Learning in neural networks with local minima

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 5 May 1992)

An attempt is made to study learning in neural networks with local minima. For small learning parameters η, the transition time from one mimimum to another is asymptotically given by $\exp(\tilde{\eta}/\eta)$, with $\tilde{\eta}$, a constant independent of η, called the reference learning parameter. A general scheme to calculate the reference learning parameter is presented. This scheme is valid for a large class of learning rules.

PACS number(s): 87.10.+e

I. INTRODUCTION

A. Context

In the past decade many learning rules for neural networks have been invented or reinvented. These learning rules, in combination with a suitable architecture, make neural networks very useful for industrial applications. Nevertheless, in many cases a good theoretical understanding of why these networks are successful or how their performance can be improved is absent. Theoretical attempts in this direction can be roughly divided in two main streams: studies on network architecture (e.g., the number of hidden units in a multilayered perceptron) and studies on the dynamics of learning processes (e.g., the learning parameter as a function of time). This paper fits in the second category.

Basically, learning is the way a network builds an internal representation of its environment. This environment consists of a set of training patterns. The functionality of the network depends on the learning rule and architecture. Examples are multilayered perceptrons with backpropagation [1] for classification, principal component analyzing networks [2] for feature extraction, Kohonen-type networks [3] for the creation of topological maps, Hebbian learning [4] for associative memory, and so on.

The learning parameter plays a similar role in all these learning rules. It sets the typical magnitude of the changes in the network stage at each presentation of a training pattern. The effect of the learning parameter on the network performance has been studied in some specific cases [5,6], but also from a more general point of view [7,8]. This general formalism can be extended to study learning processes in a changing environment. The results obtained in this study can be used to derive an algorithm for on-line learning-parameter adjustment [9]. However, as we will explain below, all these efforts fail to give us an insight on the *global* network performance.

In some important cases, of which backpropagation is the most appealing example, the learning rule is derived from an error criterion. The learning rule is chosen such that, on the average, it performs gradient descent on this error potential. This error potential can have many minima. *A priori,* there is no guarantee that the learning process will lead the network to the global minimum. Even worse, the learning rule has the tendency to drive the network into the nearest local minimum. Only because of the stochasticity, introduced by the random selection of the training patterns, there is a possibility to escape from these local minima. What is the effect of the learning parameter in this case? Common sense tells us that a larger learning parameter leads to larger fluctuations and thus to a larger escape probability. In this paper, we will try to refine and quantify these statements.

Learning processes can be described by a master equation. In solving this master equation, one could try to borrow from the general theory on stochastic processes. To a certain extent, we will follow this strategy. But, even in this field, no general (expansion) method exists to solve the master equation in unstable systems [10]. For small learning parameters, a straightforward Fokker-Planck approach seems natural [5]. However, although this approach may be appropriate in the case of one minimum, it is not appropriate in the case of several local minima. Our approach is based on two hypotheses which are supported by experimental evidence and common sense. These hypotheses give us the opportunity to calculate asymptotic expressions for transition times and stationary probabilities.

B. Definitions

The state of a neural network is specified by an N-dimensional vector $\mathbf{w}=(w_1,\ldots,w_N)^T$, called the weight vector. This vector contains the strengths of all synapses and thresholds in the network. The network is trained with examples from an environment. This environment is defined as a set of training patterns $\vec{x}$ to be taken from a subset $\Omega\subseteq\mathbb{R}^n$. The environment of the network is fixed. In other words, the probability density that the network "sees" a training pattern $\vec{x}$ is time independent. In general, this probability density $\rho(\mathbf{w},\vec{x})$ may be conditional, i.e., depend explicitly on the current network state $\mathbf{w}$. In the examples of Secs. II and IV, the learning network does indeed affect the (probability distribution of the) environment on which it is trained. Since this special aspect of the learning procedure has no influence on the methods we use, we will not emphasize it.

At distinct points in time a training pattern $\vec{x}$ is drawn

at random from the environment Ω according to the probability $\rho(\mathbf{w},\vec{x})$. This training pattern is presented to the network and a learning step takes place. The network changes its weight vector $\mathbf{w}$ to $\mathbf{w}'=\mathbf{w}+\Delta\mathbf{w}$, obeying

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

where $\mathbf{f}(\mathbf{w},\vec{x})$, the so-called "stochastic force," is an arbitrary function $\mathbf{f}$: $\mathbb{R}^N\times\mathbb{R}^n\rightarrow\mathbb{R}^N$. Equation (1) states that the new network state $\mathbf{w}'$ after the learning step is a function of the state $\mathbf{w}$ before this learning step and the randomly drawn input vector $\vec{x}$. Depending on the particular choice of the stochastic force $\mathbf{f}(\mathbf{w},\vec{x})$ learning processes of neural networks with quite different functionalities can be described.

We will restrict ourselves in this paper to a special kind of learning rules, namely those learning rules for which a twice continuous differentiable error potential $E(\mathbf{w})$ can be defined. Such an error potential exists if and only if the drift term $\mathbf{f}(\mathbf{w})$, which is just the stochastic force averaged over the set of training patterns Ω, i.e.,

$$\mathbf{f}(\mathbf{w})\equiv\int d^n x\,\rho(\mathbf{w},\vec{x})\mathbf{f}(\mathbf{w},\vec{x}) ,$$

is continuous differentiable and obeys

$$\frac{\partial f_i(\mathbf{w})}{\partial w_j}=\frac{\partial f_j(\mathbf{w})}{\partial w_i} \quad \forall\ i,j .$$

Up to an additive constant, the error potential is then unambiguously defined by

$$f_i(\mathbf{w})=-\frac{\partial E(\mathbf{w})}{\partial w_i} \quad \forall\ i . \tag{2}$$

This error potential yields a global measure of network performance: the lower $E(\mathbf{w})$, the "better" the network state $\mathbf{w}$. The approach we will follow in this paper can also be applied if there exists no error potential, so it is valid for any continuous differentiable drift $\mathbf{f}(\mathbf{w})$. The results are totally equivalent, but it is difficult to specify what makes a particular state $\mathbf{w}$ better than another.

Backpropagation [1] is a well-known example of a learning rule with an error potential. To clarify our definitions, let us consider a multilayered feedforward network with one output, $n-1$ input units, and N synapses and thresholds. In our formalism a training pattern $\vec{x}$ is a combination of the network input, say, $x_1,\ldots,x_{n-1}$, and the desired output x_n. The error potential is the quadratic distance between the network output $y(\mathbf{w},x_1,\ldots,x_{n-1})$ and the desired output x_n, averaged over the total set of training patterns

$$E(\mathbf{w})=\tfrac{1}{2}\int d^n x\,\rho(\vec{x})[y(\mathbf{w},x_1,\ldots,x_{n-1})-x_n]^2 .$$

It is straightforward to prove that this error potential is the error potential of the backprop learning rule

$$\Delta w_i=\eta[y(\mathbf{w},x_1,\ldots,x_{n-1})-x_n]\times\frac{\partial y(\mathbf{w},x_1,\ldots,x_{n-1})}{\partial w_i} .$$

Other examples of learning rules with an error potential are Hebbian learning [4,11] for attractor neural networks and some types of Kohonen-learning [3,5] for topological maps.

