LETTER TO THE EDITOR

Symmetry breaking in non-monotonic neural networks

G Boffetta†, R Monasson‡ and R Zecchina§

† Dip. di Fisica Generale e INFN, Università di Torino, Via P.Giuria 1, 10125 Torino, Italy
‡ Laboratoire de Physique Théorique, E.N.S., 24 rue Lhomond, 75231 Paris Cedex 05, France
§ Dip. di Fisica Teorica e INFN, Università di Torino, Via P.Giuria 1, 10125 Torino, Italy, and Dip. di Fisica, Politecnico di Torino, C.so Duca degli Abruzzi 24, 10129 Torino, Italy

Received 19 January 1993, in final form 22 March 1993

Abstract. The optimal performance of a non-monotonic neural network is studied by the replica method. In analogy to what happens in multi-layered networks, we show that replica symmetry breaking (RSB) is required. The distribution of the patterns stabilities, the correlations in the distribution of the internal representation and the optimal capacity per synapse ($\alpha_c \simeq 4.8$) are computed with one step of RSB.

Over the past few years, much effort has been concentrated in studying optimal storage capabilities of neural networks. Using the framework developed by Gardner–Derrida [1], many different aspects, as for instance the architecture of the networks, the nature of their synapses and the statistical distribution of the patterns, have been shown to influence storage capacity. An open and more difficult question concerns the role played by the possible dynamical schemes of the single formal neuron for the computation capabilities of neural networks with given synaptic configuration and architecture. In most of the models considered so far, the neurons are defined in terms of sigmoidal (monotonic) transfer functions describing the mean firing rate of the neurons as a function of their local field, or post-synaptic potential (in the limit of infinite gain, those transfer functions become step-like, leading to the binary neurons of spin-glass models). Recently, the introduction of non-monotonic transfer functions (instead of the usual sigmoidal or sign-type outputs) was shown to play a significant role both for associative performance (storage capacity) [2–5] and for computational capabilities (dynamical selection of optimal sub-networks) [5]. In the latter models, non-monotonicity is thought to describe an effective behaviour of the formal neurons, due to the presence of local inhibitory feedbacks (caused by inhibitory inter-neurons) which control the system dynamics by lowering the neural activity. Beside the biological motivations mentioned (which, to our knowledge, are not established experimentally) and computational arguments, the theoretical problems arising in the context of the statistical mechanics of non-monotonic networks (NMN) are worth being discussed for their general interest. This is the purpose of the present letter. The optimal associative performance (storage capacity) of NMN is deeply related to the role played, in the space of interactions, by symmetry breaking, as it is for multi-layered networks (MLN). As we shall see below, the model discussed can be mapped straightforwardly onto a two layer parity machine, which makes the connection with the theory of learning in MLN even more clear. The NMN therefore provides us with a toy model of MLN which is analytically tractable and allows for a very simple way of labelling the couplings domains of the broken symmetry phase, diversely from what happens for the classical two-layers networks.

The model we investigate here is the simple fully connected non-monotonic network with binary neurons; we study, in particular, the fractional volume [1] in the space of couplings such that the condition of learning is satisfied for all the patterns. As far as optimal capacity is concerned, our system is equivalent to a feed-forward perceptron including N inputs S_i $(i = 1, \ldots, N)$, connected to an output cell σ by real valued synapses J_i. When an input pattern ξ_i is presented, the network computes the output field $h = \boldsymbol{J} \cdot \boldsymbol{\xi}$ ($\boldsymbol{J}^2$ is normalized to 1) and σ is equal to $+1$ if $h < -\gamma - k$ or $k < h < \gamma - k$, -1 otherwise. k is the stability parameter (for simplicity we will assume hereafter $k = 0$), γ is an arbitrary threshold and we recover the usual sign function for $\gamma = \infty$. Let us consider now P pairs $(\boldsymbol{\xi}^\mu, \sigma^\mu)$ of random and unbiased binary patterns. Each pattern $\boldsymbol{\xi}^\mu$ is said to be stored if it is mapped onto its corresponding output σ^μ, or in other words, if its stability

$$\Delta^\mu \equiv \sigma^\mu \boldsymbol{J} \cdot \boldsymbol{\xi}^\mu \tag{1}$$

belongs to $]-\infty, -\gamma]$ or $[0, \gamma]$. In the large N limit, the critical capacity α_c is defined as the maximum ratio P/N below which there exists a vector of couplings storing the whole training set [1].

