Stochastic dynamics of supervised learning



1993 J. Phys. A: Math. Gen. 26 63

(http://iopscience.iop.org/0305-4470/26/1/011)

J. Phys. A: Math. Gen. 26 (1993) 63–71. Printed in the UK

# Stochastic dynamics of supervised learning

Lars Kai Hansen†, Raj Pathria‡§ and Peter Salamon‡

† The Computational Neural Network Centre, Electronics Institute B349, The Technical University of Denmark, DK-2800 Lyngby, Denmark

‡ Department of Mathematical Sciences, San Diego State University, San Diego CA 92182, USA

Received 18 October 1991, in final form 13 July 1992

**Abstract.** The stochastic evolution of adiabatic (slow) backpropagation training of a neural network is discussed and a Fokker–Planck equation for the post-training distribution function in the network space is derived. The distribution we obtain differs from the one given by Radons *et al.* Studying the character of the post-training distribution, we find that, except under very special circumstances, the distribution will be non-Gibbsian. The validity of the present approach is tested on a simple backpropagation learning system in one dimension, which can be solved analytically as well. Implications of the Fokker–Planck approach for general situations are examined in the local linear approximation. Surprisingly we find that the post-training distribution is isotropic close to its peak, hence simpler than the corresponding Gibbs distribution.

## 1. Introduction

Training neural networks is a stochastic process. The result of the training process, i.e. the network *weights*, are stochastic variables depending on the specific, random training set and the possible stochastic dynamics of the search process. A recent study by Levin *et al* [1] on generalization in neural networks has focused attention on the post-training distribution of neural network weights, i.e. on the stationary distribution for the given dynamics. It has been shown that essential information concerning the learning ability of specific model-domain complexes can be derived from appropriate distribution functions. Using analogies to *thermodynamic equilibrium*, based on an assumed equivalence between maximum likelihood estimation and error minimization, Levin *et al* show that the post-training distributions are *Gibbsian*. It is of great interest to study to what extent their discussion applies to generic training schemes, such as the backpropagation supervised training scheme of feed-forward nets [2], and to see under which conditions the search process samples a Gibbs distribution with the training set error playing the role of energy. The standard implementation of backpropagation learning is a stochastic search algorithm; its stochasticity derives from partial updates based on a random sequence of examples. Learning dynamics, in general, is an active field of research. Sompolinsky *et al* [3], in particular, have analysed a nonlinear perceptron in the thermodynamic (large-network) limit and found an interesting phase diagram for nets with discrete weights. The model investigated includes a

§ Permanent address: Department of Physics, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1.

0305-4470/93/010063+09$07.50  

generic additive Gaussian white noise term. Such terms are a characteristic of thermal systems, i.e. systems interacting with a heat bath, and they generate Gibbs distributions, with the variance of the noise term acting as *temperature*. Additive thermal noise has also been studied by Hertz *et al* [4], who argue that this could model an *unreliable teacher*. Radons *et al* [5] have studied backpropagation dynamics in terms of an approximate Fokker–Planck equation for post-training distribution in network space using the Kramers–Moyal expansion.

In this communication we follow an approach similar to [5], characterizing the stochastic learning dynamics in terms of a master equation, and derive a Fokker–Planck equation of motion for the network distribution functions in the limit of slow training (i.e. small gradient descent parameter $\eta$). We, however, employ the established strategy of statistical physics, invoking a *mesoscopic* timescale, $\tau$, and arrive at a Fokker–Planck equation of the same form as Radons *et al* but with a modified diffusion matrix. In principle, this equation can be used to study both relaxation dynamics and stationary solutions. We obtain here an expression for the stationary probability distribution which indicates that the general case is indeed non-Gibbsian; it becomes Gibbsian only when the covariance matrix of the backpropagated gradients is *isotropic* and independent of the weights whose distribution is being sought. This variance determines the diffusion matrix in our approach while in Radons *et al* the same is determined by the second moment matrix. For a problem with vanishing variance, our distribution reproduces the deterministic solution, while the distribution obtained by Radons *et al* continues to give a spread around the deterministic solution. In order to study the difference between the two versions of the Fokker–Planck equation, and to expound the limiting process, we solve a simple 1D linear learning problem in some detail. We find that our version of the distribution indeed complies with the exact solution. Proceeding then to the multivariate case, we study a *student–teacher* design and obtain local geometric properties of the peak of the stationary distribution. Surprisingly we find that the peak of the distribution is isotropic, hence much simpler than the corresponding Gibbs distribution.

