On Fokker-Planck approximations of on-line learning processes

Tom Heskes

Beckman Institute and Department of Physics,
University of Illinois at Urbana-Champaign,
405 North Mathews Avenue, Urbana, Illinois 61801, U.S.A.

present address:
Department of Medical Physics and Biophysics,
University of Nijmegen,
Geert Grooteplein 21, 6525 EZ Nijmegen, The Netherlands

Journal of Physics A, 27:5145–5160, 1994; PACS: 87.10.+e

Abstract

There are several ways to describe on-line learning in neural networks. The two major ones are a continuous-time master equation and a discrete-time random-walk equation. The random-walk equation is obtained in case of fixed time intervals between subsequent learning steps, the master equation results when the time intervals are drawn from a Poisson distribution. Following Van Kampen [1], we give a rigorous expansion of both the master and the random-walk equation in the limit of small learning parameters. The results explain the difference between the Fokker-Planck approaches proposed by Radons *et al.* [2] and Hansen *et al.* [3]. Furthermore, we find that the mathematical validity of these approaches is restricted to local properties of the learning process. Yet Fokker-Planck approaches are often suggested as models to study global properties, such as mean first passage times and stationary solutions. To check their accuracy and usefulness in these situations we compare simulations of two learning procedures with exactly the same drift vector and diffusion matrix, the only moments that are considered in a Fokker-Planck approximation. The simulations show that the mean first passage times for these two learning procedures diverge rather than converge for small learning parameters. We reach the conclusion that Fokker-Planck approaches are not accurate enough to compute global properties of on-line learning processes.

1 Introduction

1.1 Outline

On-line learning stands for learning in artificial neural networks where at each learning step one of the patterns is drawn at random from the total set of training patterns and is presented to the network. This is in constrast with batch-mode learning where the learning rule involves first an average over the whole training set and is only then applied. Batch-mode learning is deterministic, whereas on-line learning, through the random presentation of patterns, is stochastic. This stochasticity can be very helpful, e.g., to speed up learning or to escape from local minima from the error potential on which the (average) learning rule performs a gradient descent.

In section 1.2 we give a few descriptions of on-line learning processes. A discrete-time random-walk equation is obtained if the time intervals between subsequent learning steps are taken constant, a continuous-time master equation if these time intervals are Poisson distributed. Both the master and the random-walk equation cannot be solved in general.

Many researchers therefore propose to describe on-line learning processes by an approximate Fokker-Planck equation [2, 4, 5, 6, 7, 8, 3, 9]. In sections 2.1 and 2.2 , we will review the approaches suggested by Radons *et al.* [2, 9] and Hansen *et al.* [3], respectively. These two approaches differ by the form of the diffusion term. The lack of a firm (common) theoretical basis makes it difficult to judge the validity of these two approaches and to explain their difference. Van Kampen's approximation [1], however, is known to be a proper "small-fluctuations" expansion, valid for small learning parameters η. In section 2.3 we will rederive Van Kampen's expansion of a continuous-time master equation. Its derivation for the discrete-time random walk, treated in section 2.4, is somewhat more complicated. The results from these sections do not only explain the difference between the Fokker-Planck approaches of Radons and Hansen, but they also indicate that the Fokker-Planck approaches are only locally valid, i.e., on relatively short time scales or in a local neighborhood of minima of the error potential. Strictly speaking, global properties of on-line learning processes, such as mean first passage times and stationary solutions, are outside this validity regime.

Nevertheless, if viewed as models instead of as proper expansions, Fokker-Planck approaches might still be useful to describe global properties of on-line learning. Several suggestions in this direction have been made in the literature [2, 6, 5, 7, 8]. In section 3, we will discuss the accuracy of Fokker-Planck approaches in predicting mean first passage times. For the one-dimensional toy problem of section 3.2, the Fokker-Planck approaches yield closed expressions for mean first passage times that can be integrated numerically and compared with Monte-Carlo simulations of the on-line learning process. In sections 3.3 and 3.4, we describe Monte-Carlo simulations of the Kohonen learning rule and on-line backpropagation. In both cases, we compare the mean first passage times for the on-line learning process with those for the corresponding "Langevin-type" learning process. The Langevin-type learning rule is defined as the batch-mode learning rule with additive noise such that the first two moments (drift and diffusion) are completely equivalent to the first two moments of the on-line learning rule. Since Fokker-Planck approaches are based solely on these two moments of the transition matrix, they predict the same results for on-line learning and Langevin-type learning. Is this correct?

1.2 Definitions and background

At each learning step, a training pattern x is drawn at random from the total training set and presented to the network. The vector x denotes the combination of input vector and desired output vector for supervised learning or just the input vector for unsupervised learning. The weight change at iteration step i is given by

$$\Delta w_i \;=\; w_{i+1} - w_i \;=\; \eta\, f(w_i, x)\,, \tag{1}$$

with w_i the weight vector at iteration step i, which includes the strengths of all synapses and thresholds, η the learning parameter, and $f(.,.)$ the particular learning rule. In the following we will use a one-dimensional notation for simplicity. The description (1) is valid for a large class learning rules in neural network literature. Well-known examples are the (unsupervised) Kohonen learning rule [10] and the (supervised) backpropagation learning rule [11] (see sections 3.3 and 3.4).

On-line learning described by (1) is a Markov process. The probability $p_i(w)$ for the system to be in state w after i learning iterations obeys the random-walk equation [12, 2, 13]

$$p_{i+1}(w) \;=\; \int dw'\, T(w|w')\, p_i(w')\,, \tag{2}$$

with transition probability $T(w|w')$ to go from an old state w' to a new one w given by

$$T(w|w') \;=\; \int dx\, \rho(x)\, \delta(w - w' - \eta\, f(w', x))\,, \tag{3}$$

with $\rho(x)$ the probability density function of training patterns. We will denote an average with respect to $\rho(x)$ by $\langle . \rangle_x$. The average can be over a continuous distribution (as in section 3.3) as well as over a finite training set (as in section 3.4).

We still have the freedom to choose the points of time t_i of the iteration steps i. We write

$$t_{i+1} \equiv t_i + \Delta t\,.$$

There are two popular ways to choose the time intervals Δt. The most obvious choice is constant time intervals, i.e., time intervals chosen from the "distribution"

$$\varrho(\Delta t) \;=\; \delta(\Delta t - \tau)\,.$$

Then the probability $P(w,t)$ to be in state w at *time* t follows

$$P(w, t+\tau) \;-\; P(w,t) \;=\; \int dw'\, [T(w|w')\, P(w',t) \;-\; T(w'|w)\, P(w,t)]\,, \tag{4}$$

which is just the random-walk equation (2) in a different notation. For Poisson-distributed time intervals, i.e.,

$$\varrho(\Delta t) \;=\; \tau\, \exp\left[\frac{-\Delta t}{\tau}\right]\,,$$

the random-walk equation (2) transforms into the continuous-time master equation [14, 4]

$$\tau \frac{\partial}{\partial t} P(w,t) \;=\; \int dw'\, [T(w|w')\, P(w',t) \;-\; T(w'|w)\, P(w,t)]\,. \tag{5}$$

This transformation is exact for all times t and learning parameters η. It can be shown that at long times t the solutions $P(w,t)$ of the discrete-time random walk (4) and the continuous-time master equation (5) approach each other [14, 9]. The Kramers-Moyal expansion

$$\int dw'\, [T(w|w')\, P(w',t) \;-\; T(w'|w)\, P(w,t)] \;=\; \sum_{n=1}^{\infty} \frac{(-\eta)^n}{n!} \frac{\partial^n}{\partial w^n} [a_n(w) P(w,t)]\,, \tag{6}$$

with the moments $a_n(w)$ defined by

$$a_n(w) \;\equiv\; \langle f^n(w,x) \rangle_x\,,$$

is just another way to write down the master equation (5) or the random-walk equation (4) [1]. In general, neither the random-walk equation (4) nor the master equation (5) can be solved analytically. A way to proceed is to look for approximations valid for small learning parameters η.