C. State of the art

In a previous paper [7] we studied the behavior of learning rules obeying Eq. (1) for small constant learning parameters η. If the points of time of the learning steps follow a Poisson process with on the average one learning step per unit time, the evolution of the learning process as defined above is fully determined by the continuous-time master equation [12]

$$\frac{\partial P(\mathbf{w}',t)}{\partial t}=\int d^N w[T(\mathbf{w}'|\mathbf{w})P(\mathbf{w},t)-T(\mathbf{w}|\mathbf{w}')P(\mathbf{w}',t)] . \tag{3}$$

$P(\mathbf{w},t)$ denotes the probability density function of the weight vector $\mathbf{w}$ at time t. The transition probability $T(\mathbf{w}'|\mathbf{w})$ obeys

$$T(\mathbf{w}'|\mathbf{w})=\int d^n x\,\rho(\mathbf{w},\vec{x})\delta^N(\mathbf{w}'-\mathbf{w}-\eta\mathbf{f}(\mathbf{w},\vec{x})) . \tag{4}$$

This can be read as the probability measure of the set of training patterns $\vec{x}$ such that the learning rule (1) turns the old network state $\mathbf{w}$ exactly into the new one $\mathbf{w}'$.

From the master equation (3), evolution equations for the average network state and the fluctuations around this average can be derived. It is possible to prove that for large times and small learning parameters the network has a very high probability to be in the neighborhood of an attractive fixed point $\mathbf{w}^*$ of the differential equation (see also [8,13,14])

$$\frac{d\mathbf{w}(t)}{dt}=\mathbf{f}(\mathbf{w}(t)) .$$

In an error potential $E(\mathbf{w})$ exists, these fixed points $\mathbf{w}^*$ are just the minima of this error potential. In these terms, the statement above only tells us that the network will get stuck in the neighborhood of one of the minima. It cannot predict at which one nor can it give us information about the time it takes to go from one minimum to another. So far, nothing has been said about the effect of the learning parameter on the *global* performance of the network.

For stochastic processes such as simulated annealing and diffusion processes, the stationary probability distribution can be derived explicitly. In general, this is not possible for master equations of the form (3). This complicates the study of the global performance of learning rules. To make some progress, we will make some hypotheses which are motivated by simulations.

D. Outline of the paper

In Sec. II we will discuss a simple one-dimensional network with one global and one local minimum. Looking at simulations with many identical copies of this network, we will arrive at two hypotheses. The first one is worked out in Sec. III, where we calculate the shape of the probability density function in the neighborhood of local minima. These shapes will be used in Sec. IV to calculate the

transition time from one minimum to another. The derivation for one dimension is extended to general higher-dimensional learning rules. The final calculation scheme can be applied to any learning rule of the form (1). As an example, a two-dimensional network is treated in detail. In Sec. V the main results are summarized, the hypotheses are reviewed, and the applicability of our approach to practical situations is discussed.

II. THE HYPOTHESES

In this section we will introduce two hypotheses which form the starting points for our theoretical derivation. We will visualize them by means of a simple example of a one-dimensional "neural network" with a local and a global minimum.

A. An example

The network has one weight w, which is adapted according to the Grossberg learning rule [15]

$$\Delta w=\eta(x-w)\ ,$$

where η is the learning parameter and x is the input of the network, drawn at random from the environment according to a conditional probability density function $\rho(w,x)$. In the usual case, where $\rho(w,x)=\rho(x)$, the error potential of Grossberg learning is always quadratic with just one minimum at $w^*=\int dx\,\rho(x)x$. In our example the probability to draw an input x does depend on the current network state. The network senses the real environment, denoted by $\rho_0(x)$, through a Gaussian filter of width σ,

$$\rho(w,x)=\frac{1}{Z(w)}\rho_0(x)e^{-(x-w)^2/2\sigma^2}\ .$$

That is, the probability to draw an input x within a distance σ of the current network state w is enlarged, whereas an input example further away is less probable. $Z(w)$ is a normalization constant such that $\int dx\,\rho(w,x)=1\ \forall\ w$. For the "real" input distribution $\rho_0(x)$, we take a sum of two Gaussian functions with standard deviation χ and mean x_0 and $-x_0$,

$$\rho_0(x)=\frac{1}{(2\pi\chi^2)^{1/2}}\sum_{\pm}\frac{1\pm a}{2}e^{-(x\mp x_0)^2/2\chi^2}\ ,$$

with $0\le a\le 1$, an asymmetry parameter. These probability distributions are sketched in Figs. 1(a) and 1(b) for $\sigma^2=\chi^2=\frac{1}{3}$, $\operatorname{arctanh}(a)=0.05$ and $x_0=1$. The solid line in Fig. 1(a) shows the distribution $\rho(-0.4,x)$, the one in Fig. 1(b) $\rho(0.4,x)$. The dashed lines in these figures give $\rho_0(x)$. It can be seen that the real input distribution $\rho_0(x)$ is strongly deformed by the Gaussian window of the network.

It is straightforward to show that the error potential, defined in Eq. (2), has the form of the well-known Ising potential in statistical physics,

$$E(w)=\frac{\sigma^2}{2(\sigma^2+\chi^2)}w^2-\sigma^2\ln\left[\cosh\left[\frac{wx_0}{\sigma^2+\chi^2}+\epsilon\right]\right]\ . \quad (5)$$

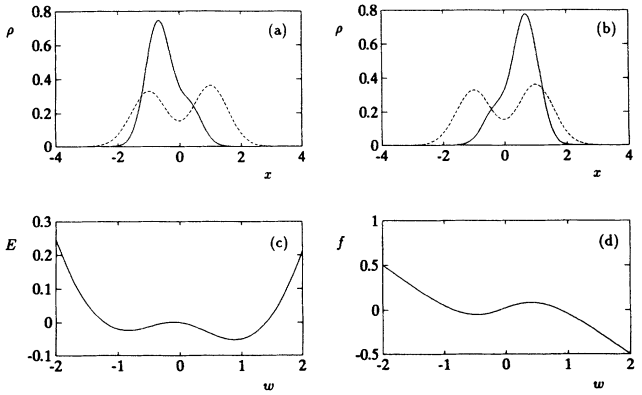

FIG. 1. Probability density $\rho(w,x)$, error potential $E(w)$, and drift $f(w)$ for $\sigma^2=\chi^2=\frac{1}{3}$, $\epsilon=0.05$, and $x_0=1$. The dashed lines show $\rho_0(x)$. (a) $\rho(-0.4,x)$. (b) $\rho(0.4,x)$. (c) $E(w)$. (d) $f(w)$.

The asymmetry introduced by $a\neq 0$ corresponds to a magnetic field of strength $\epsilon\equiv\operatorname{arctanh}(a)$. The ratio $\beta\equiv x_0^2/(\sigma^2+\chi^2)$ plays the role of the inverse temperature. In Fig. 1(c) the error potential $E(w)$ is plotted for $\beta=1.5$, $\sigma=\chi$, $\epsilon=0.05$, and $x_0=1$. The drift $f(w)$ is shown in Fig. 1(d).