The paper is organized as follows: in section 2 we establish the basic master equation and deduce from it the appropriate Fokker–Planck equation. In section 3 we illustrate our approach by analysing a simple, exactly solvable, one-dimensional problem. In section 4 we discuss general feedforward nets in the local linear approximation, while we summarize our results in section 5.

## 2. The master equation and the Fokker–Planck approximation

The standard approach to supervised training of a feed-forward network is the backpropagation of errors as devised by Rumelhart and McClelland [2]. Although the consensus is that this method is prohibitively slow for real world applications, it is still widely used and is of great conceptual interest. The method is based on the minimization of overall network error on a training set of examples, and employs a gradient descent scheme to obtain that goal. Standard practice, however, invokes a momentum-smoothed, recursive gradient descent, as described in [2]. This defines a stochastic dynamical system on the weight space. In the following we shall confine ourselves to treating the simple case without momentum smoothing:

$$w_j^{(n+1)} - w_j^{(n)} = -\eta \frac{\partial E^{(n)}}{\partial w_j^{(n)}} \equiv \eta f_j^{(n)} \qquad j = 1, \ldots, N \tag{1}$$

in which the instantaneous cost function $E^{(n)} = \frac{1}{2}(y^{(n)} - F_{w^{(n)}}(x^{(n)}))^2$ is computed from a sample of the training set $(x^{(n)}, y^{(n)}) \in \{(x, y)\}$ and the current transfer function $F_{w^{(n)}}(\cdot)$ implemented by the $N$ network weights. The total training set cost (or error) is given by $E = \sum_{\alpha=1}^{p} E^{(\alpha)}$ for a training set comprising $p$ examples. The dynamics implicit in (1) is Markovian, so the probability distribution over the dynamical variable $w$ obeys the one-step master equation [6]:

$$P^{(n+1)}(w) - P^{(n)}(w) = \int \prod_{j=1}^{N} \mathrm{d}w'_j \, [W(w, w')P^{(n)}(w') - W(w', w)P^{(n)}(w)] \quad (2)$$

where $W(w, w')$ is the probability for going from $w'$ to $w$ in a single step. Following [6], we average the equation of motion for the weights over a *mesoscopic timescale* $\tau$ which is much larger than the single-step time while still being much smaller than the time scale on which the distribution $P(w)$ can change appreciably. This provides us with the following modification of (1):

$$w_j^{(t+\tau)} - w_j^{(t)} = -\eta \sum_{n(t)}^{n(t+\tau)} \frac{\partial E^{(n)}}{\partial w_j^{(n)}}. \quad (3)$$

To derive the Fokker–Planck approximation to (2), we split the right-hand side of (3) into a coherent part and a fluctuating part:

$$-\eta \sum^{\tau} \frac{\partial E^{(n)}}{\partial w_j^{(n)}} = \eta[\tau f_j^0 + \sum^{\tau}(f_j^{(n)} - f_j^0)](w^{(n(t))}) + O\left(\tau\eta^2 \left\| \frac{\partial f_j^0}{\partial w_{j'}} \right\|\right) \quad (4)$$

with $f_j^0 = \langle -\partial E/\partial w_j \rangle$ being the average force as computed from the training set. The fluctuating part, being a sum of $\tau \gg 1$ *independent* terms, can be replaced by a zero mean Gaussian variable, $\delta_j$, with covariance matrix

$$\tau\eta^2\sigma_{jj'}^2(w) \equiv \tau\eta^2\langle(f_j^{(n)} - f_j^0)(f_{j'}^{(n)} - f_{j'}^0)\rangle. \quad (5)$$

The coarse grained dynamics is now of the Langevin type, and we can write down the Fokker–Planck equation governing the dynamics of the probability distribution at the coarse grained time scale. This standard procedure yields [6]

$$\frac{\partial P(w,t)}{\partial t} = -\sum_{j=1}^{N} \frac{\partial}{\partial w_j}[f_j^0(w)P(w,t)] + \frac{\eta}{2}\sum_{j=1}^{N}\sum_{j'=1}^{N} \frac{\partial^2}{\partial w_j \partial w_{j'}}[\sigma_{jj'}^2(w)P(w,t)]. \quad (6)$$

Note that, along with $\tau$, a factor of $\eta$ has been absorbed in the new time variable $t$. This Fokker–Planck equation is different from that obtained by Radons *et al* [5] in the form of the second term. In [5] the factor $\sigma^2(w)$ is replaced by the second moment matrix of the fluctuating part rather than the second cumulant appearing here. This difference arose because Radons *et al* employed a direct expansion of the master equation in terms of $\eta$ (instead of an expansion in the mesoscopic time scale $\tau$).