2 Fokker-Planck approximations for on-line learning

2.1 Radons' Fokker-Planck equation

Radons *et al.* [2, 9] (see also [7, 8]) truncate the Kramers-Moyal expansion (6) after two terms to obtain the Fokker-Planck equation

$$\tau \frac{\partial}{\partial t} Q(w,t) \;=\; -\eta \frac{\partial}{\partial w}[a_1(w) Q(w,t)] \;+\; \frac{\eta^2}{2} \frac{\partial^2}{\partial w^2}[a_2(w) Q(w,t)] \,. \tag{7}$$

Even though the Kramers-Moyal expansion does indeed look like an expansion in the learning parameter η, one has to be very careful with a truncation after any number of terms since the probability distribution $P(w,t)$ itself is also a function of η. This can be seen most easily by substitution of the stationary solution of the (one-dimensional) Fokker-Planck equation

$$Q_{\mathrm{stat}}(w) \propto \frac{1}{a_2(w)} \exp\left[\frac{2}{\eta} \int^w dw' \, \frac{a_1(w')}{a_2(w')}\right] \tag{8}$$

into the Kramers-Moyal expansion (6). The first two terms in this expansion exactly cancel each other (of course), but all higher order terms are of the same order of magnitude in η as the first two terms. We are by no means allowed to claim that the stationary solution (8) is some consistent approximation of the true stationary solution of the master equation (5). In [15, 16], conditions on the transition matrix $T(w|w')$ are stated that justify a full use of the Fokker-Planck approximation (7) (see also section 2.3). These conditions do not hold for the transition probability (3). For a further explanation we refer to the standard text books [15, 16] or the book chapter [13].

2.2 Hansen's Fokker-Planck equation

Hansen *et al.* [3] arrive at a slightly different Fokker-Planck equation through a quite different route. They average the dynamics of the weights (1) over a large number $1 \ll n \ll 1/\eta$ of learning steps. Neglecting higher order terms and assuming independence between subsequent weight changes, they obtain

$$w(t + n\tau) - w(t) \;=\; \eta \, n \, a_1(w) \;+\; \eta \sqrt{n \tilde{a}_2(w)} \, \xi \,, \tag{9}$$

with ξ Gaussian white noise (zero average, unit standard deviation) and

$$\tilde{a}_2(w) \;=\; \Big\langle f^2(w,x) \Big\rangle_x \;-\; \langle f(w,x) \rangle_x^2 \;=\; a_2(w) - a_1^2(w) \,.$$

Equation (9) is called a "Langevin-type" equation. It can be viewed as a discretized version of the continuous-time Langevin equation [1]. Now Hansen *et al.* state that this Langevin-type equation is (in the limit $n \to 0$) completely equivalent to the Fokker-Planck equation (7) but with $\tilde{a}_2(w)$ instead of $a_2(w)$. They suggest

$$\tau \frac{\partial}{\partial t} Q(w,t) \;=\; -\eta \frac{\partial}{\partial w}[a_1(w) Q(w,t)] \;+\; \frac{\eta^2}{2} \frac{\partial^2}{\partial w^2}[\tilde{a}_2(w) Q(w,t)] \,. \tag{10}$$

Also in this case one must be careful, since the relationship between this Fokker-Planck equation and the Langevin-type equation (9) for $n \gg 1$ is not clear. Let us try to formalize the step from the Langevin-type equation (9) to the Fokker-Planck equation (10). First we rewrite (9) as

$$w(t + \tau n' \eta^{-2}) - w(t) \;=\; n' \, \eta^{-1} \, a_1(w) \;+\; \sqrt{n' \tilde{a}_2(w)} \, \xi \,, \tag{11}$$

with $n' \equiv n\eta^2$. We can, by letting $\eta \to 0$, take the limit $n' \to 0$ on the right-hand side. Then we reach the conclusion that the *stationary* solution of Hansen's Fokker-Planck equation correctly

describes the *stationary* solution of the Langevin-type learning process (9) in the limit of small learning parameters η. Note however that, because of the assumptions made in deriving (9), this does not necessarily mean that this stationary solution is the stationary solution of the master equation (5) (see also section 3).

If we also take the limit $\tau \to 0$ on the left-hand side of (11), we can indeed arrive at the Fokker-Planck equation (10). However, since τ is nothing but our definition of time scale (we might have called it 1 from the beginning) this is not a well-defined limit. In section 2.4 we will give a systematic derivation of a (continuous-time) Fokker-Planck approximation of the random-walk equation (4) for small learning parameters η.

2.3 Van Kampen's expansion of the master equation

Intuitively, a realisation of a stochastic process can often be viewed as an average, deterministic trajectory, with stochastic fluctuations around this trajectory. This is the so-called "small-fluctuations Ansatz"

$$w \;=\; \phi(t) \;+\; \sqrt{\eta}\,\xi\,. \tag{12}$$

It says that the time-dependent stochastic variable w is given by a deterministic part $\phi(t)$ (to be determined) plus a term of order $\sqrt{\eta}$ containing the (small) fluctuations. Using Van Kampen's expansion [1] (see also [16, 13]), it is possible to obtain the precise conditions under which this intuitive picture is valid. A quick review of the expansion can be found in the appendix.

The final result of Van Kampen's expansion is a (nonlinear) differential equation for the deterministic part

$$\frac{\tau}{\eta}\frac{d\phi(t)}{dt} \;=\; a_1(\phi(t)) \tag{13}$$

and a linear Fokker-Planck equation for the probability $\Pi(\xi,t)$ of the fluctuations ξ

$$\frac{\tau}{\eta}\frac{\partial\Pi(\xi,t)}{\partial t} \;=\; -a_1^{(1)}(\phi(t))\,\frac{\partial}{\partial\xi}\left[\xi\Pi(\xi,t)\right] \;+\; \frac{1}{2}a_2(\phi(t))\,\frac{\partial^2}{\partial\xi^2}\Pi(\xi,t)\,. \tag{14}$$

This so-called linear noise approximation is only valid as long as the Ansatz (12) is justified. In the appendix it is shown that this restricts its validity to regions of weight space with $a_1^{(1)} < 0$. In regions of weight space with $a_1^{(1)} \geq 0$ it is only valid on time scales $< \mathcal{O}(1/\eta)$ [assuming that we start with a localized distribution, e.g., $P(w,0) = \delta(w - \phi(0))$].

Generalization of these results to N dimensions, i.e., N adaptive elements, is straightforward. The first moment becomes an N-dimensional drift vector, its derivative an $N \times N$-matrix $H(w)$ with components

$$H_{ij}(w) \;\equiv\; -\frac{\partial(a_1(w))_i}{\partial w_j}\,.$$

This "Hessian matrix" $H(w)$ (it is a true Hessian matrix if and only if the drift vector can be written as the gradient of some error potential or energy function, see e.g. [5]) must be positive definite for Van Kampen's expansion to be valid. Each of these so-called attraction regions defined by positive definite Hessian $H(w)$ contains one fixed-point solution of the deterministic equation (13), i.e., a solution ϕ^* with

$$a_1(\phi^*) \;=\; 0 \quad \text{ and positive definite } H(\phi^*)\,.$$

Thus, the small-fluctuations Ansatz (12) is valid inside the attraction regions, i.e., in the vicinity of the fixed-point solutions, but [on time scales $\geq \mathcal{O}(1/\eta)$] not outside of these attraction regions.