Before turning to the analytic results, it is worth observing that when the transfer function is truly non-monotonic (i.e. γ is finite), the space of the couplings storing the patterns perfectly is not connected. Let us indeed consider two synaptic vectors $\boldsymbol{J}$, $\boldsymbol{K}$ and a stored pattern $\boldsymbol{\xi}^1$, the stabilities of which are respectively lower than $-\gamma$ (with $\boldsymbol{J}$) and between 0 and γ (with $\boldsymbol{K}$). Obviously, for any path linking $\boldsymbol{J}$ to $\boldsymbol{K}$ on the unit sphere, there exist weights which do not store $\boldsymbol{\xi}^1$ (its stability would belong to $[-\gamma, 0]$). This situation is quite reminiscent of the multi-layered networks, for which finding a storage algorithm whose running time would be polynomial in N seems to be unlikely. A second consequence is that the symmetry of the replicas should be broken for the computation of α_c. The analogy with multi-layered networks is even stronger when one realizes that the present model is equivalent to a fully connected parity machine with three hidden units (whose thresholds are $-\gamma$, 0 and $+\gamma$) and where the three synaptic vectors are equal to $-\boldsymbol{J}$. The number of allowed internal representations is however four and not eight, and thus the computational abilities of the non-monotonic model should rather be compared to the two hidden units parity network, both of them dividing the input patterns space in four different regions. The study of the latter network [7] (with non-overlapping fields) showed that replica symmetry breaking is required and that the one-step RSB solution makes α_c decrease from 5.5 (symmetric calculation) to 4. Such a behaviour may be qualitatively expected in our case.

Let V be the fraction of the couplings storing the training set perfectly

$$V(\{\boldsymbol{\xi}^\mu, \sigma^\mu\}) = \frac{\displaystyle\int \mathrm{d}\boldsymbol{J}\, \delta(\boldsymbol{J}^2 - 1) \prod_{\mu=1}^{P} \theta_\gamma(\Delta^\mu)}{\displaystyle\int \mathrm{d}\boldsymbol{J}\, \delta(\boldsymbol{J}^2 - 1)} \tag{2}$$

where $\theta_\gamma(\Delta)$ is defined by

$$\theta_\gamma(\Delta) \equiv \begin{cases} 1 & \text{if } \Delta \leqslant -\gamma \text{ or } 0 \leqslant \Delta \leqslant \gamma \\ 0 & \text{if } \Delta > \gamma \text{ or } -\gamma < \Delta < 0 \end{cases} \tag{3}$$

and the stability Δ is given in (1). In the thermodynamic limit, $f = (1/N)\overline{\ln V}$ (the bar indicating the average over the pattern distribution) is assumed to be self-averaging and is computed using the replica method [1, 6]. As is customary in the replica approach, we introduce the matrix of the overlaps between the different replicas $q^{ab} = \boldsymbol{J}^a \cdot \boldsymbol{J}^b$. The replica symmetric (RS) ansatz gives

$$f_{\mathrm{rs}}(q, \alpha) = \frac{1}{2}\ln(1-q) + \frac{q}{2(1-q)} + \alpha \int_{-\infty}^{+\infty} Dx \ln\left[H\left(\frac{x\sqrt{q}}{\sqrt{1-q}}\right) - H\left(\frac{\gamma + x\sqrt{q}}{\sqrt{1-q}}\right) + H\left(\frac{\gamma - x\sqrt{q}}{\sqrt{1-q}}\right)\right] \tag{4}$$

where Dx is the Gaussian measure and $H(x) = \int_x^{+\infty} Dy$. The typical overlap q is determined by the saddle-point equation associated with (4) and the critical capacity α_c is reached when $q \to 1$ [1, 4]

$$\alpha_c(\gamma) = \left(\int_0^{\gamma/2} Dx\ x^2 + \int_{\gamma/2}^{+\infty} Dx\ (\gamma - x)^2 \right)^{-1}. \tag{5}$$

