditional probability density function $\rho(\mathbf{w},\mathbf{x})$, the Grossberg learning rule [20]

$$\Delta\mathbf{w} = \eta(\mathbf{x}-\mathbf{w})$$

performs stochastic gradient descent on the error potential

$$E(\mathbf{w}) = \sum_{i=1}^{N} \frac{w_i^2}{4} - \frac{1}{2\beta}\ln[\beta w_i + \epsilon_i] \ .$$

In other words, the learning process is such that

$$\int d^N x\, \rho(\mathbf{w},\mathbf{x})(\mathbf{x}-\mathbf{w}) = -\nabla E(\mathbf{w}) \ .$$

β and ϵ are adjustable parameters. Roughly speaking β determines the steepness of the minima and ϵ the relative depth.

First, we will discuss simulations of Grossberg learning with just one weight. The error potential with $\beta = 1.5$ and $\epsilon = 0.05$, shown in Fig. 1(a), has one global and one local minimum. The reference learning parameters can be calculated using the procedure given in [5]. We obtain (throughout the rest of the paper we will give the numerical results in three significant digits)

$$\tilde{\eta}_{12} = 0.146 \ , \quad \tilde{\eta}_{21} = 0.327 \ .$$

The difference in $\tilde{\eta}_{12}$ and $\tilde{\eta}_{21}$ reflects the fact that transitions from left to right are easier than transitions from right to left. To make the connection with Sec. IV, the nonzero eigenvalue of the two-dimensional transition matrix $\Gamma(\eta)$ obeys

$$\lambda_1(\eta) \sim \exp\left[-\frac{\tilde{\eta}_{12}}{\eta}\right] \ , \quad \eta \to 0 \ .$$

In the derivation of our cooling schedules we have only

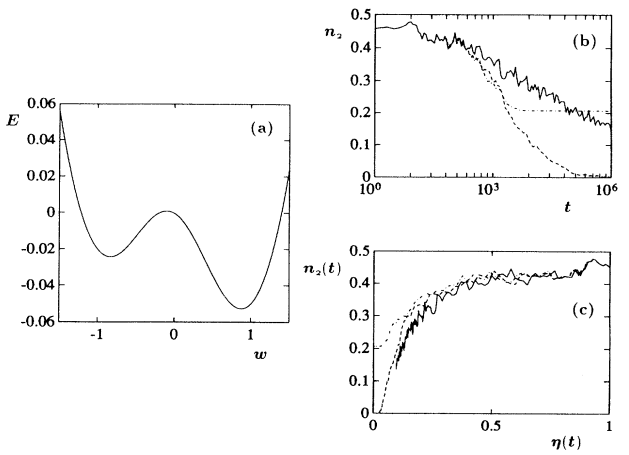

FIG. 1. (a) One-dimensional error potential $E(w)$ for $\beta=1.5$ and $\epsilon=0.05$. (b) Occupation number at the local minimum $n_2(t)$ as a function of time t. (c) Occupation number $n_2(t)$ vs the learning parameter $\eta(t)$. (b) and (c) for three different cooling schedules: $\eta^*=0.2$ (dashed line), $\eta^*=0.04$ (solid line), and $\eta^*=1$ (dash-dotted line).

studied the asymptotic behavior of the learning parameter. Any cooling schedule satisfying Eq. (18) for large times t is acceptable. In our simulations we will use cooling schedules of the form

$$\eta(t)=\frac{\eta^*}{\ln(\gamma\eta^* t+1)+\eta^*} . \tag{19}$$

This cooling schedule is such that

$$\eta(0)=1 , \quad \left.\frac{d\eta(t)}{dt}\right|_{t=0}=-\gamma .$$

γ sets the initial rate of change of the learning parameter. For large t the parameter γ becomes less and less important.

Simulations are done for three different cooling schedules: (1) near optimal: $\eta^*=0.2$, *dashed line;* (2) cooling too slowly: $\eta^*=1$, *solid line;* (3) cooling too abruptly: $\eta^*=0.04$, *dash-dotted line.* The parameter γ in Eq. (19) is kept constant at 0.01 and all 1000 independently learning networks are initialized with equal probability between -1 and 1. In this way the initial dynamics of the learning process is roughly the same for the three cooling schedules. The relative success of the cooling schedules is purely determined by their different large time behavior.