In the example of Fig. 1, the drift term has three zeros and thus the error potential has one local minimum, one local maximum and one global minimum. In general, the number of zeros of $f(w)$ depends on the variables ϵ and β. If there is too much asymmetry, i.e., if ϵ is too large, there is just one minimum. The critical $\epsilon^*(\beta)$ is given by

$$\epsilon^*(\beta)=\begin{cases}0 & \text{if } \beta\le 1\\ \sqrt{\beta(\beta-1)}-\operatorname{arccosh}(\sqrt{\beta}) & \text{if } \beta>1\ .\end{cases}$$

We will always work with two minima, a local minimum at the left and a global minimum at the right, so with $0<\epsilon<\epsilon^*(\beta)$.

B. First hypothesis

In our study of the global behavior of the learning process, we will have to make a few assumptions. In order to make their introduction plausible, we will first look at a simulation of the learning process, presented in Fig. 2. The learning process is fully determined by the master equation (3), which gives the evolution of the probability density $P(w,t)$. A histogram of the weights of many (10 000) independently operating networks yields an estimate of this probability density.

Starting with random weights, uniformly distributed between -1 and 1 [2(a) $t=1$], the probability distribution evolves quickly [2(b) and 2(c) $t=10$ and $t=100$] towards a metastable situation with two peaks [2(d) $t=1000$]. These two peaks are called mesostates. Almost all "probability mass" is concentrated in these mesostates. The mesostates are quasistationary, since there is always a small but finite probability that a large fluctuation occurs, taking a network across the maximum of the error potential. This leads to a net flow of probability mass from the

local to the global minimum [2(e) $t=10\,000$]. This flow does not seem to effect the shape of the mesostates. In the stationary situation, almost all networks are found in the neighborhood of the global minimum [2(f) and 2(g) $t=100\,000$ and $t=1\,000\,000$].

Regions in the neighborhood of local minima are called attraction regions. In these regions $f'(w)<0$. The region between two attraction regions is called a transition region. We expand the probability $P(w,t)$ by writing

$$P(w,t)=P_{\text{left}}(w,t)+P_{\text{trans}}(w,t)+P_{\text{right}}(w,t) .$$

Here $P_{\text{left}}(w,t)$ and $P_{\text{right}}(w,t)$ refer to the left and right mesostates; these probabilities are zero outside the left and right attraction regions, respectively. $P_{\text{trans}}(w,t)$ is the probability distribution in the transition region. The typical time involved in the interaction between the two mesostates, i.e., the relaxation time to the stationary situation, is much larger than the time needed to converge to the metastable situation, denoted by τ_{meso}. In our study of the long-time behavior, it is therefore quite plausible to make the assumption that the mesostates have attained a unique stationary shape, but that the relevant weights have not reached their stationary value [16]. In other words,

$$\begin{aligned} P_{\text{left}}(w,t)&=n_{\text{left}}(t)p_{\text{left}}(w) , \\ P_{\text{right}}(w,t)&=n_{\text{right}}(t)p_{\text{right}}(w) . \end{aligned} \tag{6}$$

The time-independent distributions $p_{\text{left}}(w)$ and $p_{\text{right}}(w)$ are normalized, such that the factors $n_{\text{left}}(t)$ and $n_{\text{right}}(t)$ can be viewed as occupation numbers. Equation (6) constitutes our first hypothesis. It is frequently used in the theory of stochastic processes.

C. Second hypothesis

We are interested in the occupation numbers n_{left} and n_{right} as a function of time. As can be seen from Fig. 2, after some initial time, the probability mass in the transition region is negligible in comparison with the probability mass in the attraction regions. Hence the interaction between the two mesostates can be written in the form

$$\begin{aligned} \frac{d}{dt}n_{\text{left}}(t)&=-\frac{d}{dt}n_{\text{right}}(t) \\ &=-\frac{1}{\tau_{\text{r-l}}}n_{\text{left}}(t)+\frac{1}{\tau_{\text{l-r}}}n_{\text{right}}(t) , \end{aligned} \tag{7}$$

where $1/\tau_{\text{r-l}}$ is the probability per unit time for a network in the left attraction to fluctuate across the maximum of the error potential into the right attraction region. $\tau_{\text{r-l}}$ is called the transition time from the left to the right attraction region. The solution of this set of linear differential equations is

$$\begin{aligned} n_{\text{left}}(t)&=1-n_{\text{right}}(t) \\ &=\frac{\tau_{\text{r-l}}}{\tau_{\text{r-l}}+\tau_{\text{l-r}}}+\left[n_{\text{left}}(0)-\frac{\tau_{\text{r-l}}}{\tau_{\text{r-l}}+\tau_{\text{l-r}}}\right] \\ &\quad\times\exp\left[-\frac{\tau_{\text{r-l}}+\tau_{\text{l-r}}}{\tau_{\text{r-l}}\tau_{\text{l-r}}}t\right] . \end{aligned} \tag{8}$$

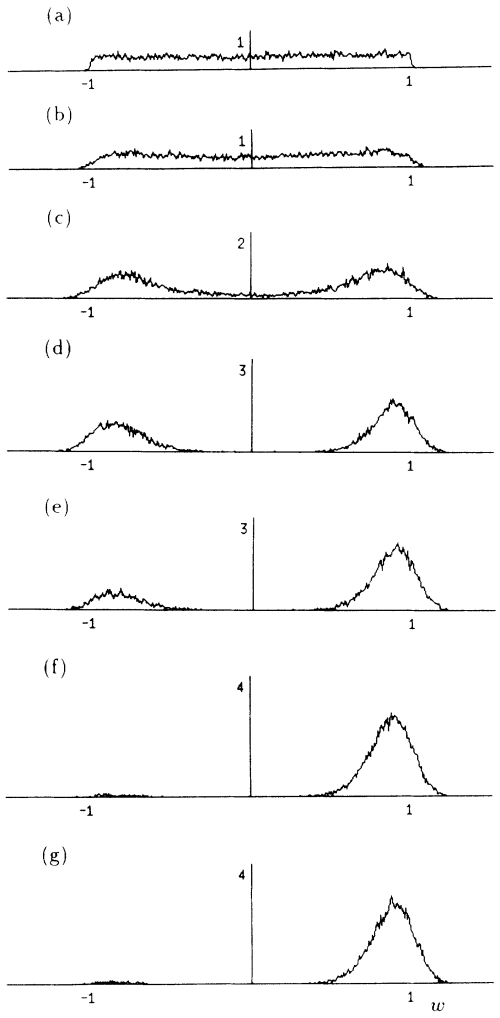

FIG. 2. Histogram found by simulation of the learning process with 10 000 networks and $\eta=0.05$. Parameters as in Fig. 1. (a) $t=1$. (b) $t=10$. (c) $t=100$. (d) $t=1000$. (e) $t=10\,000$. (f) $t=100\,000$. (g) $t=1\,000\,000$.