Now that we have made a rigorous expansion of the master equation, we can check the validity of Radons' Fokker-Planck approximation (7). If we substitute the small-fluctuations Ansatz (12) into the Fokker-Planck equation (7), then the lowest-order Fokker-Planck equation for ξ is exactly the same as the lowest-order term (14) in the linear noise expansion. In other

words, terms $\geq \mathcal{O}(\eta^3)$ in the Kramers-Moyal expansion (6) do not contribute to the linear noise approximation. In this sense the Fokker-Planck equation (7) is equivalent to Van Kampen's equation (14). However, we have to keep in mind that only the linear noise approximation is strictly valid [16]. In other words, all (nonlinear) features that arise from using the Fokker-Planck equation (7) beyond that approximation are spurious and cannot be taken seriously [15, 17]. Furthermore, it means that the mathematical validity of Radons' Fokker-Planck approach is restricted to relatively short time scales and regions of weight space with positive definite Hessian matrix, in short, restricted to local properties. Yet it is frequently used to study global properties. In section 3 we will discuss its accuracy in these situations.

2.4 Van Kampen's expansion of the random-walk equation

Van Kampen's expansion of the discrete-time random-walk equation is slightly more complicated. In the appendix we derive the linear noise approximation

$$\begin{aligned} \frac{\tau}{\eta}\frac{d\phi(t)}{dt} &= a_1(\phi(t)) \\ \frac{\tau}{\eta}\frac{\partial \Pi(\xi,t)}{\partial t} &= -a_1^{(1)}(\phi(t))\,\frac{\partial}{\partial \xi}\left[\xi \Pi(\xi,t)\right] + \frac{1}{2}\tilde{a}_2(\phi(t))\,\frac{\partial^2}{\partial \xi^2}\Pi(\xi,t)\,. \end{aligned} \tag{15}$$

The only difference with equations (13) and (14) for the linear noise approximation of the continuous-time master equation is the term $\tilde{a}_2(\phi(t))$ instead of $a_2(\phi(t))$. Furthermore, as expected, the result (15) can also be obtained by substitution of the small-fluctuations Ansatz (12) into Hansen's Fokker-Planck equation (10). The conclusion is therefore that Radons' Fokker-Planck approximation of the continuous-time master equation is as accurate as Hansen's Fokker-Planck approximation of the discrete random-walk equation. Both can be used on time scales $< \mathcal{O}(1/\eta)$ and in the so-called attraction regions. However, we should keep in mind that even in these situations only their linear noise approximations are strictly valid.

On time scales $> \mathcal{O}(1/\eta)$, the particular choice of time intervals does not matter anymore and the solutions $P(w,t)$ of the master and random-walk equation become essentially equal [14, 9]. The stationary solutions are the same. This is *not* the case for the two Fokker-Planck approximations! As argued in section 2.2, the stationary solution of Hansen's Fokker-Planck equation (10) becomes exact in the limit of small learning parameters η for all Langevin-type learning processes with additive Gaussian white noise. Radons [9] gives an example of a linear learning rule with a non-Gaussian noise distribution for which the Fokker-Planck equation (7) yields the correct stationary distribution in the limit of small learning parameters. Is there a paradox? No! Inside the attraction regions, the local relaxation time is also of order $1/\eta$ [4]. So when the two solutions $P(w,t)$ approach each other, the deterministic part $\phi(t)$ approaches a fixed-point solution ϕ^* with $a_1(\phi^*) = 0$ and thus $\tilde{a}_2(\phi^*)$ approaches $a_2(\phi^*)$, which makes the two approximations indeed equivalent. Outside of the attraction regions both approximations $Q(w,t)$ become invalid at times of order $1/\eta$, i.e., before the "true" probabilities $P(w,t)$ start to become equivalent.

3 Fokker-Planck approaches and global properties

3.1 Description of simulations

In the previous sections we have shown that the mathematical validity Fokker-Planck approaches suggested in the literature is restricted to local properties of on-line learning processes. If presented as *models* instead of as proper expansions for small learning parameters η, these models might still be useful to study global properties (see e.g. [2, 6, 5, 7, 8] for attempts in this direction). In this section we will investigate how accurate these models can be. Fokker-Planck

approaches are solely based on the first two moments of the transition matrix (3): the drift $a_1(w)$ and the diffusion $a_2(w)$. Therefore, they yield the same predictions for the "original" on-line learning process (1) and the Langevin-type equation

$$\Delta w \;=\; \eta\, a_1(w) \;+\; \eta\, \sqrt{\tilde{a}_2(w)}\, \xi\,, \tag{16}$$

which is (9) with $n = 1$. In our simulations we have Poisson-distributed time intervals for both learning procedures.

We will focus mainly on first passage times from a fixed-point solution ϕ^* of the deterministic equation (13) into some region $\mathcal{I}$ *outside* the attraction region. Mean first passage time typically scale exponentially with the reciprocal value of the learning parameter [18, 5], i.e., are (for small learning parameters η) much larger than the time scale on which the Fokker-Planck approximations can be proven to be valid. Mean first passage times for different values of the learning parameter η are calculated from Monte-Carlo simulations with an ensemble of M independently operating networks. We start with all networks at $w(0) = \phi^*$ and take network i out of the simulation when it reaches region $\mathcal{I}$ for the first time. This first passage time is denoted τ_i. The maximum likelihood value of the mean first passage time $\tau_{\rm mfp}$ is

$$\tau_{\rm mfp} \;=\; \frac{1}{M} \sum_i \tau_i\,,$$

with standard error [19]

$$\Delta\tau_{\rm mfp} \;=\; \frac{\tau_{\rm mfp}}{\sqrt{M}}\,.$$

In the following sections we will show plots of the logarithm of the mean first passage time $\tau_{\rm mfp}$ as a function of the reciprocal value of the learning parameter η. Lines in these plots are least-squares fits of the form

$$\ln \tau_{\rm mfp} \;=\; a \;+\; b\, \ln\left[\frac{1}{\eta}\right] \;+\; \frac{c}{\eta}\,, \tag{17}$$

with c called the reference learning parameter. If the learning parameter η is chosen much smaller than this reference learning parameter, the first passage times get exponentially large. We will encounter mean first passage times on the order of 10^6 learning steps.