$n_2(t)$, the occupation number of networks in the vicinity of the local minimum, is plotted as a function of time t in Fig. 1(b) and versus the learning parameter $\eta(t)$ in Fig. 1(c). If the learning parameter is cooled too abruptly ($\eta^*=0.04$, dash-dotted line), many learning systems, in this case about 20%, end up not at the global minimum but at the local minimum. If the learning parameter is cooled too slowly ($\eta^*=1$, solid line), all learning systems may still reach the global minimum (we stopped after 10^6 learning steps) but this takes a far longer time than for the (almost) optimal cooling schedule ($\eta^*=0.2$, dashed line). After 10^6 learning steps with the (almost) optimal cooling schedule only 0.1% of the networks is still at the local minimum and with the slow cooling schedule about 15%. The simulations stress the importance of having a reasonable estimate for the reference learning parameter in order to derive an acceptable cooling schedule.

How to find a cooling schedule in the case of more minima is illustrated by simulating Grossberg learning performing stochastic gradient descent on the two-dimensional error potential shown in Fig. 2(a). With parameters $\beta=2.5$, $\epsilon_1=0.4$, and $\epsilon_2=0.2$, this error potential has four minima. Following the procedure explained in [5], we obtain the matrix $\tilde{\eta}$ with reference learning parameters

$$\tilde{\eta}=\begin{bmatrix} & 0.944 & 0.543 & \infty \\ 1.97 & & \infty & 0.543 \\ 2.58 & \infty & & 0.944 \\ \infty & 2.58 & 1.97 & \end{bmatrix} .$$

The possible transitions are drawn schematically in Fig. 3(a). The reference learning parameters $\tilde{\eta}_{14}$, $\tilde{\eta}_{41}$, $\tilde{\eta}_{23}$, and $\tilde{\eta}_{32}$ are infinite since the transition times for a direct transition over the barrier in the middle grow faster than exponentially with the reciprocal value of the learning parameter (see [5] for further explanation). Straightforward calculation of the eigenvalue $\lambda_1(\eta)$ yields

$$\lambda_1(\eta)\sim\exp\left[-\frac{0.944}{\eta}\right] \quad \text{for } \eta\to 0 ,$$

so, $\eta^*=0.944$. This parameter η^* is larger than the reference learning parameters corresponding to transitions going from a higher to a lower minimum. On the other hand, it is smaller than the reference learning parameters corresponding to transitions from a lower to a

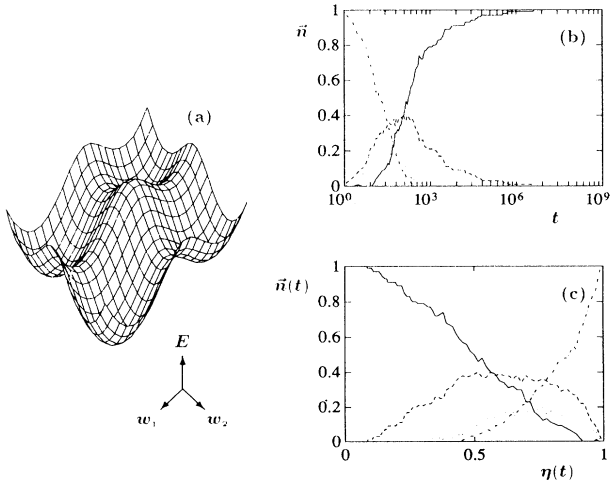

FIG. 2. (a) Two-dimensional error potential $E(w_1,w_2)$ for $\beta=2.5$, $\epsilon_1=0.4$, and $\epsilon_2=0.2$. (b) and (c) Occupation numbers $n_1(t)$ (solid line), $n_2(t)$ (dashed line), $n_3(t)$ (dotted line), and $n_4(t)$ (dash-dotted line) as a function of time t and vs the learning parameter $\eta(t)$, respectively.