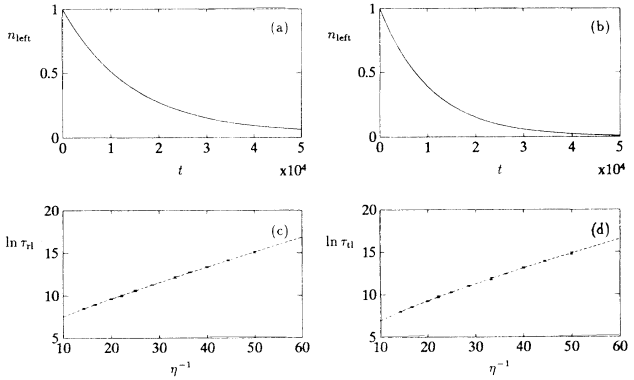

FIG. 3. (a) and (b) The occupation n_{left} as a function of time. (c) and (d) The transition time as a function of the learning parameter. Parameters as in Figs. 1 and 2. Dashed lines show the best possible fits of the form (10). In (b) and (d) the effect of the transition region is neglected.

Since for the moment we will focus on the transition from the local to the global minimum, we start our simulations with all networks in the left attraction region. During the learning process, we keep track of the occupation number $n_{\text{left}}(t)$, i.e., the fraction of the 10 000 networks that is still in the left attraction region at time t. With parameters as in Fig. 2, we obtain Fig. 3(a). After a short time, the occupation n_{left} decays exponentially. A comparison with Eq. (8) yields the experimental values for $\tau_{\text{r-l}}$ and $\tau_{\text{l-r}}$.

We will try to find mathematical expressions for these transition times. Denoting the boundary between the attraction region and the transition region by $w_{\text{t-l}}$, we can derive, using the master equation (3),

$$\begin{aligned}\frac{d}{dt}n_{\text{left}}(t)&=\frac{d}{dt}\int_{-\infty}^{w_{\text{t-l}}}dw'P(w',t)\\&=\int_{-\infty}^{w_{\text{t-l}}}dw'\int_{-\infty}^{\infty}dw[T(w'|w)P(w,t)-T(w|w')P(w',t)]\\&=-\int_{w_{\text{t-l}}}^{\infty}dw'\int_{-\infty}^{w_{\text{t-l}}}dw\ T(w'|w)P(w,t)+\int_{-\infty}^{w_{\text{t-l}}}dw'\int_{w_{\text{t-l}}}^{\infty}dw\ T(w'|w)P(w,t)\ .\end{aligned}\tag{9}$$

The first term in Eq. (9) corresponds to probability mass leaving the attraction region, the second term to mass entering this region.

Roughly speaking, the problem of calculating the transition time $\tau_{\text{r-l}}$ consists of two parts: the escape from the attraction region to the transition region and the question whether a network that managed the escape reaches the other attraction region or falls back into the attraction region it came from. In order to grasp the importance of this second part, the influence of the transition region on the transition time $\tau_{\text{r-l}}$, we have repeated our simulations with one important difference. Again starting with all networks in the left attraction region, we train the system as before. But if a network gets beyond the inflection point $w_{\text{t-l}}$ of the error potential, i.e., just in the transition region, we take it out of the simulation, put it directly in the right attraction region, and leave it there. This simulation is described by Eq. (9) with the second term deliberately set to zero. We find Eq. 3(b) instead of Fig. 3(a). Of course, the typical decay time, denoted by $\tau_{\text{t-l}}$, is much smaller than $\tau_{\text{r-l}}$.

Doing the same simulation with 10 000 networks for various learning parameters, we obtain Fig. 3(c) and Fig. 3(d), where the natural logarithm of the transition time is plotted against the reciprocal value of the learning parameter. The error bars give an indication of the error for each simulation. Transition times of the form

$$\tau_{\text{r-l}}=\frac{1}{\eta^{\alpha}}\exp\left[\frac{\tilde{\eta}}{\eta}+d\right]\tag{10}$$

are frequently encountered in the study of unstable stochastic systems [16]. Trying to fit our data points with a function of this form, we find the parameters

$$\alpha_{\text{r-l}}=0.7\pm0.4\ ,$$

$$\tilde{\eta}_{\text{r-l}}=0.16\pm0.02\ ,$$

$$d_{\text{r-l}}=4.4\pm0.7$$

for the "normal" simulations, and

$$\alpha_{\text{t-l}}=1.2\pm0.5\ ,$$

$$\tilde{\eta}_{\text{t-l}}=0.15\pm0.02\ ,$$

$$d_{\text{t-l}}=2.6\pm0.7$$

for the simulations in which we neglected the transition region. The dotted lines in these figures fit perfectly.

The close correspondence between the parameters $\tilde{\eta}$ in the two different simulations leads to the second hypothesis. Namely, that in order to compute or find a good estimate for the parameter $\tilde{\eta}_{\text{r-l}}$, we can restrict ourselves to the calculation of the transition time $\tau_{\text{t-l}}$, the average time to go from the attraction region to the transition region. In mathematical terms,

$$\lim_{\eta\to0}\eta\ln\tau_{\text{r-l}}=\tilde{\eta}_{\text{r-l}}\approx\tilde{\eta}_{\text{t-l}}=\lim_{\eta\to0}\eta\ln\tau_{\text{t-l}}\ .$$

In Sec. III we will give a theoretical argument in support of this second hypothesis.

Now, neglecting the first term in Eq. (9) and using the first hypothesis, that after a time $t\gg\tau_{\text{meso}}$ the shape of the mesostate stays the same, we find

$$\tilde{\eta}_{\text{r-l}}\approx\tilde{\eta}_{\text{t-l}}=-\lim_{\eta\to0}\eta\ln\left[\int_{w_{\text{t-l}}}^{\infty}dw'\int_{-\infty}^{w_{\text{t-l}}}dw\ T(w'|w)p_{\text{left}}(w)\right]\ .\tag{11}$$

In the rest of this paper we will concentrate on this equation, just to calculate the parameter $\tilde{\eta}$ in the expression for the transition time. There are several reasons for this restriction. First and most important of all, we will present a general scheme to calculate this parameter $\tilde{\eta}$, whereas we do not know how to predict the parameters α and d. Furthermore, even though we tried very hard to get the parameters as accurate as possible (simulations

with 10 000 networks for 11 different learning parameters with over more than 10 000 learning steps on the average), the uncertainty, especially in the parameters α and d, is relatively large. A true verification of the theoretical expressions for these parameters, if these expressions can be found, is therefore very difficult. Luckily, since $\tilde{\eta}$ is the parameter in the exponent, it is by far the most important parameter for practical purposes. We call it the "reference learning parameter." For if $\eta \ll \tilde{\eta}$, the probability to escape from the local minimum within an acceptable number of learning steps is negligible. On the other hand, if we choose η of the order $\tilde{\eta}$, the transition time will be limited.

