Preprint n. 994
January 28, 1994

Dipartimento di Fisica
Università di Roma "La Sapienza"

ROM P 994
SCW 9417

Rémi Monasson and Dominic O'Kane

**DOMAINS OF SOLUTIONS AND REPLICA SYMMETRY BREAKING IN
MULTILAYER NEURAL NETWORKS**

Domains of solutions and replica symmetry breaking in multilayer neural networks

Rémi Monasson ^{1,3} and Dominic O'Kane ^{2,4}

¹ Dipartimento di Fisica, Università di Roma "La Sapienza", 00185 Roma, Italy

² Theoretical Physics, 1 Keble Road, Oxford OX1 3NP, England

January 27, 1994

PACS numbers : 05.20 - Statistical Mechanics

64.60 - Order-Disorder and Statistical Mechanics of Model Systems

87.10 - General, Theoretical and Mathematical Biophysics

Abstract

The relationship between the geometrical structure of weight space and replica symmetry breaking (RSB) in multilayer neural networks is studied using a toy model. The distribution of sizes of the disconnected domains of solution space is computed analytically and compared to the RSB calculation of the Gardner volume. We are able to show explicitly that ergodicity breaking and RSB are not equivalent. Repeating these calculations using the cavity approach allows us to interpret the geometrical meaning of a RSB ansatz.

It is well-known [1] that a full study of the storage properties of multilayer neural networks within the Gardner framework [2] requires the breaking of replica symmetry [3]. This differs from the case of the simple perceptron, where a replica symmetric solution has been shown to be exact, and where the spin glass order parameter which emerges has a simple geometrical meaning. In multilayer neural networks, replica symmetry breaking is believed to reflect the existence of a more complicated weight space. However, a satisfactory description of this weight space and its relationship to the replica approach has yet to be established.

In this letter, we investigate the non monotonic perceptron [4], defined as a simple perceptron with normalised continuous weights J_i , $i = 1 \dots N$, with N binary inputs ξ_i , $i = 1 \dots N$, and a single binary output σ . A set of $P = \alpha N$ patterns $\{\xi^\mu, \sigma^\mu\}$ drawn independently and randomly from an unbiased distribution are correctly stored if for all $\mu = 1 \dots P$, the network $\mathbf{J}$ obeys

$$\Delta^\mu < -\gamma \quad \text{or} \quad 0 < \Delta^\mu < \gamma \quad (\gamma > 0) \quad (1)$$

³previous address : Laboratoire de Physique Théorique de l'ENS, 24 rue Lhomond, 75231 Paris Cedex 05, France (unité propre de recherche du CNRS rattachée à l'Université Paris-Sud)

⁴current address : Department of Mathematics, Imperial College, London SW7 2AZ, England

with $\Delta^\mu = \sigma^\mu \mathbf{J} \cdot \xi^\mu$. It has been shown [5] that this model is equivalent to a specific multilayer, namely a parity machine with three hidden units and appropriate thresholds. Earlier work has moreover shown that this model presents features common to all multilayer networks. The space of solutions, corresponding to the region of weight space where all P patterns are correctly stored, is disconnected. Secondly, the computation of the volume of this space of solutions

$$V_{sol} = \int d\mathbf{J} \rho(\mathbf{J}) \prod_{\mu=1}^P \Theta_\gamma(\Delta^\mu) \quad (2)$$

where $\Theta_\gamma(x) = 1$ if $x < -\gamma$ or $0 < x < \gamma$ and 0 otherwise, requires replica symmetry breaking.

We are interested in this model since unlike all multilayer networks studied to date it possesses a remarkable property : one can have a geometrical picture of the space of solutions as being composed of many disconnected domains. With $\rho(\mathbf{J})$ ensuring the constraint $\mathbf{J}^2 \leq 1$, it is possible to explicitly label each domain and to use this to compute analytically their geometrical characteristics (distribution of their sizes, total number and relative positions).