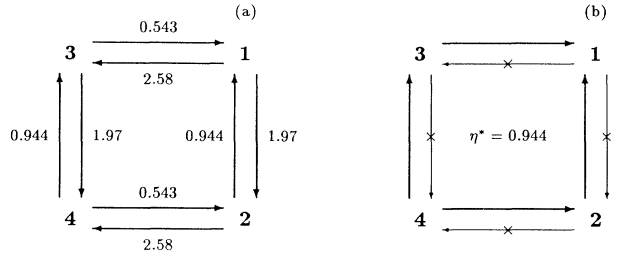

FIG. 3: (a) The reference learning parameters for the error potential shown in Fig. 2(a). (b) Transitions with $\tilde{\eta}_{\alpha\beta} > \eta^* = 0.944$ are "closed," transitions with $\tilde{\eta}_{\alpha\beta} \leq \eta^*$ are still "open."

higher minimum, since it is not necessary to go from a lower to a higher minimum on the way to the global minimum. The "open" and "closed" transitions are depicted in Fig. 3(b).

The results from learning with 100 networks, all starting at the highest local minimum, and a cooling schedule of the form (19), with parameters $\eta^* = 1$ and $\gamma = 0.01$, are given in Figs. 2(b) and 2(c) where the occupation numbers $n_1(t)$ (solid line), $n_2(t)$ (dashed line), $n_3(t)$ (dotted line), and $n_4(t)$ (dash-dotted line) are plotted a function of time t and versus the learning parameter $\eta(t)$, respectively. At the end, all networks have arrived at the global minimum.

VI. DISCUSSION

We have derived cooling schedules for learning in neural networks. The optimal cooling schedule for global optimization of on-line learning is of the form

$$\eta(t) = \frac{\eta^*}{\ln t} \quad \text{for large } t \ .$$

η^* can be calculated from the reference learning parameters for transitions between different minima. In some simple cases we were able to calculate η^* and found good agreement with simulation results.

Some comments should be made about the practical use of the theory presented in this paper.

(i) The derived cooling schedule is "exponentially slow," i.e., it takes an exponentially long time before one can be sure that the learning network has found the optimal solution. This is a fundamental problem in global optimization and is not typical for learning processes. For low-dimensional problems, a combination of cooling schedules and other techniques, e.g., multistart algorithms, might improve the speed of convergence. However, for large networks with many adaptable weights it will be unlikely to improve upon this exponentially slow cooling (see [3] for similar arguments regarding Langevin algorithms compared with other optimization techniques).

(ii) The cooling schedule will drive the networks to the most "attractive" minima. The question is whether these most attractive minima will coincide with the global minima. Let us compare stochastic learning processes with Metropolis and diffusion-type algorithms [1,3]. For both the Metropolis and the diffusion-type algorithms the stationary distribution is a Gibbs distribution which makes the global minima always the most attractive minima. The important difference between these stochastic processes and stochastic learning processes of the form (1) is that for the former the noise is the same at each minimum, whereas for the latter the noise at each minimum in general will be different [2]. Usually we will have that the higher the error potential, the more there is to learn, the larger the fluctuations in the learning rule, so the higher the noise level. Roughly speaking, the reference learning parameter for a transition from minimum α to β is proportional to the height of the barrier between α and β and inversely proportional to the local fluctuations at α. These arguments strongly suggest that the "colored noise" coming from the random presentation of patterns in on-line learning processes *helps* to find the global minimum in stochastic learning processes. Therefore violations of the postulate (13) will be rare.

(iii) Throughout this paper we assumed that we knew the reference learning parameters. To calculate these reference learning parameters, one needs detailed information about the environment. Usually, this information is not available. And if it is available it will be easier to compute the global minimum than all reference learning parameters. Therefore we do not suggest that for practical applications one should try to calculate these reference learning parameters. A solution of this problem might be a prelearning phase, during which an estimate of η^* is obtained by sampling the error surface. This is analogous to the estimation of the initial temperature for simulated annealing cooling schedules (see, e.g., [21]). In this paper we merely tried to show that there exists such a parameter η^* leading to an optimal cooling schedule and to give an idea of the factors that determine this parameter. This knowledge is meant to provide a theoretical basis for the design of practical algorithms that lead to global optimization of learning in neural networks.

