The Langevin method in the statistical dynamics of learning

1990 J. Phys. A: Math. Gen. 23 L763

(http://iopscience.iop.org/0305-4470/23/15/012)

J. Phys. A: Math. Gen. **23** (1990) L763-L766. Printed in the UK

LETTER TO THE EDITOR

The Langevin method in the statistical dynamics of learning

R Der

Karl-Marx-Universität Leipzig, Sektion Informatik, FG Neuroinformatik, Karl-Marx-Platz 10-11, Leipzig 7010, German Democratic Republic

Received 1 February 1990

Abstract. The statistical dynamics of learning in the presence of noise has been treated recently by Hertz, Krogh and Thorbergson using a Langevin model. Fluctuations introduced by this model into the space orthogonal to the space of patterns are shown to influence the dynamics of learning in a spurious way. This is avoided in the new Langevin model in which the noise is introduced directly into the cost function by using randomly modulated training patterns. Correspondingly a new kind of response function is defined. Unconstrained learning and learning under constrained thermal fluctuations are discussed as examples. Explicit expressions—e.g. for the learning time—are found to differ substantially from those given by Hertz *et al.*

The efficiency of learning procedures in layered neural networks has been of much interest recently. The problem may be traced back to that of minimising a given cost function by some kind of gradient descent. With the presence of noise this is a statistical dynamical problem which was formulated recently by Hertz, Krogh and Thorbergson (HKT) in the framework of the Langevin method, cf [1]. There the noise was introduced *ad hoc* into the dynamical equation describing the gradient descent. This leads to spurious noise contributions in the space Q orthogonal to the space P of the patterns, as will be discussed in more detail below. The present letter proposes a convenient reformulation of this method by starting from a time-dependent cost function which formulates the problem of learning under noise in a different way.

Let us consider, as in [1], a one-layer perceptron which is to associate a set of p input patterns ξ_i^μ, $\mu = 1, \ldots, p$; $i = 1, \ldots, N$ with a certain output unit taking given training values β^μ. The cost function measures the distance $d^\mu = p^\mu - \beta^\mu$ between the values β^μ of the output unit and the post-synaptic potential $p^\mu = (1/\sqrt{N}) J_i \xi_i^\mu$. For the sake of simplicity in the present letter both ξ and β are assumed to be random numbers with zero mean. Note that here sums run over all indices occurring twice (Einstein convention) of Roman or Greek letters from 1 to N or 1 to p, respectively.

The peculiarity of the cost function to be proposed consists of random modulations of the training value β^μ caused by a time-dependent stochastic process $f^\mu(t)$ added to it (noisy trainer). The complete cost function $E(t)$ then is given by

$$2E(t) = \sum_\mu (d^\mu + f^\mu(t))^2 + c \sum_i J_i^2 \tag{1}$$

where c is the chemical potential, cf [1]. $f(t)$ is a Gaussian white noise given by

$$\langle f^\mu(t) \rangle = 0 \qquad \langle f^\mu(t) f^{\mu'}(t') \rangle = 2T\delta(t-t')\delta_{\mu,\mu'} \tag{2}$$

where $\langle \ldots \rangle$ means the average over the noise and the fictitious temperature T measures the strength of the noise.

0305-4470/90/150763+04$03.50

Learning corresponds to changing the synaptic strength proportional to the negative gradient of E, $\Delta J_i \sim -\partial E/\partial J_i$. Because of the occurrence of the stochastic process $f(t)$ this gradient is noisy so that the approach to the minimum of E corresponds to a stochastic dynamics. This leads in continuous time to the generalised Langevin equation

$$\dot{J}_i = -D_{i,j}J_j + B_i + b_i(t) \tag{3}$$

where the overdot denotes the time derivative and

$$\begin{aligned} &D_{i,j} = A_{i,j} + c\delta_{i,j} \qquad A_{i,j} = (1/N)\xi_i^\mu \xi_j^\mu \\ &B_i = (1/\sqrt{N})\xi_i^\mu \beta^\mu \qquad \text{and} \qquad b_i(t) = (1/\sqrt{N})f^\mu(t)\xi_i^\mu. \end{aligned} \tag{4}$$

The constant of proportionality was included into the unit of time. The noise now acts in the space of the patterns only, whereas in [1] the noise was introduced into the Langevin equation *ad hoc* and is found to act in the entire space.

The difference between the two approaches is seen best by considering the fluctuation dissipation theorem (FDT), cf [2]. We define $\delta J_i(t) = J_i(t) - \langle J_i\rangle_{\text{eq}}$ where $\langle J_i\rangle_{\text{eq}} = D^{-1}_{i,j}B_j$, is obtained from (3) in using $\langle \dot{J}\rangle = 0$ in equilibrium, cf [1]. From the explicit (formal) solution of (3) we obtain for very large times t

$$\delta J_i(t) = (1/\sqrt{N})\int_0^t \mathrm{d}t'(\exp[-(t-t')D])_{i,j}\xi_j^\mu f^\mu(t')$$

and hence using (2)

$$\begin{aligned} C &:= (1/N)\lim_{t\to\infty}\langle \delta Ji(t)^2\rangle \\ &= (1/N)(\langle Ji^2\rangle_{\text{eq}} - (\langle Ji\rangle_{\text{eq}})^2) \\ &= (T/N)\,\mathrm{Tr}(AD^{-1}) \\ &= T(1-cG) \end{aligned} \tag{5}$$

