On stochastic dynamics of supervised learning

1993 J. Phys. A: Math. Gen. 26 3455

(http://iopscience.iop.org/0305-4470/26/14/012)

J. Phys. A: Math. Gen. 26 (1993) 3455–3461. Printed in the UK

On stochastic dynamics of supervised learning

Günter Radons
Institut für Theoretische Physik, Universität Kiel, D-2300 Kiel, Federal Republic of Germany

Received 16 March 1993

Abstract. Recently Hansen *et al* derived a Fokker–Planck equation (FPE) for the learning dynamics of neural networks, which differs from a previously given version by Radons *et al*. It is shown that the discrepancies are due to different implicit assumptions for the distribution of time intervals between the discrete learning events. Both approximations are therefore equally justified from a general point of view. The long-time properties, however, are independent of this distribution and are in general more accurately described in the original FPE of Radons *et al*. Especially, mean and variance of the synaptic parameter distributions are exact only in the latter approach.

1. Introduction

In a recent paper [1] Hansen *et al* derived a Fokker–Planck equation (FPE) for the dynamics of learning in neural networks, which in some respects differs from the one given previously by Radons *et al* [2]. The purpose of this contribution is a thorough comparison of the two proposed equations with the goal of clarifying the origin of the differences. It will turn out that both approaches are correct and which one is appropriate is dependent on application.

The main part of this paper is organized as follows. In section 2, we rederive our original FPE by invoking the general connection between discrete random walks and continuous time evolution equations of Bedeaux *et al* [3]. This is an alternative to the systematic approach of Hansen *et al*, and provides us, e.g., with an understanding of the *physical origin* of the 'spurious dependence on initial conditions' observed in [1] for an exactly solvable model. This example is treated in detail in section 3, where we also show that the different stationary solutions of both approximations can both be exact for small learning rates, though for different noise distributions.

2. Learning as random-walk and continuous-time approximations

Learning in neural networks leads to the following type of dynamical system:

$$w_{n+1} = w_n + \eta F(w_n, \xi_n) \tag{1}$$

where w_n denotes the vector of synaptic parameters at the nth update, ξ_n is the pattern vector presented at that instant and η is the learning rate. For random and uncorrelated pattern presentation (1) describes a random walk in the parameter space of neural networks with a given structure. The probability density for synaptic parameters obeys [2]

$$P_{n+1}(w) = \int T(w \mid w') P_n(w') \, \mathrm{d}w' \tag{2}$$

0305-4470/93/143455+07$07.50

with

$$T(w \mid w') = \int \delta(w - w' - \eta F(w', \xi))\rho(\xi)\,\mathrm{d}\xi$$
$$= \langle \delta(w - w' - \eta F(w', \xi)) \rangle_\xi \tag{3}$$

where $\rho(\xi)$ is the probability density for patterns ξ. Equation (2) can be rewritten in the form of a discrete 'time' master equation

$$P_{n+1}(w) - P_n(w) = \int [T(w \mid w')P_n(w') - T(w' \mid w)P_n(w)]\,\mathrm{d}w' . \tag{4}$$

Eventually one wants to replace differences in (4) by differentials to obtain a continuous time description. A consistent way for this has been achieved by Bedeaux *et al* [3]: for that purpose one has to specify how the discrete learning events occur in continuous time, i.e. one has to assign times t_n to nth parameter update. There is some freedom in this choice. If the time differences $\Delta t_n = t_{n+1} - t_n$ are drawn randomly from the distribution

$$\psi(\Delta t) = \tau_1^{-1} \exp(-\Delta t/\tau_1) \tag{5}$$

the random walk (1) is *exactly* described by the continuous time master equation

$$\frac{\partial P(w,t)}{\partial t} = \int [W(w \mid w')P(w',t) - W(w' \mid w)P(w,t)]\,\mathrm{d}w' \tag{6}$$