We see in figure 1 that α_c is maximum ($\alpha_c \simeq 10.5$) for $\gamma_{\mathrm{rs}} \simeq 1.2$; such a value for the capacity is much higher than the monotonic perceptron one ($\alpha_c = 2$, see also [4]). In order to check the reliability of the RS result, one has to analyse the transverse stability of the symmetric saddle-point [1, 9]. It turns out [5] that the condition of stability ($\alpha_c(\gamma) < 2$) is never satisfied on the critical line $\alpha_c(\gamma)$ (when γ is finite), thus, as expected, the symmetry of the replicas must be broken.

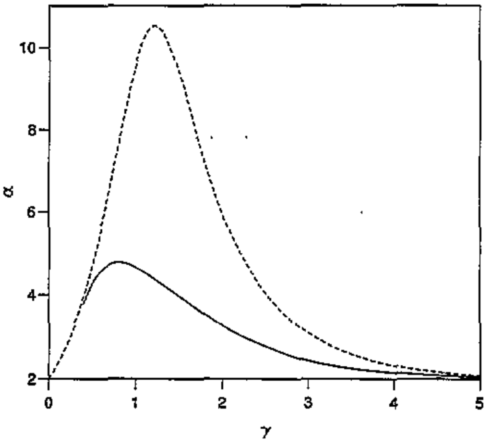

Figure 1. Critical capacity as a function of γ. The upper (dashed) curve is obtained within the RS assumption, whereas the lower one is the result of the one-step RSB computation. The maximum values of storage capacity are 10.5 and 4.8 respectively, corresponding to two different optimal choices of γ (1.2 for the RS case and 0.8 for the one-step RSB case).

Performing one step of RSB, f becomes a function of three parameters q_1, q_0 and m [6]. Being interested in the critical capacity, we focus on the limit $q_1 = 1 - \epsilon$, $\epsilon \to 0$, i.e. when the overlap between two vectors of couplings belonging to the same pure state goes to one. The saddle-point equations lead to the scaling relations $q_0 \to Q$ and $m/\epsilon \to \mu$; in particular we obtain $f_{\mathrm{rsb}}(q_1, q_0, m, \alpha) = (1/\epsilon)\ F(\mu, Q, \alpha) + \mathrm{O}(\ln \epsilon)$ where

$$F(\mu, Q, \alpha) = \frac{1}{2\mu}\ln[1 + \mu(1-Q)] + \frac{Q}{2(1+\mu(1-Q))} - \frac{\alpha}{\mu}\int_{-\infty}^{+\infty} Dx \ln \rho(x) \tag{6}$$

and

$$
\begin{aligned}
&\rho(x) = \left[p_-(x) + G_-(x) + p_0(x) + G_+(x) + p_+(x)\right]^{-1} \\
&p_-(x) = \int_{(\gamma/2 - x\sqrt{Q})/\sqrt{1-Q}}^{(\gamma - x\sqrt{Q})/\sqrt{1-Q}} Dy \exp(-\mu(\gamma - x\sqrt{Q} - y\sqrt{1-Q})^2/2) \\
&p_0(x) = \int_{-x\sqrt{Q}/\sqrt{1-Q}}^{(\gamma/2 - x\sqrt{Q})/\sqrt{1-Q}} Dy \exp(-\mu(x\sqrt{Q} + y\sqrt{1-Q})^2/2) \\
&p_+(x) = \int_{-\infty}^{-(\gamma + x\sqrt{Q})/\sqrt{1-Q}} Dy \exp(-\mu(\gamma + x\sqrt{Q} + y\sqrt{1-Q})^2/2) \\
&G_-(x) = H\left(\frac{\gamma - x\sqrt{Q}}{\sqrt{1-Q}}\right) \\
&G_+(x) = H\left(\frac{x\sqrt{Q}}{\sqrt{1-Q}}\right) - H\left(\frac{\gamma + x\sqrt{Q}}{\sqrt{1-Q}}\right).
\end{aligned}
\qquad (7)
$$