To see immediately how the network stores patterns, and how the space of solutions is disconnected, let us look at the simplest case, $P = 1$, when we are storing just one pattern. Figure 1a demonstrates that to store pattern ξ^1 , the weight vector $\mathbf{J}$ can lie in either of two disconnected domains each of which corresponds to satisfying one of the two inequalities in eqn. 1. When we present a new pattern ξ^2 , we see on figure 1b that the number of domains of synaptic vectors $\mathbf{J}$ which store both patterns correctly increases. The choice of constraint on $\mathbf{J}$ ensures that, for all P , each of these domains is labelled exclusively by one of the 2^P internal representation vectors $\tau = \{\tau^1, \dots, \tau^P\}$ where $\tau^\mu = \text{sign}(\Delta^\mu)$. Whilst 2^P is therefore the upper bound on the number of domains, we will see in the following that some of the domains may not exist, especially as P becomes large - in this case, some of the 2^P internal representation vectors are not realisable. We can now write explicitly the volume V_τ of the domain labelled by internal representation vector τ

$$V_\tau = \int_{\mathbf{J}^2 \leq 1} d\mathbf{J} \prod_{\mu=1}^P \Theta_\gamma(\Delta^\mu) \Theta(\tau^\mu \Delta^\mu) \quad (3)$$

where $\Theta(x)$ is the Heaviside function. One can verify that $\sum_\tau V_\tau = V_{sol}$, as required.

For $\alpha \geq 0$, correlations between different Δ^μ and hence different domains are negligible. One can then decouple the product over the P patterns in eqn. 3 to give

$$V_\tau \propto V_+^{\frac{1}{2}(P + \sum_\mu \tau^\mu)} V_-^{\frac{1}{2}(P - \sum_\mu \tau^\mu)} \quad \text{where} \quad V_+ = \frac{1}{2} - H(\gamma) \quad \text{and} \quad V_- = H(\gamma) \quad (4)$$

with $H(x) = \int_x^\infty \frac{dt}{\sqrt{2\pi}} \exp(-t^2/2)$. The range of domain volumes therefore lies between V_-^P and V_+^P . The number of domains with the same volume V_τ in this interval is simply $\binom{P}{\frac{1}{2}(P + \sum_\mu \tau^\mu)}$. This characterises the entire distribution of the sizes.

When α becomes finite, the correlations between different domains are no longer negligible and we resort to a method in which we extend an approach first used by [6]. It is then natural to define the "inverse size" k of a volume V_τ by $V_\tau = \exp(-Nk)$. The number of domains whose sizes lie in the range $k \rightarrow k + dk$ is defined as $\exp(Nc(k)).dk$. $c(k)$ may be seen as the

microcanonical entropy of the P Ising spin system τ whose Hamiltonian is $-\log(V_\tau)$. It is then clear that, in the thermodynamical limit $N \rightarrow \infty$, we can calculate $c(k)$ from a knowledge of the “free-energy” at inverse “temperature” R

$$g(R) = -\frac{1}{NR} \overline{\log \left(\sum_{\tau} V_{\tau}^R \right)} \quad (5)$$

using the Legendre transform identities $k = \partial(Rg(R))/\partial R$ and $c(k) = -\partial g(R)/\partial(1/R)$. The overbar denotes the average over the random pattern disorder and we assume that $c(k)$, being an extensive quantity, is self-averaging [3]. To perform the average, one makes use of the replica trick $\overline{\log x} = \lim_{n \rightarrow 0} (\overline{x^n} - 1)/n$. We do the calculation for R integer expecting that the final result is valid for any real value of R [7]. There are n blocks ($a = 1 \dots n$) of R replicas ($\rho = 1 \dots R$), and thus, the spin glass order parameter $q_{\rho\nu}^{ab}$ is the typical overlap between two weight vectors $\mathbf{J}_\rho^a$ and $\mathbf{J}_\nu^b$. Here the *replica symmetric* Ansatz, whose validity will be discussed later, becomes

$$q_{\rho\nu}^{ab} = \begin{cases} Q & \text{if } a = b \text{ and } \rho = \nu \\ q_1 & \text{if } a = b \text{ and } \rho \neq \nu \\ q_0 & \text{if } a \neq b. \end{cases} \quad (6)$$