According to the theory of stochastic processes [6], an essential prerequisite for the expansion of the master equation is to split the dynamical variable into a *coherent* part, which over times of order $\tau$ undergoes changes proportional to $\tau$, and a fluctuating part which, over a similar period of time, grows as $\tau^{\nu}$ where $\nu < 1$. In most applications, $\nu = \frac{1}{2}$, implying that the variance of the fluctuating part grows as $\tau$. This establishes the role that the two parts finally play in the set-up of the Fokker–Planck equation; the coherent part governs the first term (the *drift* term), while the variance of the fluctuating part governs the second term (the *diffusion* term).

## 3. Analytical solution of a 1D example and comparison with the Fokker–Planck solution

It is not surprising that a full time-dependent solution of the Fokker–Planck equation cannot be found in closed form unless the functions $f_j^0(w)$ and $\sigma_{jj'}^2(w)$ are especially simple. The stationary solution for the one-dimensional case, however, can be written down quite generally:

$$P^{(s)}(w) = \frac{C}{\sigma^2(w)} \exp\left(\frac{2}{\eta}\int^{w} \mathrm{d}w' \frac{f^0(w')}{\sigma^2(w')}\right) \tag{7}$$

where $C$ is the normalization constant. If $\sigma^2(w)$ is identically zero, the distribution (7) collapses into a delta-function centred at the deterministic solution $w = w^*$, where $f^0(w^*) = 0$. Note that the stationary distribution is *essentially singular* in the parameter $\eta$. If $\sigma^2(w) > 0$ and is *independent* of $w$, the resulting distribution is Gibbsian (i.e. the probability of a specific microstate $w$ depends solely on the cost of the state and the temperature):

$$P^{(s)}(w) \sim \exp(-\beta E(w)) \qquad (\beta = 2/\eta\sigma^2). \tag{8}$$

The quantity $\eta\sigma^2/2$ then plays the role of *temperature* which would have to be uniform over the entire $w$-space, to ensure thermal equilibrium. The generic case, however, is that (7) represents non-Gibbsian distributions. As we shall see in the next section on the multi-dimensional case the Gibbsian distribution is only recovered in the case in which the error function is isotropic in $w$-space.

We consider a simple neural network learning problem, namely a constant output net ($F_w(x) = w$) trained on noisy examples ($y_\alpha = u_\alpha$, with $u_\alpha$ a random variable). The cost function in this case is given by

$$E = \sum_{\alpha=1}^{p} E^{(\alpha)} = \frac{1}{2}\sum_{\alpha=1}^{p}(w - u^{(\alpha)})^2. \tag{9}$$

We choose this example because the probability distribution $P^{(n)}(w)$ can be derived exactly using simple analytical means. This, in turn, will enable us to make a direct comparison between the exact solution and the solution obtained from the Fokker–Planck equation.

The backpropagation dynamics of this problem is derived from

$$w^{(n+1)} - w^{(n)} = -\eta \frac{\partial E^{(\alpha(n))}}{\partial w}\bigg|_{w=w^{(n)}} \tag{10}$$

leading to the stochastic process

$$w^{(n+1)} - w^{(n)} = -\eta(w^{(n)} - u^{(n)}) \tag{11}$$

where the $u^{(n)}$ are assumed to be independent and identically distributed *Gaussian* noise impulses with zero mean and given variance $\sigma_e^2$.

The solution, for a given history of $u^{(n)}$ and starting at $w^{(0)}$, is given by

$$w^{(n)} = (1-\eta)^{n-1}\left[\sum_{m=0}^{n-1} u^{(m)}\eta(1-\eta)^{-m}\right] + w^{(0)}(1-\eta)^n. \tag{12}$$

The probability distribution of $w$ at a given instant of the process is a function of the initial value $w^{(0)}$ and can be computed from

$$P^{(n)}(w|w^{(0)}) = \int_{-\infty}^{\infty} \prod_{m=0}^{n-1} \mathrm{d}u^{(m)}\, p(u^{(m)})\delta(w^{(n)} - g(\{u^{(m)}\})) \tag{13}$$

where

$$g(\{u^{(n)}\}) = (1-\eta)^{n-1}\left[\sum_{m=0}^{n-1} u^{(m)}\eta(1-\eta)^{-m}\right] + w^{(0)}(1-\eta)^n. \tag{14}$$