3.2 One-dimensional toy problem

The learning rule is the one-dimensional Grossberg learning rule [20]

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

which tends to the average $\langle x \rangle_x$ over all inputs if x is drawn independently from the network state w. However, by choosing the probability to draw a particular input x as a function of the current network state w, i.e., $\rho(x|w)$ instead of $\rho(x)$, various attractive points can be introduced [5]. We choose an underlying probability distribution

$$\rho_0(x) \;=\; \frac{1+\gamma}{2}\,\delta(x-1) \;+\; \frac{1-\gamma}{2}\,\delta(x+1)\,,$$

i.e., there are only two possible inputs; for $\gamma > 0$ the probability to draw $x = 1$ is higher than the probability to draw $x = -1$. Now we apply a Gaussian window such that the probability to receive a particular input is enlarged if the weight is closer to this input:

$$\rho(x|w) \;=\; \frac{1}{Z(w)}\,\rho_0(x)\,\exp\left[-\frac{\beta(x-w)^2}{2}\right]\,,$$

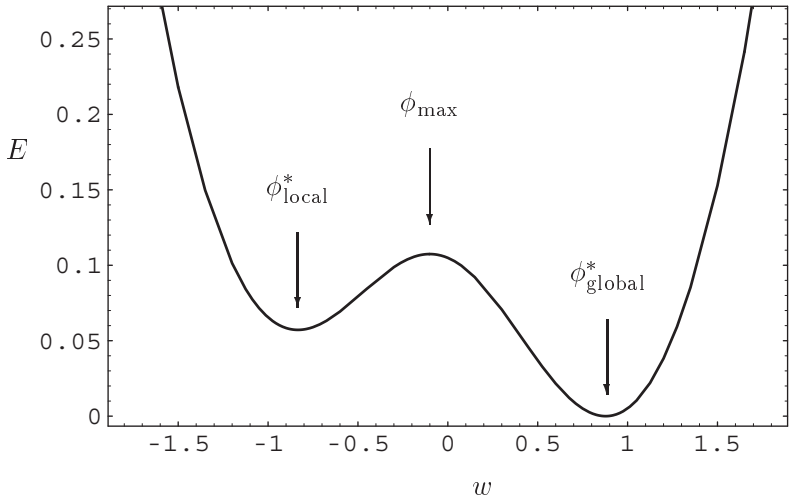

Figure 1: Error potential E of the one-dimensional toy problem as a function of the weight w.

with $Z(w)$ a normalization constant to normalize $\rho(x|w)$. Straightforward calculations yield the "jump moments"

$$a_1(w) \equiv \langle x - w \rangle_x = \int dx\, \rho(x|w)\, f(w,x) = \tanh[\beta w + \epsilon] - w$$
$$a_2(w) \equiv \left\langle (x-w)^2 \right\rangle_x = 1 + w^2 - 2w \tanh[\beta w + \epsilon] = \frac{1}{\cosh^2[\beta w + \epsilon]} + a_1^2(w)\,,$$

with $\epsilon \equiv \operatorname{arctanh} \gamma$. In our simulations we work with $\beta = 1.5$ and $\epsilon = 0.05$. The error potential $E(w)$ defined by

$$-\frac{dE(w)}{dw} = a_1(w) \quad \text{and} \quad E(\phi^*_{\rm global}) = 0\,,$$

is plotted in figure 1.

We collect the first passage times through the local maximum $\phi_{\max}$ of $N = 3000$ networks starting from the local minimum $\phi^*_{\rm local}$ [figure 2(a)] and from the global minimum $\phi^*_{\rm global}$ [figure 2(b)]. Mean first passage times predicted by Radons' Fokker-Planck equation (7) are obtained by numerical integration of [1]

$$\tau_{\rm mfp} = \frac{2}{\eta^2} \int_{\phi^*}^{\phi_{\max}} dw\, \left[a_2(w) Q_{\rm stat}(w)\right]^{-1} \int_{-\infty}^{w} dw'\, Q_{\rm stat}(w')\,, \tag{18}$$

with the stationary solution (8), and similarly for Hansen's suggestion (10). The figures indicate that (18) yields a quite accurate prediction for the Langevin-type equation (16). Since $a_1^2(w) \ll a_2(w)$ for all $\phi^*_{\rm local} \le w \le \phi^*_{\rm global}$, the difference between the two Fokker-Planck approaches is small. However, the mean first passage times for the Langevin-type equation are different from those of the "true" on-line learning process. And, most important of all, the graphs seem to diverge, rather than to converge for small learning parameters η.

In [5] we suggested that one might be able to estimate the slope of these graphs, i.e., the reference learning parameters c. The model we presented is based on the following two assumptions.

1. The shape of the probability distribution inside attraction regions is given by Gaussians that follow from a (local) application of Van Kampen's expansion. The simplification here is that we assume the Gaussian shape in the whole attraction region, not just in a neighborhood of order η of the fixed-point solution.

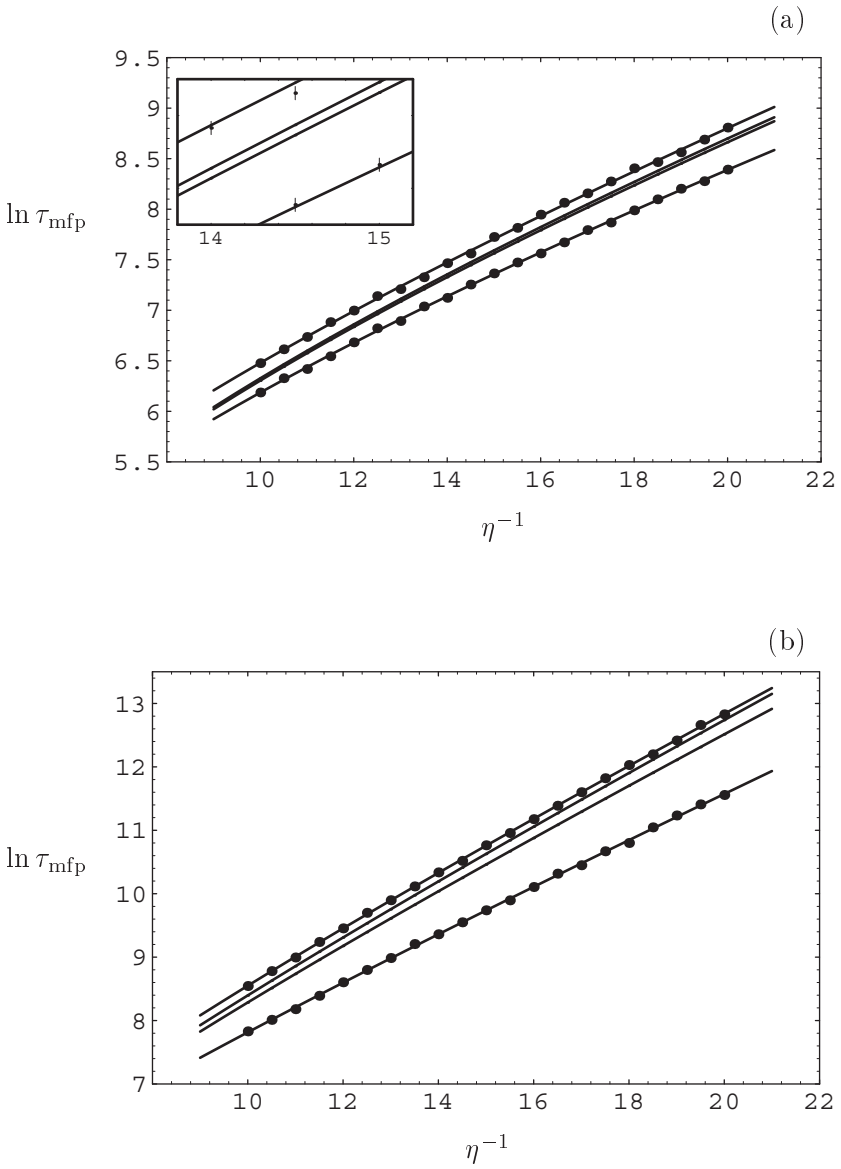

Figure 2: Logarithm of the mean first passage time versus reciprocal value of the learning parameter. Circles represent simulations, lines fits of the form (17). Error bars are on the order of the point sizes (see inset). From top to bottom: Langevin-type learning, Hansen's Fokker-Planck equation, Radons' Fokker-Planck equation, and on-line learning. (a) Starting from the local minimum. Graphs for the two Fokker-Planck equations are almost on top of each other (see inset). (b) Starting from the global minimum.

	local		global	
on-line	0.138		0.326	
Langevin	0.156		0.353	
Radons	0.177	(0.161)	0.351	(0.354)
Hansen	0.180	(0.163)	0.363	(0.365)
Heskes		(0.233)		(0.639)

Table 1: Reference learning parameters c for the mean first passage times starting from the local and the global minimum. Terms in parentheses denote theoretical predictions.