III. MESOSTATES

In this section we will calculate the shapes of the mesostates for small learning parameters. To this end, we will use Van Kampen's system size approximation. This approximation of the master equation is valid for transition probabilities that can be written in the form [16]

$$T(\mathbf{w}'|\mathbf{w}) = T(\mathbf{w};\mathbf{w}'-\mathbf{w}) = \Omega^N \Psi\left[\frac{\mathbf{W}}{\Omega};\Delta\mathbf{W}\right], \qquad (12)$$

where Ω is a large parameter, in Van Kampen's terms the system size, and the jump $\Delta\mathbf{W}=\mathbf{W}'-\mathbf{W}$ is "extensive," i.e., independent of Ω. To prove that the transition probability (4) obeys Eq. (12), we rewrite

$$T(\mathbf{w}'|\mathbf{w}) = \frac{1}{\eta^N}\int d^n x\, \rho(\mathbf{w},\vec{x})\delta^N\left[\frac{\mathbf{w}'-\mathbf{w}}{\eta} - \mathbf{f}(\mathbf{w},\vec{x})\right]$$
$$= \frac{1}{\eta^N}\int d^n x\, \rho(\eta\mathbf{W},\vec{x})\delta^N(\Delta\mathbf{W} - \mathbf{f}(\eta\mathbf{W},\vec{x})),$$

with $\mathbf{W}\equiv\mathbf{w}/\eta$. Identifying the system size Ω as $1/\eta$, we find Eq. (12). However, Van Kampen's system size expansion is only valid in a neighborhood of a minimum $\mathbf{w}^*$ where the Hessian matrix $H(\mathbf{w})$ with elements

$$H_{ij}(\mathbf{w}) \equiv \frac{\partial^2 E(\mathbf{w})}{\partial w_i \partial w_j}$$

is positive definite [17]. This is the general definition of the attraction region. At the minimum itself H is always positive definite. Regions with one or more negative eigenvalues of the Hessian are called transition regions.

The result of the expansion is quite simple: the asymptotic expansion of the stationary probability distribution for large Ω, i.e., small learning parameters, is a Gaussian with the average and covariance matrix given by the stable fixed points of a set of coupled nonlinear differential equations. In one dimension they are written

$$\frac{1}{\eta}\frac{d}{dt}\langle w\rangle = f(\langle w\rangle) + \tfrac{1}{2}f''(\langle w\rangle)\Sigma^2,$$
$$\frac{1}{\eta}\frac{d}{dt}\Sigma^2 = 2f'(\langle w\rangle)\Sigma^2 + \eta D(\langle w\rangle), \qquad (13)$$

where $\langle w\rangle$ is the average of the mesostate and $\Sigma^2 \equiv \langle w^2\rangle - \langle w\rangle^2$ denotes the variance. The diffusion $D(w)$ is a measure of the fluctuations of the learning rule. The general definition is

$$D_{ij}(\mathbf{w}) \equiv \int d^n x\, \rho(\mathbf{w},\vec{x}) f_i(\mathbf{w},\vec{x}) f_j(\mathbf{w},\vec{x}).$$

The macroscopic equations (13) given an indication of the fundamental difference between an attraction region and a transition region. In the attraction region $f'(w)<0$, and thus the fluctuations tend to some equilibrium value, proportional to the learning parameter. In the transition region, on the other hand, $f'(w)>0$, and thus the fluctuations show a tendency to explode, independent of the value of the learning parameter. This important difference between attraction and the transition regions is a strong argument in favor of the second hypothesis, which claims that in order to calculate the reference learning parameter, the influence of the transition region can be neglected. It also explains why the probability mass in the transition regions (for small learning parameters η and after some initial time) is negligible in comparison with the probability mass in the attraction regions. The same arguments apply in higher-dimensional cases. Even if the Hessian has only one negative eigenvalue, the fluctuations will show a tendency to explode in the direction of the corresponding eigenvector.

A generalization of the set of equations (13) to more dimensions is straightforward [7]. The stable fixed points of these macroscopic equations are given by

$$\langle\mathbf{w}\rangle = \mathbf{w}^* + O(\eta), \quad \Sigma^2 = \eta K + O(\eta^2),$$

where the normalized covariance matrix K is the solution of the matrix equation

$$HK + KH = D. \qquad (14)$$

The curvature $H \equiv H(\mathbf{w}^*)$ and the diffusion $D \equiv D(\mathbf{w}^*)$ are evaluated at the minimum. The typical relaxation time to such a situation is

$$\tau_{\text{meso}} = \frac{1}{\eta\lambda_{\min}},$$

with $\lambda_{\min}$ the smallest eigenvalue of the matrix H. According to the first hypothesis, after a time of this order, the shape of the mesostate remains constant.

Finally, as a result of Van Kampen's system size expansion, up to lowest order in η the mesostate can be written

$$p_{\text{meso}}(\mathbf{w}) = \frac{C_{\text{meso}}}{(2\pi\eta)^{N/2}(\text{Det}K)^{1/2}} \times \exp\left[-\frac{(\mathbf{w}-\mathbf{w}^*)^T K^{-1}(\mathbf{w}-\mathbf{w}^*)}{2\eta}\right] \times \mathbb{1}_{\text{meso}}(w). \qquad (15)$$

The function $\mathbb{1}_{\text{meso}}(w)$ is equal to 1 in the attraction region of the minimum $\mathbf{w}^*$ and equal to 0 outside this region. The constant C_{meso} ensures the proper normalization of the mesostate. For small learning parameters η, the error introduced by taking $C_{\text{meso}}=1$ is negligible.

IV. TRANSITION TIMES

In this section we will calculate the (most dominant term of the) transition time from one minimum to another. We will start with the calculation for a one-dimensional "neural network." Later we will extend this calculation to higher-dimensional systems. A two-dimensional network will be treated in more detail.

A. The one-dimensional case

According to Eq. (15), the shape of the left mesostate in the one-dimensional example discussed in Sec. II obeys

$$p_{\text{left}}(w)=\left[\frac{\lambda_l}{\pi\eta D_l}\right]^{1/2}\exp\left[-\frac{\lambda_l(w-w_l^*)^2}{\eta D_l}\right],$$

with λ_l the second derivative of the error potential and D_l the diffusion at the minimum w_l^*. This shape and the specific form of the transition probability can be substituted into Eq. (11), yielding

$$\tilde{\eta}_{\text{t-l}}=-\lim_{\eta\to 0}\eta\ln\left\{\int_{w_{\text{t-l}}}^{\infty}dw'\int_{-\infty}^{w_{\text{t-l}}}dw\int dx\,\rho(w,x)\delta(w'-w-\eta f(w,x))\exp\left[-\frac{\lambda_l(w-w_l^*)^2}{\eta D_l}\right]\right\}. \tag{16}$$

In the term between braces, denoted as $\mathfrak{T}$, we integrate over w' and write the integration over w in the form of a theta function [$\Theta(x)=1$ if $x>0$, $\Theta(x)=0$ if $x<0$],

$$\mathfrak{T}=\int dw\int dx\,\Theta(w_{\text{t-l}}-w)\Theta(w-w_{\text{t-l}}+\eta f(w,x))\rho(w,x)\exp\left[-\frac{\lambda_l(w-w_l^*)^2}{\eta D_l}\right].$$

So, we have to integrate the function $\rho(w,x)\exp[\ ]$ over all pairs (w,x) that obey

$$w<w_{\text{t-l}}<w+\eta f(w,x),$$

i.e., for which the weight before learning is in the attraction region and after learning in the transition region. If we make the substitution $z=(w_{\text{t-l}}-w)/\eta$ and write out the exponent, we obtain

$$\mathfrak{T}=\eta A(\eta)\exp\left[-\frac{\lambda_l(w_{\text{t-l}}-w_l^*)^2}{\eta D_l}\right],$$

with

$$A(\eta)\equiv\int_0^{\infty}dz\int dx\,\Theta(f(w_{\text{t-l}}-\eta z,x)-z)\rho(w_{\text{t-l}}-\eta z,x)\exp\left[\frac{2\lambda_l(w_{\text{t-l}}-w_l^*)z-\lambda_l\eta z^2}{D_l}\right].$$