Note that since $g$ is linear in the random variables $u^{(n)}$, we can use the expression

$$\delta(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{d}\theta\, \mathrm{e}^{\mathrm{i}x\theta} \tag{15}$$

to integrate over the random variables. Introducing the function $\alpha^{(n)}(\eta) = [1-(1-\eta)^{2n}]/[1-(1-\eta)^2]$, we obtain

$$P^{(n)}(w|w^{(0)}) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{d}\theta\, \mathrm{e}^{\mathrm{i}\theta(w - w^{(0)}(1-\eta)^n) - \eta^2\alpha^{(n)}(\eta)(\theta^2\sigma_e^2/2)} \tag{16}$$

$$= \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{(w - w^{*(n)})^2}{2\sigma_n^2}\right) \tag{17}$$

where

$$w^{*(n)} = w^{(0)}(1-\eta)^n \tag{18}$$

$$\sigma_n^2 = \alpha_n(\eta)\eta^2\sigma_e^2 \tag{19}$$

Equations (16)–(19), being exact, hold for *all* $n$. The stationary distribution can now be obtained by letting $n \to \infty$:

$$P_\infty(w) = \frac{1}{\sqrt{2\pi\sigma^2(\infty)}} \exp\left(-\frac{w^2}{2\sigma^2(\infty)}\right) \qquad \sigma^2(\infty) = \eta^2\sigma_e^2/(1-(1-\eta)^2) \tag{20}$$

We shall now compare the foregoing results with those following from the Fokker–Planck equation (7). With $f^0(w) = -w$ and $\sigma^2(w) = \sigma_e^2$, equation (6) can readily be integrated to give [6]

$$P(w,t) = \frac{1}{\sqrt{2\pi\bar{\sigma}^2(t)}} \exp\left(-\frac{(w - w^*(t))^2}{2\bar{\sigma}^2(t)}\right) \tag{21}$$

where

$$w^*(t) = w^{(0)}e^{-t} \qquad \bar{\sigma}^2(t) = \frac{\eta\sigma_e^2}{2}(1 - e^{-2t}). \tag{22}$$

And the corresponding stationary distribution follows by letting $t \to \infty$:

$$P^{(s)}(w) = \frac{1}{\sqrt{2\pi\bar{\sigma}^2(\infty)}} \exp\left(-\frac{w^2}{2\bar{\sigma}^2(\infty)}\right) \qquad \bar{\sigma}^2(\infty) = \frac{\eta\sigma_e^2}{2} \tag{23}$$

consistent with the general solution (16)–(19). A comparison with the Fokker–Planck results is now straightforward. To make the desired transition (from a *discrete* picture, in terms of $n$, to a *continuous* picture in terms of $t$) one must take $n \gg 1$ and $\eta \ll 1$, so that the process entails a large number of steps, each step resulting in a small change in the variable $w$. Equations (18)–(19) may then be written as

$$w^{*(n)} \simeq w^{(0)}e^{-\eta n} \qquad \sigma_n^2 \simeq \tfrac{1}{2}\eta\sigma_e^2(1 - e^{-2\eta n}). \tag{24}$$

With $\eta n$ corresponding to $t$, the two descriptions are clearly identical; we note that for *relatively low* $n$ (such that while $n \gg 1$, $\eta n \ll 1$), equations (24) reduce to

$$w^{*(n)} - w^{(0)} \simeq -w^{(0)}\eta n \qquad \sigma_n^2 \simeq \eta^2\sigma_e^2 n \tag{25}$$

so that both *drift* and *diffusion* in the variable $w^{(n)}$ are proportional to the first power in $n$. Recognizing that $n$ in this short duration of the process is a measure of the mesoscopic time $\tau$, equations (25) are in agreement with expressions (4)–(5) that form the basis of our passage from the master equation (2) to the Fokker–Planck equation (6).

In contrast, Radons *et al* [5] carried out an expansion of the single-step master equation (2) in powers of $\eta$, without invoking a mesoscopic time scale $\tau$ (which in a sense amounts to putting $\tau = 1$), with the result that in expression (6) they retained the term $(f^0(w))^2$ as well as $\sigma^2(w)$. Applying their equation to the problem studied in this section, one cannot obtain a full, time-dependent solution in closed form. The stationary solution, however, turns out to be

$$P^{(s)}(w) \sim \frac{1}{(w^2 + \sigma_e^2)^{1/\eta+1}} \tag{26}$$

which may be compared with the exact result (23).