with transition rate

$$W(w \mid w') = \tau_1^{-1} T(w \mid w') . \tag{7}$$

τ_1 is the mean time difference between subsequent learning steps. In the following we set $\tau_1 = 1$ which means that we measure time t in units of τ_1. The probability distribution $P(w,t)$ in (6) is now defined for all times t by $P(w,t) = \sum_{n=0}^{\infty} \Phi(n,t)P(w,n)$ where $\Phi(n,t)$ is the probability that exactly n learning steps have occured at time t. For the exponential density (5) one has $\Phi(n,t) = (1/n!)(t/\tau_1)^n \exp(-t/\tau_1)$, and (6) with (7) follow from (4) and $\tau_1(\partial/\partial t)\Phi(n,t) = (1 - \delta_{n,0}) \cdot \Phi(n-1,t) - \Phi(n,t)$. For probability densities $\psi(\Delta t)$ differing from the Poisson law (5), equation (6) becomes exact at long times. Especially, the stationary distribution $P(w)$ of (2) and (6) are identical and independent of the law ψ. In the context of learning in neural networks the above approach was used recently by Heskes and Kappen in [4].

In a further step we exploit the fact that the transition rate W depends on the learning rate η which may be assumed to be small ($\eta \ll 1$). Thus one can write W as a power series in η [2]. This series, although similar to the Kramers–Moyal expansion [5], differs from the latter in an important aspect: it is a systematic expansion with respect to a small parameter, and it is therefore allowed to truncate this series. Neglecting terms of order η^3 one obtains the Fokker–Planck equation of Radons *et al* [2,6]

$$\frac{\partial P(w,t)}{\partial t} = -\eta \sum_l \frac{\partial}{\partial w_l}[F_l(w)P(w,t)] + \frac{\eta^2}{2} \sum_{kl} \frac{\partial^2}{\partial w_k \partial w_l}[D_{kl}(w)P(w,t)] \tag{8}$$

with

$$F_l(w) = \langle F_l(\xi, w) \rangle_\xi$$
$$D_{kl}(w) = \langle F_k(\xi, w)F_l(\xi, w) \rangle_\xi . \tag{9}$$

In contrast, Hansen *et al* obtain the same FPE with $D_{kl}(w)$ replaced by

$$\tilde{D}_{kl}(w) = \langle (F_k(\xi, w) - \langle F_k(\xi, w)\rangle_\xi)(F_l(\xi, w) - \langle F_l(\xi, w)\rangle_\xi)\rangle_\xi \,. \tag{10}$$

It will be shown below that mean value and fluctuations of the variable w are treated correctly in our FPE. This and the above derivation show that in general there is no need for introducing a mesoscopic time scale τ as in [1]. It appears that the averaging over τ is only necessary for $\psi(\Delta t) = \delta(\Delta t - \tau_1)$ since otherwise one would not get a smooth, differentiable time evolution in this case.

3. Exact results

With the following exactly solvable example, also treated in [1], one can gain much insight into the physical implications of both proposed approaches and compare the quality of the corresponding approximations. This one-dimensional example is defined by

$$F(w, \xi) = -w + \xi \tag{11}$$

which turns (1) into a simple linear iterated map with additive noise. Let us first consider properties of the problem which are independent of the assignment of the times t_n. This are the stationary solution and related quantities.

The exact asymptotic solution $P(w) = \lim_{n\to\infty} P_n(w)$ of (2) with F as in (11) obeys the integral equation

$$P(w) = \int \rho(w/\eta - w'(1-\eta)/\eta) P(w')\, \mathrm{d}w'/\eta \,. \tag{12}$$

In terms of characteristic functions $\hat{p}(k) = \int \mathrm{e}^{\mathrm{i}kw} P(w)\, \mathrm{d}w$ and $\hat{\rho}(k)$ one gets

$$\hat{p}(k) = \hat{\rho}(\eta k) \cdot \hat{p}((1-\eta)k) \tag{13}$$

which is solved for $|\, 1 - \eta \,| < 1$ by

$$\hat{p}(k) = \prod_{l=0}^{\infty} \hat{\rho}(\eta(1-\eta)^l k) \,. \tag{14}$$

Hansen *et al* treated the case of a Gaussian density $\rho(\xi)$ with variance σ which, e.g. via (14) and Fourier transformation, results also in a Gaussian $P(w)$ with variance Σ. The quantities Σ and σ are related by

