Sensibility of neural networks

1987 J. Phys. A: Math. Gen. 20 L85

(http://iopscience.iop.org/0305-4470/20/2/008)

J. Phys. A: Math. Gen. **20** (1987) L85–L88. Printed in the UK

LETTER TO THE EDITOR

Sensibility of neural networks

R Nemeth†

Department of Atomic Physics, Eötvös University, Puskin u 5-7, Budapest, H-1088, Hungary

Received 6 November 1986

Abstract. We introduce and investigate a susceptibility-like quantity which characterises the sensibility of neural models. It can be calculated without the replicas in the Hopfield model and may be useful in the investigation of highly non-linear learning rules.

In recent years there has been much interest in the field of neural spin models originally invented by Hopfield (1982). Different features and approximate solutions of such models have been widely discussed in the literature (Amit *et al* 1985a, b, Nadal *et al* 1986, Kinzel 1985, Weisbuch and Fogelman-Soulie 1985, Hopfield *et al* 1983, Personnaz *et al* 1986). These models are constructed to learn and retrieve information which is encoded into a spin configuration (pattern). In the ideal case it should recognise all the learned patterns but in reality the number of learned patterns is lower than the number of spins in the models investigated so far.

In this letter we are looking for the largest value of patterns at which all of them are still recognised and would like to see how the system behaves in the neighbourhood of this critical value. We will use the Hopfield algorithm which uses an Ising model with zero-temperature Glauber dynamics. The Hamiltonian is

$$H = -\sum_{(ik)} J_{ik} s_i s_j \tag{1}$$

and for J_{ik} we have

$$J_{ik} = \left(\sum\nolimits_a s_i^a s_k^a\right) N^{-1}. \tag{2}$$

Here N denotes the number of spins, P the number of learned patterns, a indicates the patterns and i, j are the spins (neurons) with $s_i^a = \pm 1$. Let y_i^a denote the sign of the product of the spin and the local field

$$y_i^a = \mathrm{sgn}\left(s_i^a \sum_{k \neq i} J_{ik} s_k^a\right). \tag{3}$$

The spin changes its sign if $y_i^a = -1$ (we investigate only the learned configurations).

A state will live for a long time (or metastable) if $y_i^a = 1$ but we do not demand that for all the y_i^a. Instead of the y_i^a we use an order parameter-like quantity Y^a:

$$Y^a = \left(\sum_i y_i^a\right) N^{-1} \tag{4}$$

† Present address: Institut für Theoretische Physik, Cologne University, 5000 Köln 41, West Germany.

0305-4470/87/020085+04$02.50

and we say that the pattern is metastable if $Y^a \to 1$ if $N \to \infty$. With the help of Y^a we can express the portion of 'wrong' neurons N^*/N as

$$N^*/N = (1 - Y^a)/2. \tag{5}$$

We will use random patterns (Hopfield 1982) and $\langle \ldots \rangle$ will denote the average over the randomness and we drop the a indices since the average does not depend on the given pattern. So

$$\langle Y \rangle = \left(\sum_i \langle y_i \rangle \right) N^{-1} \tag{6}$$

where $\langle Y \rangle$ is a magnetisation-like quantity so we can call the susceptibility, χ:

$$\chi c = \left\langle \left(\sum_i y_i \right)^2 \right\rangle N^{-1} - \left(\sum_i \langle y_i \rangle \right)^2 N^{-1}. \tag{7}$$

This quantity characterises the sensibility of the memory to new patterns to be learned. On the other hand, Y is a random variable and for the description of a random variable we have to use at least the first two cumulants. (This would be enough in the thermodynamic limit if the y_i were independent.)

Now we would like to calculate this quantity χ. For y_i^a we have from (2) and (3)

$$y_i^a = \mathrm{sgn}\left[\left(s_i^a \sum_{k \neq i} \sum_{(b \neq a)} s_i^b s_k^b s_k^a \right) N^{-1} + (N-1)/N \right] \tag{8}$$

where we separated the $b = a$ part from the sum. Now all the terms in the double sum are independent random variables ($s_i^a = \pm 1$ with equal probability) so the central limit theorem is applicable and we can write for the average of y_i^a

$$\langle y_i \rangle = 1 - (2\alpha/\pi)^{1/2} \exp(-1/2\alpha) \tag{9}$$

with $\alpha = P/N$ in the $\alpha \ll 1$ case using the limiting form of the function $\mathrm{erf}(x)$.