(sum over i implied) where Tr denotes the trace of a matrix and $G := (1/N)\,\mathrm{Tr}(D^{-1})$ obeys the equation

$$G = 1/(c + \alpha/(1+G)) \tag{6}$$

$\alpha = p/N$, as derived earlier by HKT, cf also [3]. Using (6) we may obtain from (5)

$$C = \alpha TG/(1+G). \tag{7}$$

Equation (7) is an anomalous FDT. An FDT of the usual kind is obtained by way of introducing a new response function r which represents the response of the post-synaptic potential (the signal) p^μ with respect to a variation of the training value β^μ of the output unit, i.e.

$$r := (1/N)\partial\langle p\mu\rangle_{\text{eq}}/\partial\beta\mu \tag{8}$$

where $\langle p^\mu\rangle_{\text{eq}} = (1/\sqrt{N})\langle J_i\rangle_{\text{eq}}\xi_i^\mu$, $\langle J_i\rangle_{\text{eq}} = D^{-1}_{i,j}B_j$, cf equation (12) of [1]. r is easily evaluated as $r = \alpha G/(1+G)$, hence

$$C = Tr \tag{9}$$

is a FDT of the usual kind so that r is seen to be the proper response function conjugate to the correlation function C.

Using (6), r is found to obey the exact equation

$$r = 1/(1 + c/(\alpha - r)) \tag{10}$$

with solution

$$2r = 1 + \alpha + c - \sqrt{(1+\alpha+c)^2 - 4\alpha}. \tag{11}$$

The negative radical is appropriate here since from (8) we immediately verify that for $c \to \infty$ we have $r = \alpha/c$ in agreement with (11). The most important difference between our response function r and the G of HKT is found in the fact that G has a pole at $c = 0$ whereas r is regular there.

Central system quantities are readily expressed in terms of our response function r. For instance the analogue of the Edwards-Anderson order parameter $q = (1/N)(\langle J_i \rangle_{eq})^2$ (cf equation (25) of [1]) is

$$q = -r' = r^2/(\alpha - r^2) \tag{12}$$

where $r' = \partial r/\partial c$ was obtained from (10). For the learning time τ we find the expression

$$\tau = -r'/r = q/r. \tag{13}$$

The essential difference between the two approaches is seen clearly from considering τ at small $c > 0$ where $\tau = (1-\alpha)^{-1} + O(c)$ in our theory. Here τ is the time of learning corresponding to the relaxation process in pattern space P. For the HKT model we obtain $\tau = c^{-1} + O(c)$. This is the relaxation time of the slowly decaying component of J in Q space (orthogonal to the patterns) which has nothing to do with the dynamics of learning.

On more general grounds the difference between the two approaches discloses in the different meanings of the corresponding FDTs, i.e. (9) in our case and

$$C = TG \tag{14}$$

in the HKT case. The FDTs allow us to express q and τ directly in terms of C which is the mean square deviation of J caused by the noise from its equilibrium value. Let us consider learning under constrained thermal fluctuations, i.e. we assume $C = S^2$, S^2 given. Then using (9) we find $r = S^2/T$ and hence

$$q = \alpha/(\alpha_0 - \alpha) \qquad \tau = (\sqrt{\alpha_0})/(\alpha_0 - \alpha). \tag{15}$$

Here $\alpha_0 = (S^2/\alpha T)^2$ whereas $\alpha_0 = (1 + T/S^2)^2$ in [1].

The case of unconstrained learning is included now since α_0 is continuous at $c = 0$ where $\alpha_0 = 1$. Note also that with $T \to 0$ also $S^2 \to 0$, since there are no fluctuations without noise. Hence, α_0 is always finite.

Equations (15) connect the characteristic quantities q and τ with S^2, i.e. with the strength of the synaptic fluctuations. These expressions are seen to be essentially different from the corresponding expressions of HKT. Obviously, this results from the different meanings of S^2 connected with the different nature of the fluctuations entering the theory. In our case S^2 measures the fluctuations in P space which are directly connected with the dynamics of learning via the FDT (9). In HKT S^2 comprises also the fluctuations in Q space which are spurious with respect to the learning procedure. Nevertheless they are never negligible and even prevail for small c.

The difference between the two approaches is particularly transparent if the learning starts from the *tabula rasa* condition ($J(0) = 0$) or from the Hopfield matrix, i.e. $J(0) = B$, cf (4). Then in our model the Q space is empty at all times whereas in the HKT model there are fluctuations in Q space with a strength increasing in time from zero up to its equilibrium value which is of order $1/c$ for small c. Hence a convenient description of the statistical dynamics of learning should always include fluctuations

only in the space of the patterns as has been done in the present letter. More details and further applications of the present approach, e.g. to the model by Opper [2], cf also [3], will be given in a forthcoming paper.

References

[1] Hertz J A, Krogh A and Thorbergson G I 1989 *J. Phys. A: Math. Gen.* **22** 2133
[2] Opper M 1989 *Europhys. Lett.* **8** 389
[3] Englisch H and Pastur L A 1989 Spectral Analytic Approach to the ADALINE Learning for Neural Nets *Preprint* KMU-NTZ-89-11 (Leipzig) (*Proc. Int. Workshop on Neurocomputing and Attention, Moscow 1989* submitted)