The saddle-points equations with respect to μ, Q and ϵ

$$
\left.\frac{\partial F}{\partial \mu}\right|_{\mu_c, Q_c, \alpha_c} = \left.\frac{\partial F}{\partial Q}\right|_{\mu_c, Q_c, \alpha_c} = F|_{\mu_c, Q_c, \alpha_c} = 0 \qquad (8)
$$

give the critical values μ_c, Q_c and α_c. The RS solution is recovered when $\mu_c = 0$ or $Q_c = 1$; discarding this trivial saddle-point, we have solved numerically the three coupled equations (8) and the resulting critical capacity is plotted in figure 1. The maximum $\alpha_c \simeq 4.8$ is reached for $\gamma_{\text{rsb}} \simeq 0.8$. It is worth noticing that the critical capacity we find is larger than the one ($\alpha_c = 4$) of the two hidden units parity machine.

The increase of α_c above the perceptron storage capacity may be understood using a geometrical approach due to Cover [10]. Each set of N synaptic couplings defines a dichotomy of the training set, labelled by the outputs of the P patterns. The number of such different dichotomies is $C_\gamma(P, N)$. Let us consider a new pattern 'A'. The dichotomies of the first P patterns may be divided in three different types:

(i) The $D_{\neq}$ dichotomies (type I) defining three hyperplanes ($\Delta = -\gamma, 0, \gamma$) such that none of them can include A.
(ii) The D_0 dichotomies (type II) defining three hyperplanes such that the central hyperplane ($\Delta = 0$) may contain A.
(iii) The D_γ dichotomies (type III) defining three hyperplanes such that the central hyperplane ($\Delta = 0$) cannot contain A but one of the two others ($\Delta = \pm\gamma$) may include A.

We obviously have $C_\gamma(P, N) = D_{\neq} + D_0 + D_\gamma$. Each type I dichotomy will provide only one dichotomy of the whole training set, whereas types II and III will give two dichotomies each. We obtain $C_\gamma(P+1, N) = D_{\neq} + 2D_0 + 2D_\gamma = C_\gamma(P, N) + D_0 + D_\gamma$. Type II dichotomies are equivalent to the dichotomies in the hyperplane including the origin O and orthogonal to the line (OA) and thus $D_0 = C_\gamma(P, N-1)$. In contrast to the usual perceptron case ($\gamma = 0$ or $\gamma = \infty$) we have the additional term $D_\gamma > 0$; therefore one may conclude that $\alpha_c \geqslant 2$ due to the larger number of dichotomies allowed by the separating hyperplanes.

As we have seen, RSB leads to a substantial modification (reduction) of α_c. In order to shed some light on the physical meaning of replica symmetry breaking, we have analysed

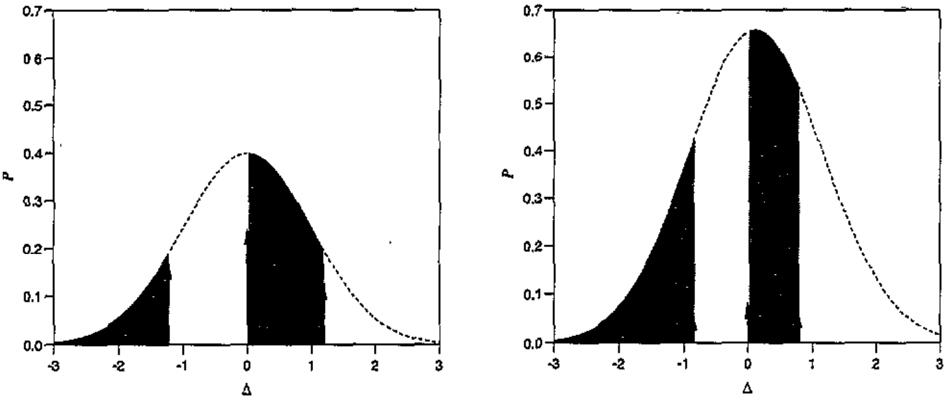

Figure 2. Distributions of the stabilities for the RS ansatz (a) and for one step of RSB (b), for $\gamma = \gamma_{rs}$ and γ_{rsb} respectively. The lengths of the arrows are equal to the Dirac weights p_0, p_+ and p_-.