To evaluate the $\langle y_i y_k \rangle$ average we separate (8) into two further parts:

$$y_i^a = \mathrm{sgn}\left[\left(s_i^a \sum_{b \neq a} s_i^b s_k^b s_k^a + s_i^a \sum_{b \neq a} \sum_{r \neq i,k} s_i^b s_r^b s_r^a + (N-1) \right) N^{-1} \right] \tag{10a}$$

$$y_k^a = \mathrm{sgn}\left[\left(s_k^a \sum_{b \neq a} s_k^b s_i^b s_i^a + s_k^a \sum_{b \neq a} \sum_{r \neq i,k} s_k^b s_r^b s_r^a + (N-1) \right) N^{-1} \right] \tag{11a}$$

or with

$$x = \left(s_i^a \sum_{b \neq a} s_i^b s_k^b s_k^a \right) N^{-1} \qquad z = \left(s_i^a \sum_{b \neq a} \sum_{r \neq i,b} s_i^b s_r^b s_r^a \right) N^{-1}$$

$$w = \left(s_k^a \sum_{b \neq a} \sum_{r \neq i,b} s_k^b s_r^b s_r^a \right) N^{-1}$$

$$y_i = \mathrm{sgn}(x + z + (N-1)/N) \tag{10b}$$

$$y_k = \mathrm{sgn}(x + w + (N-1)/N) \tag{11b}$$

where z and w are independent variables and have the same density function $f(z)$ for both of them, namely

$$f(z) = \exp(-z^2/2\alpha)/(2\alpha\pi)^{1/2} \tag{12}$$

and the distribution of x is also Gaussian but the width of it is P/N^2.

In the $N \to \infty$ limit y_i^a and y_k^a become independent random variables since the effect of x is $o(1/N)$ but because of the square in the RHS of (7) we have to take into account this correlation as well. Taking into account the correlations we can get for χ:

$$\chi \sim \exp(-1/2\alpha)/\alpha \tag{13}$$

and if we neglect the correlation of y_k^a and y_i^a

$$\chi = (\alpha/2\pi)^{1/2} \exp(-1/2\alpha) \tag{14}$$

in (13) we keep only the leading term for small α. So we can see that χ is non-zero even if we suppose the independence of y_i^a.

The essential singularity in expression (9) reminds of us the zero-temperature result of Amit *et al* (1985b). They have found for the portion of the wrong neurons the same as (16). We can calculate the average of Y^a:

$$\langle Y \rangle = 1 - (\alpha/2\pi)^{1/2} \exp(-1/2\alpha) \tag{15}$$

or with N^*:

$$N^*/N = (\alpha/2\pi)^{1/2} \exp(-1/2\alpha). \tag{16}$$

However, because all the cumulants have to go with $\exp(-1/2\alpha)$ and the cumulants higher than second-order are not negligible it means some differences to the result of Amit *et al* (1985b).

In order to reproduce their findings we assume the independence of the y_i^a. Then we can apply the central limit theorem and can see a rapid increase of the distribution function from zero to one at $\langle Y \rangle = 1 - 2(\alpha/2\pi)^{1/2} \exp(-1/2\alpha)$. This transition-like behaviour was also seen in the replica symmetric solution which we can reproduce in this way. The two kinds of derivations used different approximations, one in each case, so these approximate steps have to correspond to each other. Therefore replica symmetry breaking corresponds to spatial correlations of the neurons. The occurrence of replica symmetry breaking is a remote analogy to the occurrence of non-zero correlation length in ferromagnets if we increase the temperature from zero to a finite value. However, in the case of the memory model the disordering effect is due to the increase of the number of the patterns taught (frustration). The spatial ordering of spins can be characterised in the Sherrington-Kirkpatrick model (Mezard and Virasoro 1985), but the transition to the replica symmetry breaking phase was not investigated in that paper.

Since the increase of $\langle Y \rangle$ and χ is very slow at the transition point it may explain why the computer simulations work rather well and give finite α_c. Besides, the weakness of the transition strengthens the assumption that the replica symmetric solution gives results which are very close to the exact ones.

We used this method (Geszti and Nemeth 1986) to investigate a more complicated model (Geszti 1986) where the replica trick cannot be applied in an easy way. It works rather well in that highly non-linear case so we guess that this method may be useful in the investigation of other non-linear models as well.

I thank T Geszti, J Kertesz and I Kondor for useful discussions.

References

Amit D J, Gutfreund H and Sompolinsky H 1985a *Phys. Rev.* A **32** 1007
—— 1985b *Phys. Rev. Lett.* **55** 1530
Geszti T 1986 *Phys. Lett.* **114A** 336
Geszti T and Nemeth R 1986 unpublished
Hopfield J J 1982 *Proc. Natl Acad. Sci.* **79** 2554
Hopfield J J, Feinstein D I and Palmer R G 1983 *Nature* **304** 158
Kinzel W 1985 *Z. Phys.* B **60** 205
Mezard M and Virasoro M A 1985 *J. Physique* **46** 1293
Nadal J P, Toulouse G, Changeux J P and Dehane S 1986 *Europhys. Lett.* **1** 535
Personnaz L, Guyon I, Dreyfus G and Toulouse G 1986 *J. Stat. Phys.* **43** 411
Weisbuch G and Fogelman-Soulie F 1985 *J. Physique Lett.* **46** L623