2. The reference learning parameter is hardly affected by what happens outside the attraction regions. Thus it can be calculated by considering the first passage times from the fixed-point solution ϕ^* to the boundary of the attraction region $\phi_{\rm bnd}$.

If these assumptions are valid, then the reference learning parameter c obeys [5]

$$c \;=\; \frac{a_1^{(1)}(\phi^*)\,(\phi_{\rm bnd}-\phi^*)^2}{a_2(\phi^*)}\,. \tag{19}$$

This description is also a Fokker-Planck approach in the sense that it only uses information about the drift and the diffusion. The local Gaussian probabilities only depend on the derivative of the drift and the diffusion at the fixed-point solution. This approach does therefore not take into account the full dependence of the drift and diffusion on the weights, in contrast with the Fokker-Planck approaches of Radons and Hansen. In the limit of high barriers, the reference learning parameter for the Fokker-Planck first mean passage time (18) converges to the Arrhenius factor

$$c \;=\; -2\int_{\phi^*}^{\phi_{\rm max}} dw\;\frac{a_1(w)}{a_2(w)}\,.$$

Table 1 shows that the Arrhenius factors (terms in parentheses) resulting from Radons' and Hansen's Fokker-Planck approaches are far better estimates of the reference learning parameters for on-line learning than the prediction (19). In this case, the full Fokker-Planck equations are therefore better models to predict reference learning parameters than the model presented in [5].

Let us define the stationary occupation numbers

$$n_{\rm local} \;\equiv\; \int_{-\infty}^{\phi_{\rm max}} dw\; P_{\rm stat}(w) \qquad \text{and} \qquad n_{\rm global} = 1 - n_{\rm local}\,.$$

They obey the "detailed-balance" condition

$$Q \;\equiv\; \frac{n_{\rm global}}{n_{\rm local}} \;=\; \frac{\tau_{\rm global}}{\tau_{\rm local}}\,.$$

For small learning parameters η, the stationary probability distribution is sharply peaked in the neighborhood of the minima, and $\tau_{\rm local}$ and $\tau_{\rm global}$ are the mean first passage times through the local maximum starting from the local and the global minimum, respectively. Figure 3 is figure 2(a) substracted from figure 2(b), i.e., shows $\ln Q$ as a function of η^{-1}. The graphs for Langevin-type learning and Hansen's Fokker-Planck equation are on top of each other. This is in perfect agreement with section 2.2 where we derived that the stationary solutions of the Fokker-Planck equation (7) and the Langevin-type learning rule (16) are equivalent for small learning parameters η.

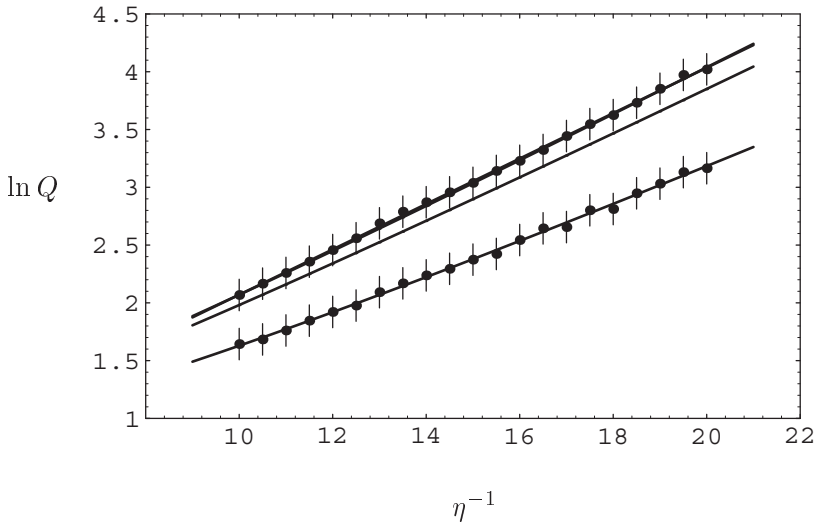

Figure 3: Logarithm of stationary occupation at global minimum divided by stationary occupation at local minimum versus reciprocal value of the learning parameter. Circles represent simulations, lines fits of the form (17). From top to bottom: Langevin-type learning, Hansen's Fokker-Planck equation, Radons' Fokker-Planck equation, and on-line learning. Graphs for Langevin-type learning and Hansen's Fokker-Planck equation are on top of each other.

3.3 Kohonen learning rule

The Kohonen learning rule [10] tries to capture important features of self-organizing processes. Properties of the Kohonen learning procedure have been studied in great detail. In this context, Ritter and Schulten [12] were the first to use a master equation for the description of on-line learning processes.

Here we will study a network with three units, each having one weight. The network state vector is written $w = (w_1, w_2, w_3)^T$. Inputs x are drawn with equal probability from the interval $[0, 1]$:

$$\rho(x) \;=\; \theta(x)\,\theta(1-x)\,.$$

First the "winner" $\kappa(x)$ is determined. It is the unit with weight $w_{\kappa(x)}$ closest to the input x:

$$\Big(w_{\kappa(x)} - x\Big)^2 \;\leq\; (w_i - x)^2 \quad \forall_i\,.$$

The weights are then updated by

$$\Delta w_i \;=\; \eta\; h_{i,\kappa(x)}\,(x - w_i) \quad \text{with} \quad h_{ij} \;=\; \delta_{ij} \;+\; \sigma\delta_{i,j\pm1}\,.$$

So, not only the winner is updated (with strength 1), but also its nearest neighbor(s) (with strength σ). By writing the determination of the winning unit as a product of θ-functions, it is easy to see that the Kohonen learning rule is of the form (1).

A weight vector is called "ordered" if $w_1 \leq w_2 \leq w_3$ or $w_1 \geq w_2 \geq w_3$, and disordered otherwise. For $\sigma = 0.1$, the value that we use in our simulations, there are both ordered and disordered fixed-point solutions ϕ^* of the deterministic equation (13). We start with all 500 networks at the disordered fixed-point solution ϕ^* for which $w_2 < w_1 < w_3$, and take a network out of the simulation if it reaches the region $\mathcal{I}$ with $w_1 < w_2 < w_3$. We perform these simulations for both the original on-line learning rule (1) and the Langevin-type learning rule (16) with the same drift *vector* and diffusion *matrix*. The results are shown in figure 4. Here it is even more clear that the on-line learning rule and the Langevin-type learning rule