how the network stores the patterns, i.e. the distribution of the stabilities (1), comparing the results with those found assuming unbroken symmetry. Within the RS ansatz, the distribution of the stabilities Δ is easily obtained on the critical line $\alpha_c(\gamma)$ [8]

$$\mathcal{P}_{rs}(\Delta) = p_- \ \delta(\Delta+\gamma) + p_0 \ \delta(\Delta) + p_+ \ \delta(\Delta-\gamma) + \theta_\gamma(\Delta)\ \frac{e^{-\Delta^2/2}}{\sqrt{2\pi}}$$
$$p_- = H\left(\frac{\gamma}{2}\right) - H(\gamma) \qquad p_0 = \frac{1}{2} - H\left(\frac{\gamma}{2}\right) \qquad p_+ = H(\gamma) \tag{9}$$

and the result is plotted in figure 2(a) for $\gamma = \gamma_{rs}$. One can notice three Dirac peaks, half of the patterns belonging to the three separating hyper-planes. The stabilities of the remaining patterns obey a Gaussian law with zero mean value. When $\gamma = \gamma_{rs}$, the fractions of patterns stored with negative and positive stabilities are equal to 0.27 and 0.73, and the values of the weights p_-, p_0 and p_+ are 0.16, 0.23 and 0.11 respectively. The latter stabilities distribution does not appear to be consistent since one can expect that in an N-dimensional space, the number of patterns satisfying a system of linear equations $\Delta^\mu = \pm\gamma, 0$ must be lower than N (i.e. we expect to find $(p_0 + p_+ + p_-)P \leqslant N$). For the RS distribution this is clearly not the case because, as we have seen, $p_0 + p_+ + p_- = 0.5$, implying that $0.5\alpha_c N \approx 5.25N$ patterns should be stored on the hyperplanes (it is worth remembering that for the simple perceptron one finds N such patterns).

Performing the same calculation with one step of RSB, we find

$$\begin{aligned}\mathcal{P}_{rsb}(\Delta) &= p_- \ \delta(\Delta+\gamma) + p_0 \ \delta(\Delta) + p_+ \ \delta(\Delta-\gamma) \\ &\quad + \theta_\gamma(\Delta) \int_{-\infty}^{+\infty} Dx \ \rho(x) \ \frac{\exp(-(\Delta + x\sqrt{q})^2/2(1-q))}{\sqrt{2\pi(1-q)}}\end{aligned} \tag{10}$$
$$p_{-,+,0} = \int_{-\infty}^{+\infty} Dx \ \rho(x) \ p_{-,+,0}(x).$$

Numerical computation of (10) leads to the results reported in figure 2(b); when $\gamma = \gamma_{rsb}$ we find for p_-, p_0 and p_+ 0.06, 0.08 and 0.06 respectively. The comparison with the

RS distribution (figure 2(*a*)) is straightforward and shows a reduction of the fraction of patterns memorized on the hyperplanes: such a new fraction becomes consistent with the geometrical argument discussed in that being $p_0 + p_+ + p_- \approx 0.2$, the number of patterns stored on the hyperplanes is $0.2\alpha_c N \approx 0.96N < N$.