As regards time dependence, we may look at the mean and the variance of $w$, which can be evaluated *exactly* from the evolution equations [6]

$$\frac{d}{dt}\langle w\rangle = \langle a_1(w)\rangle \tag{27}$$

and

$$\frac{\mathrm{d}}{\mathrm{d}t}\sigma^2(w) = 2\langle (w - \langle w\rangle)a_1(w)\rangle + \langle a_2\rangle \tag{28}$$

where $a_1(w) = -w$, while $a_2(w) = \eta\sigma_e^2$ in our version of the theory but $\eta(w^2+\sigma_e^2)$ in the version of Radons *et al.* Integrating (27) and (28), we find that, whereas $\langle w\rangle$ in either case is given by $w^{(0)}\mathrm{e}^{-t}$, $\sigma^2(w)$ in the case of Radons *et al* turns out to be

$$[\sigma^2(w)]_{\mathrm{R}} = \tfrac{1}{2}\eta\sigma_e^2[1 - \mathrm{e}^{-(2-\eta)t}] + (w^{(0)})^2[\mathrm{e}^{-(2-\eta)t} - \mathrm{e}^{-2t}] \tag{29}$$

$$\simeq \tfrac{1}{2}\eta\sigma_e^2[1 - \mathrm{e}^{-2t}] + (w^{(0)})^2\eta t\mathrm{e}^{-2t} \qquad (\eta \ll 1). \tag{30}$$

Comparing (29)–(30) with (22) and (24), we find that $\sigma^2(w)$ following from the formulation of Radons *et al* possesses a spurious dependence on the initial conditions which persists even if $\sigma_e^2$ is set equal to zero. Thus, even if the stochasticity of the process is removed, the variable $w$ continues to be statistical in character. However, since both approaches are approximate, we cannot rule out the possibility that there exist noise models, e.g. models violating our assumption of a finite variance of the noise, for which the Fokker–Planck equation of [5] provides a more accurate description of the relaxation phenomena.

Before closing this section we note that, since the *diffusion* term $\sigma^2(w)$ for the problem studied here was independent of $w$, the stationary solution was expected to be Gibbsian. A closer look at equation (20) or (23) reveals that the solution is indeed Gibbsian although, in view of the fact that the drift term is linear (and hence the cost-function quadratic) in $w$, the final distribution takes the form of a *Gaussian* distribution.

## 4. General feedforward networks

As an application of the Fokker–Planck scheme in the multivariate case, we consider now the case of a general feedforward network *close to* the conclusion of the training process. In this regime the network is supposedly fairly well settled close to a minimum of the error function. Hence, we can make a local linear approximation valid for small fluctuations around the minimum†. We will consider a standard feedforward net parametrized by a set of $N$ weights $w$, implementing a real-valued function of $L$ input variables $x$:

$$y(x) = F_w(x). \tag{31}$$

This *student* network is trained on a database generated by a *teacher* network with weights $w^*$:

$$y(x^{(\alpha)}) = F_{w^*}(x^{(\alpha)}) + \nu^{(\alpha)}. \tag{32}$$

Further we assume that the noise $\nu$ is independent of the input $x$. The cost function, $E$, is the sum of squares as previously. The Fokker–Planck equation is specified by the first two moments of the distribution of the backpropagation gradients

$$f_j^0 = \left\langle -\frac{\partial E}{\partial w_j} \right\rangle \tag{33}$$

† Note that in the vicinity of the maximum of the stationary distribution, i.e. where $f^0 = 0$, the two forms of the Fokker–Planck equation agree.

and

$$\sigma^2_{jj'} = \left\langle \left( -\frac{\partial E}{\partial w_j} - f^0_j \right) \left( -\frac{\partial E}{\partial w_{j'}} - f^0_{j'} \right) \right\rangle . \tag{34}$$