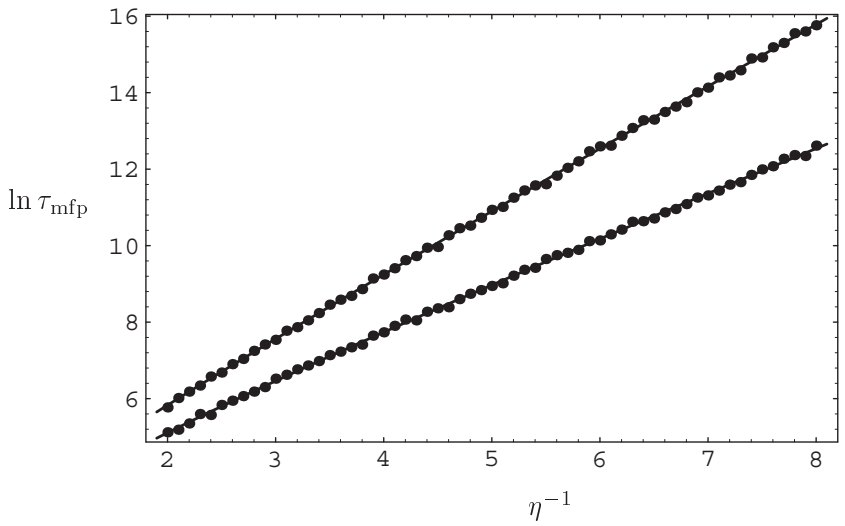

Figure 4: Logarithm of mean first passage times from a disordered fixed-point solution into an ordered region versus reciprocal value of the learning parameter. Circles represent simulations, lines fits of the form (17). Error bars are on the order of the point sizes. On-line learning (upper line) and Langevin-type learning (lower line).

give different results for small learning parameters η. We obtain a reference learning parameter $c = 1.56$ for on-line learning and $c = 1.09$ for Langevin-type learning. Note that, in contrast with the one-dimensional toy problem of section 3.2, in this example the reference learning parameter for on-line learning is the largest one. It does not make sense to look at the mean first passage times for Langevin-type learning as an upper or lower bound for on-line learning: they are just completely different.

3.4 Backpropagation

Backpropagation [11] is a popular supervised learning rule for multi-layered perceptrons. In several papers [2, 6, 7, 8, 21, 22] properties of on-line backpropagation have been studied using Fokker-Planck approaches.

Simulations are performed on the network shown in figure 5(a). Nine adaptive elements are combined in the weight vector $w = (w_{10}, w_{11}, w_{12}, w_{20}, w_{21}, w_{22}, w_{30}, w_{31}, w_{32})^T$. The network has two variable inputs: x_1 and x_2. Thresholds are incorporated by defining $x_0 \equiv y_0 \equiv -1$. Outputs of the hidden units and the output unit are given by

$$y_i \;=\; \tanh\left[\sum_{j=0}^{2} w_{ij} x_j\right] \quad \text{and} \quad y_3 \;=\; \tanh\left[\sum_{j=0}^{2} w_{ij} y_j\right] ,$$

respectively. Training patterns are three-dimensional vectors $x^\mu = (x_1^\mu, x_2^\mu, x_3^\mu)^T$. The components x_1^μ and x_2^μ give the input values of the network for pattern μ, the component x_3^μ the desired output value. At each learning step one of the patterns, say μ, is drawn at random from the training set. The on-line learning rule for this pattern x^μ then follows the gradient of the error

$$E(w, x^\mu) \;\equiv\; \frac{1}{2}\left[y_3(w, x_1^\mu, x_2^\mu) - x_3^\mu\right]^2 \;+\; \frac{\lambda}{4}\sum_{i=0}^{2}\sum_{j=0}^{2}\left[w_{ij}^2 - \alpha\right]^2 ,$$

with $\alpha = 0.1$ and $\lambda = 0.01$. Incorporation of the second term, the so-called bias, has a few advantages among which there are prevention of local minima with infinite weights and reduction of training times [23, 24].

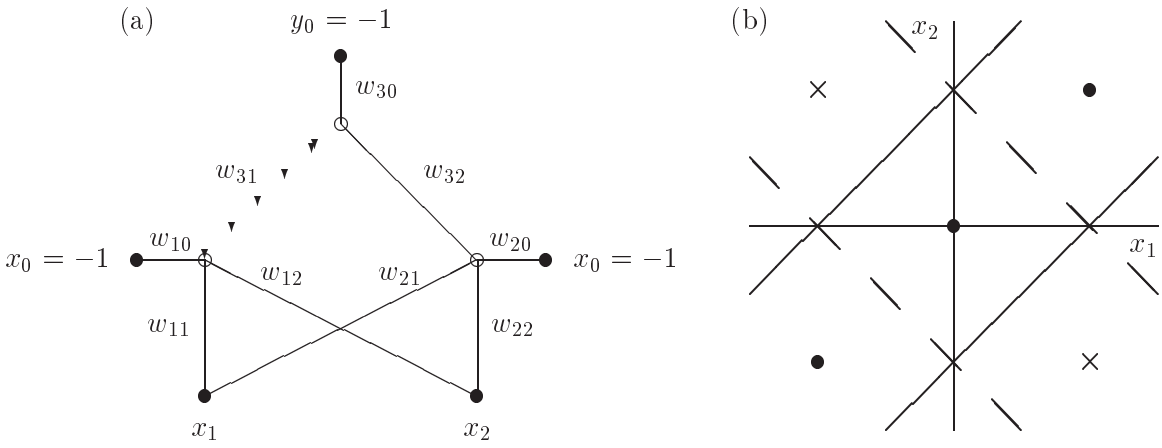

Figure 5: (a) Network structure. (b) XOR problem with one additional pattern.

Following reference [25], we choose the set of five training patterns sketched in figure 5(b). Circles indicate negative desired output $x_3^\mu = -0.8$, crosses positive output $x_3^\mu = 0.8$. It is the usual XOR truth table with an additional pattern at the origin. Now the total error potential ($E(w, x^\mu)$ averaged over all five patterns) has not only global minima, but also deep local minima. The (thick) solid lines in figure 5(b) show the separation lines of the hidden units that lead to the optimal solution (all five patterns correctly classified), the dashed lines those corresponding to the local minima (one pattern misclassified). For symmetry reasons, there are 8 local and 8 global minima.

All 500 networks start at a local minimum where the pattern $x = (0, 0, -0.8)^T$ is misclassified. First passage times into a region $\mathcal{I}$ where all five patterns are correctly classified are collected for both on-line learning and Langevin-type learning. The results are shown in figure 6. Again it is evident that Langevin-type learning yields very different mean first passage times than on-line learning, *especially* for small learning parameters η. Here we find reference learning parameters $c = 1.45$ for Langevin-type learning and $c = 2.22$ for on-line learning.

4 Two conclusions

The Fokker-Planck approaches suggested by Radons and Hansen are equally valid: Radons' Fokker-Planck equation is a *locally* valid approximation of the continuous-time master equation, Hansen's Fokker-Planck equation is a *locally* valid approximation of the discrete-time random-walk equation. Drift and diffusion, the only two moments that are taken into account by a Fokker-Planck approach, are not sufficient for a precise calculation of *global* properties of on-line learning processes.

Acknowledgments

I would like to thank Andreas Herz and Christian Kurrer for many useful comments on an earlier version of this manuscript. This work was supported by a grant from the National Institutes of Health (P41RR05969) to Klaus Schulten and (in the final stage) by the Dutch Foundation for Neural Networks (SNN).