ACKNOWLEDGMENTS

This work was partly supported by the Dutch Foundation for Neural Networks and the Canon Foundation in Europe.

APPENDIX

In this Appendix we will try to find an expression for $\kappa_1(\eta)$, i.e., for the smallest nonzero eigenvalue of the matrix $\Gamma(\eta)$. Let us consider the characteristic equation of the matrix $\Gamma(\eta)$:

$$0 = \det[\Gamma(\eta) - \lambda] = \sum_{n=0}^{M} c_{M-n}(\eta)(-\lambda)^n = \prod_{i=0}^{M-1} [\lambda_i(\eta) - \lambda] \ . \tag{A1}$$

Typically, $c_n(\eta)$ is the sum of a product over n transition probabilities, so schematically

$$c_n(\eta)\equiv\frac{(-)^{M-n}}{(M-n)!}\frac{\partial^{M-n}}{\partial\lambda^{M-n}}\det[\Gamma(\eta)-\lambda]\Big|_{\lambda=0}$$
$$=\sum_{\text{products}}\prod_{n\text{ terms}}\frac{1}{\tau_{\alpha\beta}(\eta)}\ . \tag{A2}$$

In terms of the eigenvalues λ_i the coefficient c_n reads

$$c_n=\sum_{\{i_1,\ldots,i_n\}}\lambda_{i_1}\cdots\lambda_{i_n}\ ,$$

where the sum is over all possible combinations $\{i_1,\ldots,i_n\}$ containing n distinct elements of the set $\{1,\ldots,M\}$. Since all eigenvalues are positive, we have $c_n\geq 0$. By simply writing out, we deduce

$$c_n^2-c_{n-1}c_{n+1}=\sum_{\{i_1,\ldots,i_{n-1}\}}\sum_{\{j_1,\ldots,j_n\}}\lambda_{i_1}\cdots\lambda_{i_{n-1}}\lambda_{j_1}\cdots\lambda_{j_n}\left[\sum_{k\notin\{i_1,\ldots,i_{n-1}\}}\lambda_k-\sum_{l\notin\{j_1,\ldots,j_n\}}\lambda_l\right]$$
$$\geq 0\ , \tag{A3}$$

since there are more constraints on the sum over l than on the sum over k. The inequality (A3) leads to the ordering

$$0=\frac{c_M}{c_{M-1}}<\frac{c_{M-1}}{c_{M-2}}\leq\cdots\leq\frac{c_1}{c_0}\ . \tag{A4}$$

In the limit $\eta\to 0$ the transition times given in Eq. (4) are dominated by the reference learning parameters $\tilde{\eta}_{\alpha\beta}$. Just as in Sec. III, we can neglect the influence of the parameters $A_{\alpha\beta}$ and $d_{\alpha\beta}$ in our search for a "lowest-order" cooling schedule of the form (11). Furthermore, in Eq. (A2), only the largest term in the sum will survive for small learning parameters η. So, we can always find a positive parameter $\tilde{\eta}_n$ such that

$$\frac{c_{n+1}(\eta)}{c_n(\eta)}\sim\exp\left[-\frac{\tilde{\eta}_n}{\eta}\right]\quad\text{for }\eta\to 0\ . \tag{A5}$$

Let us substitute the guess

$$\lambda=\frac{c_{i+1}(\eta)}{c_i(\eta)} \tag{A6}$$

in the characteristic equation (A1). Making use of the ordering (A4), we note that the $(M-i)$th term and the $(M-i-1)$th term are the largest terms in the sum. Since these terms exactly cancel, we conclude that the guess (A6) indeed yields (up to leading order) all eigenvalues of the matrix $\Gamma(\eta)$. Combining Eqs. (A2) and (A4)–(A6), we obtain the smallest nonzero eigenvalue

$$\lambda_1(\eta)\sim\exp\left[-\frac{\eta^*}{\eta}\right]\quad\text{for }\eta\to 0\ ,$$

with

$$\eta^*=-\lim_{\eta\to 0}\eta\ln\left|\frac{\frac{\partial}{\partial\lambda}\det[\Gamma(\eta)-\lambda]|_{\lambda=0}}{\frac{1}{2}\frac{\partial^2}{\partial\lambda^2}\det[\Gamma(\eta)-\lambda]|_{\lambda=0}}\right|\ .$$