$$\Sigma^2 = \frac{\eta}{2-\eta}\, \sigma^2 \,. \tag{15}$$

The stationary solution of the Fokker–Planck approximation (8) is obtained in one dimension as $P(w) = N \cdot D^{-1}(w) \exp(2\int^w F(w')/D(w')\, \mathrm{d}w'/\eta)$. For the example (11), $F(w) = -w$ (for $\langle \xi \rangle = 0$) and $D(w) = \sigma^2$ in the version of Hansen *et al* [1]. Thus their FPE yields also a Gaussian for $P(w)$ with variance $\eta\sigma^2/2$, which means that their approximation becomes asymptotically exact in the limit $\eta \to 0$. In contrast the FPE of Radons *et al* leads with $D(w) = w^2 + \sigma^2$ to

$$P(w) = N \cdot (w^2 + \sigma^2)^{-(1/\eta)-1} \tag{16}$$

where N is a normalization constant.

There are several remarks to be made. Firstly, the FPE of Hansen *et al* always leads to a Gaussian $P(w)$ independent of the noise distribution $\rho(\xi)$. Thus their stationary solution is approximate for all non-Gaussian $\rho(\xi)$. This is important for learning in neural networks, since there the noise due to the random pattern presentation is typically non-Gaussian but rather bounded and discrete. The discreteness of $\rho(\xi)$ may even lead to very irregular, multifractal equilibrium distributions as shown in [6].

Secondly, the variances Σ of $P(w)$ and σ of $\rho(\xi)$ are always exactly related as in (15), independently of the choice of $\rho(\xi)$. This may be seen by differentiating (13) twice with respect to k. Since $\hat{p}$ and $\hat{\rho}$ are moment generating functions [5], the result for $k = 0$ relates the moments of P and ρ as in (15). Now it is important to note that $P(w)$ of (16) fulfills (15) exactly for all η where a stationary solution exists (i.e. $0 < \eta < 2$). This follows from a direct calculation or from (21) below. As a consequence the divergence of Σ for $\eta \to 2$ is also correctly taken into account. This is the most one can expect from a Fokker–Planck approximation for a noise distribution not further specified with prescribed variance σ, since the tails of $P(w)$ and the higher moments depend on the explicit form of $\rho(\xi)$. It follows that in the allowed range of learning rates and for most $\rho(\xi)$ the stationary solution of the FPE of Radons *et al* approximates to the exact solution better than the one of Hansen *et al*. For the quality of our FPE in the non-linear case see [6].

Thirdly, there is also a noise distribution $\rho(\xi)$ where the stationary solution (16) becomes asymptotically exact for $\eta \to 0$. This may be seen as follows. Solving (13) for $\hat{\rho}(k)$ and inserting the Fourier transform [7] $\hat{p}(k) = \text{const} \cdot |k|^{1/2+1/\eta} \cdot K_{1/2+1/\eta}(\sigma|k|)$ of $P(w)$ from (16) yields

$$\hat{\rho}_\eta(k) = \frac{\hat{p}(k/\eta)}{\hat{p}((1-\eta)k/\eta)} = (1-\eta)^{-1/2-1/\eta} \frac{K_{1/2+1/\eta}(\sigma|k|/\eta)}{K_{1/2+1/\eta}(\sigma(1-\eta)|k|/\eta)} \tag{17}$$

where $K_\nu(x)$ is the modified Bessel function of the third kind [8]. Expression (17) is the Fourier transform of a probability density if by Bochner's theorem [9] the right-hand side is positive-definite. Numerical Fourier inversion of (17) indicates that for $0 < \eta < 2$ $\hat{\rho}_\eta(k)$ is indeed the characteristic function of a density $\rho_\eta(\xi)$ [10]. Thus there exists for every density $P(w)$ of the form (16) a noise distribution $\rho_\eta(\xi)$ which exactly generates $P(w)$ as an invariant density of (2) with $F(\xi, w)$ as in (11). $\rho_\eta(\xi)$ still depends on the learning rate η. For $\eta \to 0$, however, $\rho_\eta(\xi)$ becomes asymptotically independent of η. This follows from the asymptotic expansion of $K_\nu(x)$ for large orders and large arguments [8]. One obtains

$$\lim_{\eta \to 0} \hat{\rho}_\eta(k) = \hat{\rho}(k) = \exp(1 - \sqrt{1 + \sigma^2 k^2}) \tag{18}$$

which is the characteristic function of the probability density [7]

$$\rho(\xi) = \frac{e}{\pi\sigma} (1 + \xi^2/\sigma^2)^{-1/2} \, K_1[(1 + \xi^2/\sigma^2)^{1/2}]. \tag{19}$$