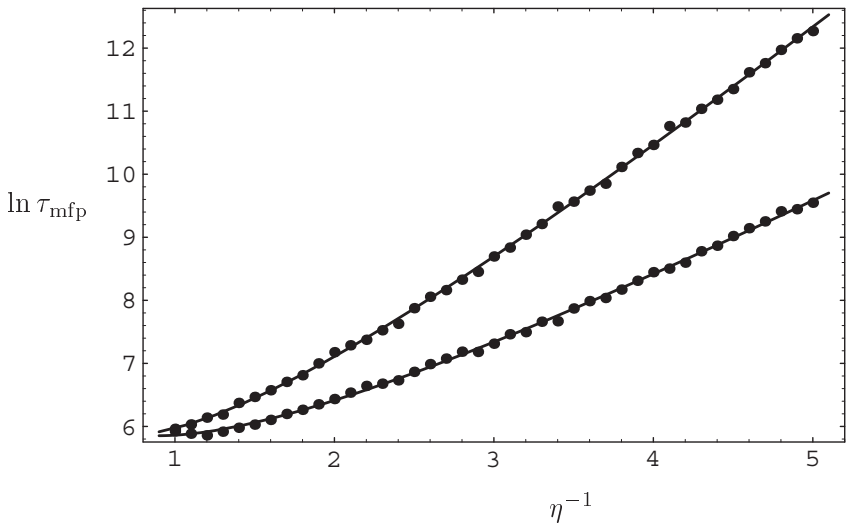

Figure 6: Logarithm of mean first passage times from a local minimum into a region with all five patterns correctly classified versus reciprocal value of the learning parameter. Circles represent simulations, lines fits of the form (17). Error bars are on the order of the point sizes. On-line learning (upper line) and Langevin-type learning (lower line).

Appendix

First, we will give a quick review of Van Kampen's expansion of the continuous-time master equation (5).

1. We start with the "small-fluctuations Ansatz" (12) and define the function $\Pi(\xi,t)$ as the probability $P(w,t)$ in terms of the new variable ξ:

$$\Pi(\xi,t) \;\equiv\; P(\phi(t)+\sqrt{\eta}\xi,t)\,.$$

2. The time derivative of the $\Pi(\xi,t)$ consists of two parts:

$$\frac{\partial\Pi(\xi,t)}{\partial t} \;=\; \frac{\partial P(w,t)}{\partial w}\,\frac{d\phi(t)}{dt} \;+\; \frac{\partial P(w,t)}{\partial t} \;=\; \frac{1}{\sqrt{\eta}}\,\frac{d\phi(t)}{dt}\,\frac{\partial\Pi(\xi,t)}{\partial\xi} \;+\; \frac{\partial P(w,t)}{\partial t}\,.$$

3. We rewrite the Kramers-Moyal expansion (6) in terms of $\Pi(\xi,t)$ and obtain

$$\tau\frac{\partial\Pi(\xi,t)}{\partial t} \;=\; \frac{\tau}{\sqrt{\eta}}\frac{d\phi(t)}{dt}\frac{\partial\Pi(\xi,t)}{\partial\xi} \;+\; \sum_{n=1}^{\infty}\frac{(-1)^n\,\eta^{n/2}}{n!}\frac{\partial^n}{\partial\xi^n}\left[a_n(\phi(t)+\sqrt{\eta}\xi)\,\Pi(\xi,t)\right]\,.$$

4. We choose the function $\phi(t)$ such that the lowest order terms in η on the right-hand side cancel and obtain the deterministic equation

$$\frac{\tau}{\eta}\frac{d\phi(t)}{dt} \;=\; a_1(\phi(t))\,.$$

5. We make a Taylor expansion of $a_n(\phi(t)+\sqrt{\eta}\xi)$ in powers of $\sqrt{\eta}$. After some rearrangements we obtain

$$\frac{\tau}{\eta}\frac{\partial\Pi(\xi,t)}{\partial t} \;=\; \sum_{m=2}^{\infty}\sum_{n=1}^{m}\frac{(-)^n\,\eta^{(m-2)/2}}{n!\,(m-n)!}\,a_n^{(m-n)}(\phi(t))\frac{\partial^n}{\partial\xi^n}\left[\xi^{m-n}\,\Pi(\xi,t)\right]\,,$$

where $a_n^{(l)}(\phi)$ stands for the l-th derivative of $a_n(\phi)$ with respect to the argument ϕ.

6. In the limit $\eta \to 0$ only the term $m = 2$ remains on the right-hand side. This is called the linear noise approximation. The remaining differential equation for $\Pi(\xi, t)$ is the Fokker-Planck equation

$$\frac{\tau}{\eta}\frac{\partial \Pi(\xi,t)}{\partial t} \;=\; -a_1^{(1)}(\phi(t))\,\frac{\partial}{\partial \xi}\left[\xi \Pi(\xi,t)\right] \;+\; \frac{1}{2}a_2(\phi(t))\,\frac{\partial^2}{\partial \xi^2}\Pi(\xi,t)\,. \tag{A.1}$$

7. From (A.1) we can calculate the dynamics of the average of the fluctuations $\langle \xi \rangle_t$ and of the square of the fluctuations $\langle \xi^2 \rangle_t$:

$$\begin{aligned}\frac{\tau}{\eta}\frac{\partial \langle \xi \rangle_t}{\partial t} &= a_1^{(1)}(\phi(t))\;\langle \xi \rangle_t \\ \frac{\tau}{\eta}\frac{\partial \langle \xi^2 \rangle_t}{\partial t} &= 2a_1^{(1)}(\phi(t))\;\left\langle \xi^2 \right\rangle_t \;+\; a_2(\phi(t))\,.\end{aligned} \tag{A.2}$$

8. We started with the Ansatz that ξ is of order 1. From equation (A.2) we conclude that the final result is consistent with the Ansatz, if both evolution equations converge, i.e., if

$$a_1^{(1)}(\phi(t)) \;<\; 0\,.$$

Next we will make a similar expansion of the discrete-time random-walk equation (4). The subsequent steps in this derivation can be compared with the corresponding steps above.

1. Again we start with the "small-fluctuations Ansatz" (12) and define the function $\Pi(\xi, t)$ as the probability $P(w, t)$ in terms of the new variable ξ.

2. This step is more complicated for a difference equation than for a differential equation. We have to make a Taylor expansion:

$$\Pi(\xi,t+\tau) \;-\; \Pi(\xi,t) \;=\; \sum_{l=0}^{\infty}\frac{\left[\phi(t+\tau)-\phi(t)\right]^l}{l!}\,\frac{\partial^l P(w,t+\tau)}{\partial w^l} \;-\; P(w,t) \;=$$
$$\sum_{l=1}^{\infty}\frac{\left[\phi(t+\tau)-\phi(t)\right]^l}{l!}\frac{\partial^l P(w,t)}{\partial w^l} \;+\; \sum_{l=0}^{\infty}\frac{\left[\phi(t+\tau)-\phi(t)\right]^l}{l!}\,\frac{\partial^l}{\partial w^l}\left[P(w,t+\tau)-P(w,t)\right]\,.$$

3. We replace the term $P(w, t+\tau) - P(w, t)$ by the (same) Kramers-Moyal expansion (6) in terms of $\Pi(\xi, t)$.

4. The deterministic equation is the nonlinear difference equation

$$\frac{1}{\eta}\left[\phi(t+\tau)-\phi(t)\right] \;=\; a_1(\phi(t))\,. \tag{A.3}$$

5. After making the Taylor expansion and some more rearrangements we obtain

$$\frac{1}{\eta}\left[\Pi(\xi,t+\tau)-\Pi(\xi,t)\right] \;=\; -\sum_{l=2}^{\infty}\frac{(l-1)\,a_1^l(\phi(t))\,\eta^{(l-2)/2}}{l!}\,\frac{\partial^l \Pi(\xi,t)}{\partial \xi^l} \;+$$
$$\sum_{l=0}^{\infty}\sum_{m=2}^{\infty}\sum_{n=1}^{m}\frac{(-1)^n\,a_1^l(\phi(t))\,\eta^{(l+m-2)/2}}{l!\,n!\,(m-n)!}\,a_n^{(m-n)}(\phi(t))\,\frac{\partial^n}{\partial \xi^n}\left[\xi^{m-n}\,\Pi(\xi,t)\right]\,.$$