We compute the moments to first order around the local minimum:

$$f^0_j = \sum_{j'} A_{jj'}(w_{j'} - w^*_{j'}) \tag{35}$$

and

$$\sigma^2_{jj'} = \sigma^2_\nu A_{jj'} \tag{36}$$

where $A_{jj'}$ is the *correlation matrix* of the random derivatives of the network output:

$$A_{jj'} = \left\langle \frac{\partial F_w(x)}{\partial w_j} \frac{\partial F_w(x)}{\partial w_{j'}} \right\rangle \Bigg|_{w=w^*} \tag{37}$$

and $\sigma^2_\nu$ is the variance of the additive noise. The resulting Fokker–Planck equation is then found to be

$$\frac{\partial P(w,t)}{\partial t} = -\sum_{j,j'=1}^{N} A_{jj'} \frac{\partial}{\partial w_j}[(w_{j'} - w^*_j) P(w,t)]$$
$$+ \frac{\eta \sigma^2_\nu}{2} \sum_{j,j'=1}^{N} A_{jj'} \frac{\partial^2}{\partial w_j \partial w_{j'}} P(w,t). \tag{38}$$

The stationary solution for a Fokker–Planck equation of this form is given by a simple isotropic Gaussian [6]:

$$P(w) = Z^{-1} \exp\left( -\beta \sum_{j=1}^{N} (w_j - w^*_j)^2 \right) \tag{39}$$

where $\beta = 2/\eta\sigma^2_\nu$, and $Z$ is the normalization constant. It is quite remarkable that the characteristics of the cost function, $A_{jj'}$, are totally absent from the close neighbourhood of the peak of the distribution. The physical reason is that the geometric structure of the average cost function, which determines the drift term, is exactly 'cancelled' by the structure of the diffusion term. Incidentally we note that the isotropy of the *weight covariance matrix* has been derived by quite different means by Widrow and Stearns for the so-called LMS (least mean squares) scheme [7].

A Gibbs distribution, on the other hand, is for the present problem, and in the local linear approximation, given by

$$P(w) = Z^{-1} \exp\left( -\frac{\beta}{2} \sum_{jj'} A_{jj'}(w - w^*)_j (w - w^*)_{j'} \right). \tag{40}$$

We note that we obtain a Gibbs distribution only in the special case where the $A$ matrix is itself isotropic i.e. proportional to the unit.

## 5. Concluding remarks

It has been shown that the standard Fokker–Planck approach of statistical physics can be a valuable tool in understanding learning dynamics, in the limit of slow training. We have pointed out that the stationary distribution in the weight space of a neural network, after training by backpropagation, is typically non-Gibbsian. In the one-dimensional case the general form of the stationary distribution can be given. In the multivariate case we found, surprisingly, that the peak of the stationary distribution is isotropic.

## Acknowledgments

We wish to acknowledge the inspiring discussions at the Neural Net Workshop of the Telluride Summer Research Center 1990, where this work was initiated. Furthermore, we thank Benny Lautrup for useful discussions and for informing us about the work of Radons *et al* [5], and Jan Larsen for a pointer to Widrow and Stearns. We thank the anonymous reviewers for very valuable suggestions. LKH and RKP wish to thank the Interdisciplinary Research Center of San Diego State University for partial support. LKH is supported by the Danish Natural Science and Technical Research Councils through the Computational Neural Network Centre (CONNECT), while RKP is supported by the Natural Sciences and Engineering Research Council of Canada.

## References

[1] Levin E, Tishby N and Solla S 1989 A statistical approach to learning and generalization in layered neural networks *Proc. 2nd Ann. Workshop on Computational Learning Theory (COLT'89)* ed R Rivest, D Haussler and M K Warmuth (San Mateo, CA: Morgan Kaufmann) pp 280–95; 1990 *Proc. IEEE* **78** 1574

[2] Rumelhart D E and McClelland J L 1986 Back-propagation of errors *Parallel Distributed Processing. Explorations in the Microstructure of Cognition* vols 1–2 (Cambridge, MA: MIT)

[3] Sompolinsky H, Tishby N and Seung H S 1990 Learning from examples in large neural networks *Phys. Rev. Lett.* **65** 1683–6

[4] Hertz J, Thorbergson S and Krogh A 1989 Phase transitions in simple learning *J. Phys. A: Math. Gen.* **22** 2133

[5] Radons G, Schuster H G and Werner D 1990 Drift and diffusion in backpropagation learning *Parallel Processing in Neural Systems and Computers* ed R Eckmiller *et al* (Amsterdam: North-Holland) p 261–4

[6] Van Kampen N G 1983 *Stochastic Processes in Physics and Chemistry* (Amsterdam: North-Holland)

[7] Widrow B and Stearns S D 1985 *Adaptive Signal Processing* (Englewood Cliffs, NJ: Prentice-Hall)