This density decays exponentially as $\exp[-|\xi|/\sigma + O(\ln(\xi))]$ for large ξ and plays the same role for $P(w)$ of (16) as the Gaussian density does for the Gaussian law $P(w)$ in the approximation of Hansen *et al*. For comparison both distributions $\rho(\xi)$ are depicted in figure 1. We see that the stationary distributions of both approaches can become exact in

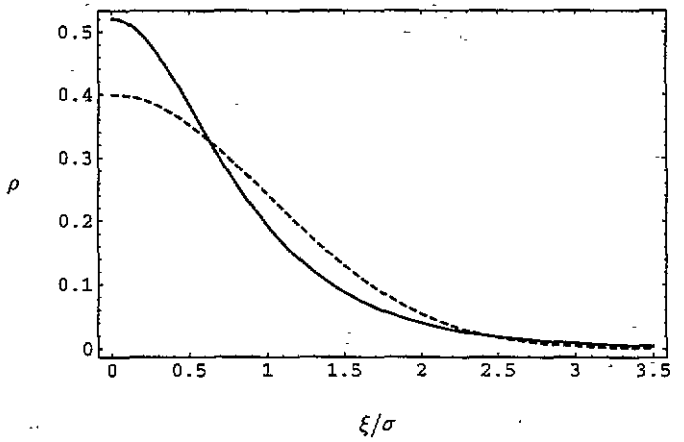

Figure 1. The noise distribution $\rho(\xi)$, equation (19), for which the stationary solution (16) becomes exact at small learning rates is depicted (full curve). For comparison the broken curve is the Gaussian density $\rho(\xi)$ which results in the stationary solution of the FPE of Hansen *et al* for $\eta \ll 1$.

the limit $\eta \to 0$, and therefore from this point of view none of the two approximations is preferable.

Now we turn to the dynamical evolution of mean $w(t) = \langle w \rangle_t = \int w P(w,t)\,dw$ and variance $\Sigma(t) = \langle w^2 \rangle_t - \langle w \rangle_t^2$ of the synaptic parameters. From the FPE (8) with (9) one obtains for the exactly solvable example (11)

$$w(t) = w(0)\,\mathrm{e}^{-\eta t} \tag{20}$$

and

$$\Sigma(t) = \frac{\sigma^2 \eta}{2-\eta}(1 - \mathrm{e}^{-\eta(2-\eta)t}) + w^2(0)(\mathrm{e}^{-\eta(2-\eta)t} - \mathrm{e}^{-2\eta t}). \tag{21}$$

Note that in [1] the prefactor $\sigma^2\eta/(2-\eta)$ is incorrectly stated as $\sigma^2\eta/2$ (in their equation (29)). The correct version may also be found in [4], who treat the same example under the name Grossberg learning. In contrast, the FPE of Hansen *et al* yields a different result for the variance

$$\Sigma_{\mathrm{H}}(t) = \frac{\sigma^2 \eta}{2}(1 - \mathrm{e}^{-2\eta t}) \tag{22}$$

where η in this notation is not absorbed in the time scale.

First, we can see from (21) that for $t \to \infty$, $\Sigma(t)$ approaches the exact limit Σ of (15). This happens by no means accidentally but is a consequence of the fact that (21) is *exact* for all times t and for all η! This is true since (20) and (21) can also be derived from the continuous-time master equation (6). Moreover, this coincidence is not restricted to the special example above but also true for the general multidimensional, non-linear case: with respect to mean and variance our FPE (8) with (9) is not an approximation to the continuous-time master equation (6) but equivalent. This was already correctly stated in [4].