6. In the limit $\eta \to 0$ only the term $l = 2$ remains in the first sum and the term $l = 0$ and $m = 2$ in the second sum. So, finally we arrive at the Fokker-Planck equation

$$\frac{1}{\eta}[\Pi(\xi,t+\tau)-\Pi(\xi,t)] = -a_1^{(1)}(\phi(t))\frac{\partial}{\partial \xi}\left[\xi\Pi(\xi,t)\right]+\frac{1}{2}[a_2(\phi(t))-a_1^2(\phi(t))]\frac{\partial^2}{\partial \xi^2}\Pi(\xi,t). \tag{A.4}$$

7. The evolution equations are now

$$\begin{aligned} \frac{1}{\eta}\left[\langle\xi\rangle_{t+\tau} - \langle\xi\rangle_t\right] &= a_1^{(1)}(\phi(t))\ \langle\xi\rangle_t \\ \frac{1}{\eta}\left[\left\langle\xi^2\right\rangle_{t+\tau} - \left\langle\xi^2\right\rangle_t\right] &= 2a_1^{(1)}(\phi(t))\ \left\langle\xi^2\right\rangle_t + \tilde{a}_2(\phi(t))\,. \end{aligned} \tag{A.5}$$

8. The validity of the expansion is again restricted to local properties.

By considering the limit of small learning parameters η, we can now transform the difference equations (A.3) and (A.4) into differential equations. To see this, let us compare the difference equation (A.3) and the differential equation

$$\frac{\tau}{\eta}\frac{d\tilde{\phi}}{dt} = a_1(\tilde{\phi}(t))\,. \tag{A.6}$$

From this differential equation we obtain the difference

$$\tilde{\phi}(t+\tau) - \tilde{\phi}(t) = \sum_{n=1}^{\infty}\frac{\tau^n}{n!}\frac{d^n\tilde{\phi}(t)}{dt^n} \equiv \sum_{n=1}^{\infty}\frac{\eta^n}{n!}b_n(\tilde{\phi}(t))\,, \tag{A.7}$$

with the functions $b_n(\phi)$ obeying the recurrence equation

$$b_{n+1}(\phi) = a_1(\phi)\,b_n^{(1)}(\phi) \quad \text{and} \quad b_1(\phi) = a_1(\phi)\,.$$

Since all $b_n(\phi)$ are independent of η, expression (A.7) is a proper expansion in the learning parameter η, and can be written

$$\frac{1}{\eta}[\tilde{\phi}(t+\tau) - \tilde{\phi}(t)] = a_1(\tilde{\phi}(t)) + \mathcal{O}(\eta)\,.$$

So, up to the order that is taken into account by the linear noise expansion anyway, the solution $\tilde{\phi}(t)$ of the differential equation (A.6) is equivalent to the solution $\phi(t)$ of the difference equation (A.3), provided, of course, that we start with $\tilde{\phi}(0) = \phi(0)$. The same procedure also applies to (A.4) for the probability $\Pi(\xi,t)$ and to (A.5) for the moments $\langle\xi\rangle$ and $\langle\xi^2\rangle$. This finally leads to the set of equations (15).

References

[1] N. van Kampen. *Stochastic Processes in Physics and Chemistry.* North-Holland, Amsterdam, 1992.

[2] G. Radons, H. Schuster, and D. Werner. Fokker-Planck description of learning in backpropagation networks. In *International Neural Network Conference 90 Paris*, pages 993–996, Dordrecht, 1990. Kluwer Academic.

[3] L. Hansen, R. Pathria, and P. Salamon. Stochastic dynamics of supervised learning. *Journal of Physics A*, 26:63–71, 1993.

[4] T. Heskes and B. Kappen. Learning processes in neural networks. *Physical Review A*, 44:2718–2726, 1991.

[5] T. Heskes, E. Slijpen, and B. Kappen. Learning in neural networks with local minima. *Physical Review A*, 46:5221–5231, 1992.

[6] W. Finnoff. Diffusion approximations for the constant learning rate backpropagation algorithm and resistance to local minima. In S. Hanson, J. Cowan, and L. Giles, editors, *Advances in Neural Information Processing Systems 5*, pages 459–466, San Mateo, 1993. Morgan Kaufmann.

[7] T. Leen and J. Moody. Weight space probability densities in stochastic learning: I. Dynamics and equilibria. In S. Hanson, J. Cowan, and L. Giles, editors, *Advances in Neural Information Processing Systems 5*, pages 451–458, San Mateo, 1993. Morgan Kaufmann.

[8] G. Orr and T. Leen. Weight space probability densities in stochastic learning: II. Transients and basin hopping times. In S. Hanson, J. Cowan, and L. Giles, editors, *Advances in Neural Information Processing Systems 5*, pages 507–514, San Mateo, 1993. Morgan Kaufmann.

[9] G. Radons. On stochastic dynamics of supervised learning. *Journal of Physics A*, 26:3455–3461, 1993.

[10] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69, 1982.

[11] D. Rumelhart, G. Hinton, and R. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.

[12] H. Ritter and K. Schulten. Convergence properties of Kohonen's topology conserving maps: fluctuations, stability, and dimension selection. *Biological Cybernetics*, 60:59–71, 1988.

[13] T. Heskes and B. Kappen. On-line learning processes in artificial neural networks. In J. Taylor, editor, *Mathematical Foundations of Neural Networks*, pages 199–233. Elsevier, Amsterdam, 1993.

[14] D. Bedeaux, K. Lakatos-Lindenberg, and K. Shuler. On the relation between master equations and random walks and their solutions. *Journal of Mathematical Physics*, 12:2116–2123, 1971.

[15] N. van Kampen. *Stochastic Processes in Physics and Chemistry*. North-Holland, Amsterdam, 1981.

[16] C. Gardiner. *Handbook of Stochastic Methods*. Springer, Berlin, second edition, 1985.

[17] N. van Kampen. The validity of nonlinear Langevin equations. *Journal of Statistical Physics*, 25:431–442, 1981.

[18] H. Kushner. Robustness and approximation of escape times and large deviations estimates for systems with small noise effects. *SIAM Journal of Applied Mathematics*, 44:160–182, 1984.

[19] J. Mathews and R. Walker. *Mathematical Methods of Physics*. Addison-Wesley, Redwood City, 1970.

[20] S. Grossberg. On learning and energy-entropy dependence in recurrent and nonrecurrent signed networks. *Journal of Statistical Physics*, 48:105–132, 1969.

[21] T. Heskes. Stochastics of on-line backpropagation. In *Proceedings of the European Symposium on Artificial Neural Networks '94*, pages 223–228, 1994.

[22] W. Wiegerinck, A. Komoda, and T. Heskes. Stochastic dynamics of learning with momentum in neural networks. *Journal of Physics A*, 27:4425–4437, 1994.

[23] S. Hanson and L. Pratt. A comparison of different biases for minimal network construction with back-propagation. In D. Touretzky, editor, *Advances in Neural Information Processing Systems 1*, pages 177–185. Morgan Kaufmann, 1989.

[24] A. Kramer and A. Sangiovanni-Vincentelli. Efficient parallel learning algorithms for neural networks. In D. Touretzky, editor, *Advances in Neural Information Processing Systems 1*, pages 40–48. Morgan Kaufmann, 1989.

[25] M. Gori and A. Tesi. On the problem of local minima in backpropagation. *IEEE Transactions on PAMI*, 14:76–86, 1992.