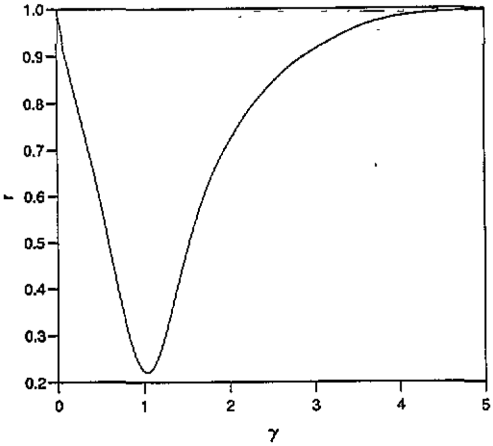

Figure 3. Typical overlap r between different domains defined by (13), as a function of γ. For $\gamma \to 0$ or ∞ the RS solution becomes exact and r tends to one. The value of γ that minimizes the overlap r ($\gamma \approx 1.1$ and $r \approx 0.22$) is close (but not equal) to the optimal value found in maximizing the critical capacity.

Let us now consider a vector $\boldsymbol{J}$ storing all the patterns. We can define the internal representation of the training set with respect to $\boldsymbol{J}$ as the P-component binary vector $\boldsymbol{R_J}$ such that

$$(\boldsymbol{R_J})^{\mu} = \mathrm{sign}(\sigma^{\mu}\boldsymbol{J}\cdot\boldsymbol{\xi}^{\mu}). \tag{11}$$

Replica symmetry breaking implies that there exist different disconnected domains $\mathcal{D}_i$ of synaptic weights, each of them storing all the patterns. As a result of the connectivity of each single domain, two couplings vectors $\boldsymbol{J}$ and $\boldsymbol{K}$ belonging to the same $\mathcal{D}_i$ give identical internal representations of the training set, i.e. $\boldsymbol{R_J} = \boldsymbol{R_K} = \boldsymbol{R}_i$. For a continuous normalization condition of the type $J_i \in [-1/\sqrt{N}; 1/\sqrt{N}]$, one may furthermore prove that two different domains $\mathcal{D}_i$ and $\mathcal{D}_j$ must store at least one identical pattern with two stabilities of opposite signs. Hence, the normalized overlap r_{ij} between their internal representations vectors

$$r_{ij} = \frac{1}{P}\,\boldsymbol{R}_i\cdot\boldsymbol{R}_j \tag{12}$$

is in this case strictly lower than 1 (with the classical constraint $\sum_i J_i^2 = 1$ used here, we have only $r_{ij} \leqslant 1$). Using a replica calculation, we have computed the mean overlaps $r_{\alpha\beta}$ between the domains belonging to two pure states α and β. With one step of breaking, on the critical line, $r_{\alpha\alpha} = 1$ and $r = r_{\alpha\beta}$ ($\alpha \neq \beta$) is given by

$$r = \int_{-\infty}^{+\infty} Dx\;\rho^2(x)\;\left[(p_0(x) + G_+(x) + p_+(x)) - (p_-(x) + G_-(x))\right]^2. \tag{13}$$

From the plot in figure 3, one notices that close to the optimal value of the threshold γ, there is a strong effect of symmetry breaking corresponding to a small overlap ($r \sim 0.2$) between different domains, while for very small or very large values of γ (i.e. when the network recovers the standard perceptron and the RS solution becomes exact) the mutual overlap of the various domains tends to one.

This work was partially supported by the *Annethe*-INFN Italian project.

References

[1] Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257
Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271

[2] Morita M, Yoshizawa S and Nakano H 1990 *IEICE Trans.* **J73-D-II** 232

[3] Yoshizawa S, Morita M and Amari S 1992 Capacity of associative memory using a non-monotonic neuron model *Preprint*

[4] Kobayashi K 1991 *Network* **2** 237

[5] Boffetta G, Monasson R and Zecchina R 1993 *Int. J. Neural Sys.* in press

[6] Mézard M, Parisi G and Virasoro M 1987 *Spin Glass Theory and Beyond* (Singapore: World Scientific)

[7] Barkai E, Hansel D and Kanter I 1990 *Phys. Rev. Lett.* **65** 2312

[8] Kepler T and Abbott L 1988 *J. Physique* **49** 1657

[9] de Almeida J and Thouless D 1978 *J. Phys. A: Math. Gen.* **11** 983

[10] Cover T 1965 *IEEE Trans.* **EC-14** 326