Going back to Eq. (16), we find

$$\tilde{\eta}_{\text{t-l}}=\frac{\lambda_l(w_{\text{t-l}}-w_l^*)^2}{D_l}-\lim_{\eta\to 0}\eta\ln A(\eta).$$

Assuming continuity of $A(\eta)$ at $\eta=0$, this second term can be neglected if $A(0)<\infty$, i.e., if

$$\int dx\,\Theta(f(w_{\text{t-l}},x))\rho(w_{\text{t-l}},x)\times\exp\left[\frac{2\lambda_l(w_{\text{t-l}}-w_l^*)f(w_{\text{t-l}},x)}{D_l}\right]<\infty.$$

This sufficient but not necessary condition is fulfilled if the probability to make very large steps decays faster than exponentially, so, for example, if this probability is a Gaussian as in the example of Sec. II, or if the stochastic force has an upper limit, as in practical situations. Therefore, we will not consider cases in which this condition is violated.

Summarizing, the reference learning parameter $\tilde{\eta}_{\text{r-l}}$ for the transition from the left to the right attraction region obeys

$$\tilde{\eta}_{\text{r-l}}\approx\tilde{\eta}_{\text{t-l}}=\frac{\lambda_l(w_{\text{t-l}}-w_l^*)^2}{D_l}, \tag{17}$$

with λ_l the curvature of the error potential and D_l the fluctuations in the stochastic force, both at the position of the left minimum w_l^*. $w_{\text{t-l}}$ stands for the inflection point of the error potential at the left side of the potential maximum. Note that a similar expression is valid for the transition time from right to left. Furthermore, since in the derivation we never used explicit information about the learning rule, the result is applicable to any one-dimensional learning rule that can be written in the form (1).

Once the transition times from the local to the global minimum and vice versa are known, the stationary occupation numbers can be calculated. Equation (7) yields

$$\frac{n_{\text{left}}(\infty)}{n_{\text{right}}(\infty)}=\frac{\tau_{\text{r-l}}}{\tau_{\text{l-r}}}.$$

In the limit of small learning parameters, we find

$$\lim_{\eta\to 0}\eta\ln\left[\frac{n_{\text{left}}(\infty)}{n_{\text{right}}(\infty)}\right]=-(\tilde{\eta}_{\text{l-r}}-\tilde{\eta}_{\text{r-l}})\approx-\frac{\lambda_r(w_{\text{t-r}}-w_r^*)^2}{D_r}+\frac{\lambda_l(w_{\text{t-l}}-w_l^*)^2}{D_l}.$$

Thus, the final stationary probability to find the network in the neighborhood of the local or the global minimum does not directly depend on the value of the error potential at either minimum. Instead, it depends on the curvature, the squared distance to the inflection point, and the fluctuations in the learning rule. The product of the curvature and the squared distance can be viewed as a rough measure of the difference in the error potential between the minimum and the inflection point. In general, there is no direct relation between the error potential and the diffusion. So, it might even be possible to construct an example in which the stationary occupation number at a local minimum is larger than at the global minimum.

We compare Eq. (17) with the simulations performed in Sec. II,

$\tilde{\eta}_{\text{t-l}}=0.146$, theory

$\tilde{\eta}_{\text{r-l}}=0.16\pm0.02$, simulations 1

$\tilde{\eta}_{\text{t-l}}=0.15\pm0.02$, simulations 2 .

The parameter $\tilde{\eta}_{\text{t-l}}$ calculated from Eq. (17) yields a good estimate for $\tilde{\eta}_{\text{r-l}}$ found by simulations.

B. Higher-dimensional learning rules

We will extend the derivation given above to N-dimensional learning rules. For any network state $\mathbf{w}$, the Hessian matrix $H(\mathbf{w})$ has N real eigenvalues that can be either positive, zero, or negative. The weight space can be divided in simply connected regions by counting the number of positive eigenvalues. In attraction regions, all eigenvalues are positive. In transition regions, at least one eigenvalue must be negative. Boundaries between attraction and transition regions are characterized by the presence of at least one eigenvalue equal to zero, and all others positive. Denoting the occupation number of the attraction region α by $n_\alpha(t)$, we can generalize Eq. (7) to

$$\frac{dn_\alpha(t)}{dt}=\sum_\beta\left[\frac{n_\beta(t)}{\tau_{\alpha\beta}}-\frac{n_\alpha(t)}{\tau_{\beta\alpha}}\right], \tag{18}$$

where $\tau_{\alpha\beta}$ is the transition time from attraction region β to α. Again, in writing down Eq. (18), we implicitly make two assumptions. We use the first hypothesis that the shape of the mesostates in the attraction regions is independent of time. Furthermore, we assume that the probability mass in the transition regions is negligible. The matrix with elements $1/\tau_{\alpha\beta}$ is a stochastic matrix. It has, at least, one eigenvalue equal to 1. The corresponding right eigenvector is the stationary distribution. The next to largest eigenvalue yields the relaxation time. In order to calculate the stationary distribution or the relaxation time, (the asymptotic expansions of) all matrix elements must be known.

Let us consider an "easy" transition from attraction region $\mathcal{L}$ to another attraction region $\mathcal{R}$, through a transition region $\mathcal{T}$. A transition is called "easy" if $\mathcal{T}$ is a transition region joining $\mathcal{L}$ and $\mathcal{R}$ in which the Hessian $H(\mathbf{w})$ has only one negative eigenvalue. For simplicity, $\mathcal{T}$ is supposed to be the only way to go from $\mathcal{L}$ or $\mathcal{R}$ without changing the sign of more than one eigenvalue, i.e., $\mathcal{T}$ is typical for the transition from $\mathcal{L}$ to $\mathcal{R}$. For small learning parameters, only this path will give a contribution. If there are more transition regions satisfying this condition, one should calculate the reference learning parameters for these different transition regions separately. The smallest reference learning parameter then yields the "easiest" transition from $\mathcal{L}$ to $\mathcal{R}$. If there is no transition region with just one negative eigenvalue connecting two attraction regions, the only reasonable way to go from one region to the other is through a succession of easy transitions.