A lower bound for η^* follows from

$$\lambda_1(\eta)\leq\frac{1}{M-1}\sum_{n=0}^{M-1}\lambda_n(\eta)$$
$$=\frac{1}{M-1}\mathrm{Tr}\Gamma(\eta)=\frac{1}{M-1}\sum_{\alpha}\sum_{\beta(\neq\alpha)}\frac{1}{\tau_{\alpha\beta}}\ .$$

In the limit $\eta\to 0$ only the largest transition probabilities, i.e., the smallest transition times, survive and thus

$$\eta^*\geq\tilde{\eta}_{\min}\ ,$$

with

$$\tilde{\eta}_{\min}\equiv\min_{\{\alpha,\beta\}}\{\tilde{\eta}_{\alpha\beta}\}\ .$$

To find a lower bound for $\lambda_1(\eta)$, we take the smallest possible $c_{M-1}(\eta)$ and the largest possible $c_{M-2}(\eta)$. From Eq. (A2) we obtain

$$\lambda_1(\eta)\geq\frac{\exp[-(M-1)\tilde{\eta}_{\max}/\eta]}{\exp[-(M-2)\tilde{\eta}_{\min}/\eta]}\quad\text{for }\eta\to 0\ ,$$

with $\tilde{\eta}_{\max}$ the largest finite reference learning parameter, i.e.,

$$\tilde{\eta}_{\max}\equiv\max_{\{\alpha,\beta|\tilde{\eta}_{\alpha\beta}<\infty\}}\tilde{\eta}_{\alpha\beta}\ .$$

An upper bound for η^* is thus

$$\eta^*\leq\tilde{\eta}_{\min}+(M-1)(\tilde{\eta}_{\max}-\tilde{\eta}_{\min})\ .$$

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

[2] T. Heskes and B. Kappen, Phys. Rev. A **44**, 2718 (1991).

[3] B. Gidas, in *Disordered Systems and Biological Organization,* edited by E. Bienenstock, F. Fogelman Soulié, and G. Weisbuch (Springer-Verlag, Berlin, 1986), pp. 321–326.

[4] H. Kushner, SIAM J. Appl. Math. **44**, 160 (1984).

[5] T. Heskes, E. Slijpen, and B. Kappen, Phys. Rev. A **46**, 5221 (1992).

[6] L. Ljung, IEEE Trans. Autom. Control, **AC-22**, 551 (1977).

[7] C. Kuan and K. Hornik, IEEE Trans. Neural Networks **2**, 484 (1991).

[8] H. Kushner, SIAM J. Appl. Math. **47**, 169 (1987).

[9] H. Ritter and K. Schulten, Biol. Cybernetics **60**, 59 (1988).
[10] D. Rumelhart, G. Hinton, and R. Williams, Nature (London) **323**, 533 (1986).
[11] D. Hebb, *The Organization of Behavior* (Wiley, New York, 1949).
[12] T. Kohonen, Biol. Cybernetics **43**, 59 (1982).
[13] N. van Kampen, *Stochastic Processes in Physics and Chemistry* (North-Holland, Amsterdam, 1981).
[14] S. Shinomoto and Y. Kabashima, J. Phys. A **24**, L141 (1991).
[15] A. Berman and R. Plemmons, *Nonnegative Matrices in the Mathematical Sciences* (Academic, New York, 1979).
[16] O. Catoni, Ann. Probability **20**, 1109 (1992).
[17] B. Hajek, Math. Operations Res. **13**, 311 (1988).
[18] J. Tsitsiklis, *Stochastic Differential Systems, Stochastic Control Theory and Applications,* edited by W. Fleming and P. Lions (Springer-Verlag, New York, 1988), p. 583–599.
[19] J. Tsitsiklis, Math. Operations Res. **14**, 70 (1989).
[20] S. Grossberg, J. Stat. Phys. **48**, 105 (1969).
[21] E. Aarts and P. van Laarhoven, in *Heidelberg Colloquium on Glassy Dynamics,* edited by J. van Hemmen and I. Morgenstern (Springer-Verlag, Berlin, 1987), pp. 288–307.