At this stage it is appropriate to remember that in deriving (6) one utilizes the freedom of regarding the update events as being generated by a Poisson process in continuous time. Again, for different choices of the interval distribution such as $\psi(\Delta t) = \delta(\Delta t - \tau_1)$, which means equal time intervals τ_1 between updates, the variance $\Sigma(t)$ of (21) becomes exact at large times in a well defined way [3]. Further, we recognize now the *physical origin* of what the authors of [1] call a 'spurious dependence on the initial conditions' of Σ: $\Sigma(t)$ splits into two parts:

$$\Sigma(t) = \Sigma_\rho(t) + \Sigma_\psi(t) \tag{23}$$

where the second term Σ_ψ, which survives for $\sigma = 0$, simply reflects the randomness in the times t_n of the pattern presentations! In contrast Σ_ρ stems from the randomness in choosing a member from the pattern ensemble ρ at a given time t_n. It turns out that both contributions increase linearly at short times $\Sigma(t) = \eta^2(\sigma^2 + w^2(0)) \cdot t$ (i.e. fluctuations in w increase as $t^{1/2}$ in accordance with the general theory of stochastic processes [5]) and that $\Sigma_\psi(t)$ decays exponentially at large times t. Σ_ψ is negligible only for $t \gg 1/\eta$, where at the same time Σ_ρ has almost reached its asymptotic value Σ of (15). This remains true for arbitrary small values of the learning rate η. Conversely, for $t < 1/\eta$ the variance $\Sigma(t)$ is always significantly affected by the choice of the interval distribution $\psi(\Delta t)$ since Σ_ψ and Σ_ρ are of the same order.

These remarks explain the different forms for $\Sigma(t)$. Equation (22) is obtained for the special choice $\psi(\Delta t) = \delta(\Delta t - \tau_1)$ in the limit of small η. In this case there are no fluctuations in $\Delta t = t_n - t_{n-1}$ and therefore $\Sigma_\psi(t)$ vanishes identically for all times t. For all other distributions ψ one expects a contribution $\Sigma_\psi(t)$ similar to the one in (21).

4. Summary

We have seen that the evolution of an ensemble of neural networks is strongly influenced by the way the discrete learning events are distributed in continuous time. There is no *a priori* prescription for this assignement, which results in an ambiguity in the dynamical behaviour in physical time. Correspondingly, different choices for the time interval distribution lead to different Fokker–Planck approximations as those presented in [1] and [2]. None of them is preferable from this point of view. The long-time behaviour of the network ensemble, however, is not affected by the above choice. The resulting stationary parameter distributions are more accurately described in the approach of Radons *et al* [2, 6] since in addition to the mean, the variance is also exact in this version, whereas the results of Hansen *et al* are valid only at small learning rates. Finally we mention that the given arguments and results are not restricted to supervised learning but also apply to unsupervised learning processes such as Kohonen's self-organizing maps [11].

References

[1] Hansen L K, Pathria R and Salomon P 1993 *J. Phys. A: Math. Gen.* **26** 63

[2] Radons G, Schuster H G and Werner D 1990 *Parallel Processing in Neural Systems and Computers* ed R Eckmiller *et al* (Amsterdam: North-Holland) p 261; *Proc. Int. Neural Network Conf. INNC-90-Paris* (Dordrecht: Kluwer) p 993

[3] Bedeaux D, Lakatos-Lindenberg K and Shuler K E 1971 *J. Math. Phys.* **12** 2116

[4] Heskes T M and Kappen B 1991 *Phys. Rev.* A **44** 2718

[5] Van Kampen N G 1981 *Stochastic Processes in Physics and Chemistry* (Amsterdam: North-Holland)
[6] Radons G, Schuster H G and Werner D 1993 *Phys. Lett.* **174A** 293
[7] Oberhettinger F 1973 *Fourier Transforms of Distributions and Their Inverses* (New York: Academic)
[8] Abramowitz M and Stegun I A 1972 *Handbook of Mathematical Functions* (New York: Dover) ch 9
[9] Lukacs E 1970 *Characteristic Functions* 2nd edn (London: Griffin)
[10] This means that $P(w)$ of (16) belongs to the class of self-decomposable distributions (see [9])
[11] Ritter H and Schulten K 1988 *Biol. Cybern.* **60** 59