The boundary between the attraction and transition regions is denoted by $\mathcal{TL}$. According to the second hypothesis, the reference learning parameter $\tilde{\eta}_{\text{r-l}}$ for the transition from attraction region $\mathcal{L}$ to $\mathcal{R}$ is approximately equal to the parameter $\tilde{\eta}_{\text{t-l}}$ that appears in the transition time from attraction region $\mathcal{L}$ to transition region $\mathcal{T}$. Comparing with Eq. (11), we now have

$$\tilde{\eta}_{\text{t-l}}=-\lim_{\eta\to0}\eta\ln\left\{\int_{\mathcal{T}}d^Nw'\int_{\mathcal{L}}d^Nw\int d^nx\,\rho(\mathbf{w},\vec{x})\delta^N(\mathbf{w}'-\mathbf{w}-\eta\mathbf{f}(\mathbf{w},\vec{x}))p_l(\mathbf{w})\right\}.$$

In the term between braces, we have to integrate over all $\mathbf{w}$ and $\vec{x}$ such that

$$\mathbf{w}\in\mathcal{L}\text{ and }\mathbf{w}+\eta\mathbf{f}(\mathbf{w},\vec{x})\in\mathcal{T}\ .$$

Just as in the preceding paragraph, the term between brackets can be divided in two parts: a contribution from the boundary $\mathcal{TL}$ between the attraction and the transition region and a rest term like the term $A(\eta)$ in the one-dimensional case. Again, it is easy to show that these rest terms can be neglected, except for those cases where the probability to make very large steps does not decay fast enough. However, there is an important difference: the boundary $\mathcal{TL}$ between the attraction region $\mathcal{L}$ and the transition region $\mathcal{T}$ is no longer a point, but an $(N-1)$-dimensional manifold. We obtain

$$\tilde{\eta}_{\text{t-l}}=-\lim_{\eta\to0}\eta\ln\left\{\int_{\mathcal{TL}}d^{N-1}w\exp\left[-\frac{(\mathbf{w}-\mathbf{w}_l^*)^TK_l^{-1}(\mathbf{w}-\mathbf{w}_l^*)}{2\eta}\right]\right\}.$$

This integral can be approximated using the method of steepest descent. The largest contribution for small learning parameters is found when the term between brackets has a maximum on $\mathcal{TL}$. In other words, the largest contribution for small learning parameters comes from the "easiest" path from the local minimum to the transition region. In the determination of this easiest path, the normalized covariance matrix K_l accounts for the effect of the local fluctuations. So, finally,

$$\tilde{\eta}_{\text{r-l}} \approx \tilde{\eta}_{\text{t-l}} = \inf_{\mathbf{w}\in\mathcal{TL}} \left[\frac{(\mathbf{w}-\mathbf{w}_l^*)^T K_l^{-1} (\mathbf{w}-\mathbf{w}_l^*)}{2} \right] . \quad (19)$$

In the next paragraph we will discuss a two-dimensional example. There it will become clear how this expression can be used to obtain quantitative results that can be compared with results from simulations.

C. A two-dimensional example

The learning rule discussed in Sec. II can be generalized to N dimensions. The network senses its environment through a, now N-dimensional, Gaussian filter of width σ. For the "real" distribution $\rho_0(\mathbf{x})$, we take a sum of M Gaussian functions, all with variance χ, positions $\mathbf{m}_\alpha$, $\forall\ \alpha=1,\ldots,M$ and relative weights $r_\alpha \geq 0$ such that $\sum_{\alpha=1}^{M} r_\alpha = 1$,

$$\rho_0(\mathbf{x}) = \frac{1}{(2\pi\chi)^{N/2}} \sum_{\alpha=1}^{M} r_\alpha \exp\left[-\frac{(\mathbf{x}-\mathbf{m}_\alpha)^2}{2\chi^2} \right] .$$

The error potential corresponding to the Grossberg learning rule

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

is

$$E(\mathbf{w}) = -\sigma^2 \ln\left\{ \sum_{\alpha=1}^{M} r_\alpha \exp\left[-\frac{(\mathbf{w}-\mathbf{m}_\alpha)^2}{2(\sigma^2+\chi^2)} \right] \right\} .$$

We will restrict ourselves to a two-dimensional example and to an input distribution consisting of four Gaussian functions, obeying

$$\mathbf{m}_1=(1,1),\quad \mathbf{m}_2=(1,-1) ,$$

$$\mathbf{m}_3=(-1,1),\quad \mathbf{m}_4=(-1,-1) ,$$

$$r_1=\frac{(1+a_1)(1+a_2)}{4},\quad r_2=\frac{(1+a_1)(1-a_2)}{4} ,$$

$$r_3=\frac{(1-a_1)(1+a_2)}{4},\quad r_4=\frac{(1-a_1)(1-a_2)}{4} .$$

Using the definitions $\epsilon_i \equiv \operatorname{arctanh}(a_i)$ for $i=1,2$ and $\beta \equiv 1/(\sigma^2+\chi^2)$, the error potential can be written

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

We will work with $0 < \epsilon_2 \leq \epsilon_1 < \epsilon^*(\beta)$, such that there are always four minima, with one global minimum in the neighborhood of (1,1). The matrix of second derivatives, the Hessian $H(w_1,w_2)$, is diagonal

$$H_{ij}(w_1,w_2) = \beta\sigma^2 \left[1 - \frac{\beta}{\cosh^2(\beta w_i + \epsilon_i)} \right] \delta_{ij} . \quad (20)$$

This makes it easy to divide the weight space into attraction regions and transition regions. The diffusion matrix $D(w_1,w_2)$ obeys

$$D_{ij}(w_1,w_2) = \{(\beta\sigma^2)^2 [w_i^2 - 2w_i \tanh(\beta w_i + \epsilon_i) + 1] + \beta\sigma^2\chi^2\}\delta_{ij} . \quad (21)$$

For our figures and numerical solutions, we choose $\beta=2.5$, $\epsilon_1=0.4$, $\epsilon_2=0.2$, and $\sigma=\chi$. The error potential is plotted in Fig. 4(a). The attraction and transition regions are shown in Fig. 4(b).

Let us consider the transition from attraction region 2 at the lower right corner to attraction region 1 in the neighborhood of the global minimum at the upper right corner. To obtain the reference learning parameter for this transition, we have to go through the following steps (we give the numerical results for our specific example in three significant digits).

(1) Calculate (numerically) the position of the local minimum $\mathbf{w}_2^*$,

$$\mathbf{w}_2^* = (0.994, -0.978) .$$

(2) Substitute this into Eqs. (20) and (21) to obtain the Hessian H_2 and the diffusion matrix D_2 at this minimum,

$$H_2 = \begin{bmatrix} 0.484 & 0 \\ 0 & 0.445 \end{bmatrix}, \quad D_2 = \begin{bmatrix} 0.103 & 0 \\ 0 & 0.111 \end{bmatrix} .$$

(3) Use Eq. (14) to calculate the normalized covariance matrix K_2 and its inverse K_2^{-1},

$$K_2 = \begin{bmatrix} 0.106 & 0 \\ 0 & 0.125 \end{bmatrix}, \quad K_2^{-1} = \begin{bmatrix} 9.40 & 0 \\ 0 & 8.02 \end{bmatrix} .$$

(4) Determine the boundary 12 between the attraction and transition region,

$$12 = \{(w_1,w_2) \in \mathbb{R}^2 | w_1 > 0.253 \ \wedge\ w_2 = -0.493\} .$$

(5) Solve Eq. (19),

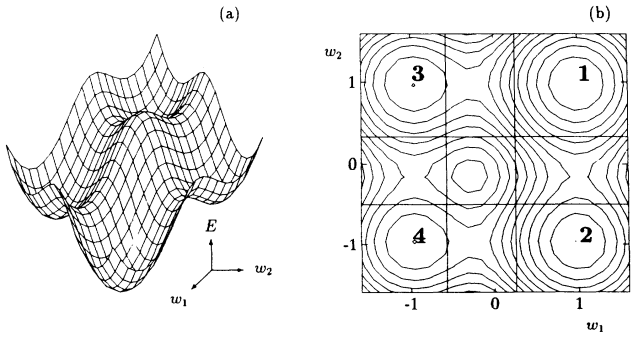

FIG. 4. (a) Error potential $E(\mathbf{w})$ for $\beta=2.5$, $\epsilon_1=0.4$, $\epsilon_2=0.2$, and $\sigma=\chi$. (b) Contour plot showing the attraction and transition regions.

$$\tilde{\eta}_{12} = \inf_{w_1 > 0.253, w_2 = -0.493} \left[\frac{9.40(w_1 + 0.994)^2 + 8.02(w_2 + 0.978)^2}{2} \right],$$

to obtain the reference learning parameter

$$\tilde{\eta}_{12} = 0.944 .$$

Similar calculations yield the reference learning parameters for other transitions. We compare two of them with simulations, similar to the first type discussed in Sec. II, so without neglecting the transition region. The results from simulations with an ensemble of 100 networks are given in Figs. 5(a) and 5(b). The best possible fits in these figures yield

	$\tilde{\eta}_{12}$	$\tilde{\eta}_{13}$
simulations	1.1±0.2	0.5±0.1
theory	0.944	0.543

Again there is a close correspondence between theory and simulations.

V. SUMMARY AND DISCUSSION

A better understanding of the global performance of on-line learning neural networks is very important, both from a theoretical and a practical point of view. In this paper we studied the effect of the learning parameter on the transition time from one minimum to another. Using Van Kampen's system size expansion we showed that the transition time grows exponentially with $\tilde{\eta}/\eta$. With a learning parameter much smaller than the so-called reference learning parameter $\tilde{\eta}$, it is almost impossible to go from this minimum to the other one within a reasonable number of learning steps. Starting from two hypotheses, supported by both simulations and theoretical arguments, we presented a general scheme to calculate this reference learning parameter. It depends on the local fluctuations in the learning rule, the local curvature of the error potential, and the distance between the minimum and the boundary of its attraction region. Correction terms must be included only if the probability to make very large steps does not decay fast enough. This will rarely occur in practical situations. Simulations confirm the theoretical results.

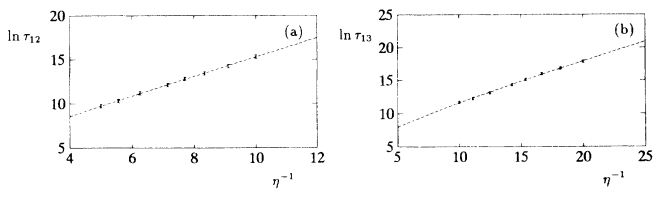

FIG. 5 The transition time as a function of the learning parameter for the two-dimensional learning rule. Parameters as in Fig. 4. Dashed lines show the best possible fits of the form (10). (a) τ_{12}. (b) τ_{13}.

The correctness of the theory depends on the validity of the two hypotheses. The first hypothesis claims that the interaction between the mesostates does not affect their shape. The local convergence to a Gaussian distribution as predicted by Van Kampen's expansion takes place on a time scale of order $1/\eta$. The time scale corresponding to the global interaction is of order $\exp(\tilde{\eta}/\eta)$. The existence of two distinct time scales, of which the time scale concerned with the maintenance of the local shape is the smallest, makes the first hypothesis very plausible.

The second hypothesis states that in order to calculate or estimate the reference learning parameter, the influence of the transition region can be neglected. In other words, we assume that the path from the minimum to the inflection point is much "harder" than the path from the inflection point to the maximum because of the larger fluctuations in the transition region. This assumption is only valid if the total drift for both paths is of the same order of magnitude. It is possible to construct error potentials for which this condition is violated. The results for the error potentials used in the simulations are promising. Further studies on error potentials for learning rules in neural networks must yield a better insight into the validity of the second hypothesis.

The final theoretical result is simple and elegant. Nevertheless, its usefulness in practical calculations is limited for several reasons. First of all, we assumed throughout the whole paper that the error potential and the diffusion matrix can be calculated. They depend not only on the learning rule and the network structure, but also on the set of training patterns. Therefore, *a priori* knowledge of the input probability distribution is required for a precise calculation of the reference learning parameter. If this information is not available, we may try to estimate the reference learning parameter from the statistics of the network weights during training. In [9], this strategy is followed to obtain a reasonable learning parameter in a changing environment. The same approach can be used to estimate the normalized covariance matrix and the position of the minimum. The problem is how to subtract more global information, e.g., the positions of boundaries between attraction and transition regions, from the statistics of the weights.

But even if the error potential and the diffusion matrix are known, numerical calculation of boundaries between attraction and transition regions can be very difficult. Since the Hessian and diffusion matrix in the two-dimensional example discussed in Sec. IV are diagonal, this problem did not appear. In practice, the error potential and the diffusion do not have such a nice symmetry.

Furthermore, in order to calculate the stationary distribution or the relaxation time, the transition times between all possible pairs of minima must be calculated. For large networks with many minima this seems a hopeless task. The challenge remains to apply our calculation scheme to a more practical example.

ACKNOWLEDGMENTS

This work was partly supported by the Dutch Foundation for Neural Networks and the Human Frontiers Science Program Organization. We would like to thank Peter Johannesma for stimulating discussions.

[1] D. Rumelhart, G. Hinton, and R. Williams, Nature **323**, 533 (1986).
[2] E. Oja, J. Math. Biol. **15**, 267 (1982).
[3] T. Kohonen, Biol. Cybernet. **43**, 59 (1982).
[4] D. Hebb, *The Organization of Behavior* (Wiley, New York, 1949).
[5] H. Ritter and K. Schulten, Biol. Cybernet. **60**, 59 (1988).
[6] D. Clark and K. Ravishankar, Neural Networks **3**, 87 (1990).
[7] T. Heskes and B. Kappen, Phys. Rev. A **44**, 2718 (1991).
[8] C. Kuan and K. Hornik, IEEE Trans. Neural Networks **2**, 484 (1991).
[9] T. Heskes and B. Kappen, Phys. Rev. A **45**, 8885 (1992).
[10] N. Van Kampen, *Stochastic Processes in Physics and Chemistry* (North-Holland, Amsterdam, 1981).
[11] J. Hopfield, Proc. Natl. Acad. Sci. U.S.A. **79**, 2554 (1982).
[12] D. Bedeaux, K. Lakatos-Lindenberg, and K. Shuler, J. Math. Phys. **12**, 2116 (1971).
[13] H. Kushner and D. Clark, *Stochastic Approximation Methods for Constrained and Unconstrained Systems* (Springer, New York, 1978).
[14] L. Ljung, IEEE Trans. Autom. Control. **AC-22**, 551 (1977).
[15] S. Grossberg, J. Stat. Phys. **48**, 105 (1969).
[16] C. Gardiner, *Handbook of Stochastic Methods*, 2nd ed. (Springer, Berlin, 1985).
[17] If there exists no error potential, the matrix H, defined as $H_{ij}(\mathbf{w}) \equiv -\partial f_i(\mathbf{w})/\partial w_j$, is no longer symmetric. The analysis remains valid if Eq. (14) is changed into $HK + KH^T = D$. Furthermore, to divide the weight space in attraction and transition regions, one has to consider the real parts of the eigenvalues of the matrix H.