Q is simply the squared magnitude of the weight vector, q_1 is the typical overlap between two weight vectors in the same domain (they have the same internal representation vector), q_0 is the typical overlap between two weight vectors each in a different domain (they have different internal representation vectors). We find

$$g(R) = \text{Extr}_{Q, q_1, q_0} \left\{ -\frac{1}{2}(1 + \log 2\pi) - \frac{q_0}{2(Q - q_1 + R(q_1 - q_0))} - \frac{R-1}{2R} \log(Q - q_1) - \frac{1}{2R} \log(Q - q_1 + R(q_1 - q_0)) - \frac{\alpha}{R} \int Dx \log \left[\int Dy (H_+^R + H_-^R) \right] \right\} \quad (7)$$

where $Dx = \frac{\exp(-x^2/2)}{\sqrt{2\pi}} dx$ and

$$H_+ = H \left(\frac{x\sqrt{q_0} + y\sqrt{q_1 - q_0}}{\sqrt{Q - q_1}} \right) - H \left(\frac{\gamma + x\sqrt{q_0} + y\sqrt{q_1 - q_0}}{\sqrt{Q - q_1}} \right) ; \quad H_- = H \left(\frac{\gamma - x\sqrt{q_0} - y\sqrt{q_1 - q_0}}{\sqrt{Q - q_1}} \right)$$

from which we have computed the distribution curves shown in figure 2 for different values of α in the case $\gamma = 1$. The most obvious trend is that as the number of stored patterns increases, the number of domains $\exp(Nc(k))$ of a given size k decreases, indicating that domains are shrinking and may possibly vanish. The smallest domains are progressively eliminated by the addition of new patterns until at a certain value α_s , the smallest domains have a vanishing volume and are in fact those which are most numerous.

As $R \rightarrow 0$ (the top of the curves), $c(k_0) = \frac{1}{N} \log \mathcal{N}$, where $\mathcal{N}$ is the total number of domains. Figure 3 shows the dependence of $c(k_0)$ on α for two values of γ . We see two distinct regimes. For $\alpha < \alpha_s$, $c(k_0) = \alpha \log 2$ signifying that $\mathcal{N} = 2^P$ (discarding polynomial factors). All possible domains exist. In other words, for each possible choice of τ it is possible to find a $\mathbf{J}$ to store all ξ^μ with internal representation τ^μ ($V_\tau > 0$). When $\alpha > \alpha_s$, $c(k_0)/\alpha$ decreases with increasing α and becomes zero for some value α_d - there exist some internal representation vectors τ which are not realisable by the network ($V_\tau = 0$). To a numerical precision of 0.01 %, we found that

α_d coincides with the storage capacity α_c computed with one-step of RSB [5]. The mathematical source of this close agreement can be seen by comparing the equation for $g(R)$ when $R \rightarrow 0$ to the one-step replica symmetry broken expression for the logarithm of the Gardner volume

$$\frac{1}{N} \overline{\log V_{sol}} = \text{Extr}_{Q, q_1, q_0, m} \left\{ \frac{1}{2}(1 + \log 2\pi) + \frac{q_0}{2(Q - q_1 + m(q_1 - q_0))} + \frac{m-1}{2m} \log(Q - q_1) + \frac{1}{2m} \log(Q - q_1 + m(q_1 - q_0)) + \frac{\alpha}{m} \int Dx \log \left[\int Dy (H_+ + H_-)^m \right] \right\} \quad (8)$$

in the limit $\alpha \rightarrow \alpha_c$. We will see later how the cavity approach [3, 8] can explain in geometrical terms the reason for the similarity between expressions (7) and (8).

Let us now see how this description of the space of solution relates to replica symmetry breaking. The volume of solutions V_{sol} is dominated by domains of size k_1 corresponding to the point of the curve $c(k)$ where the slope $R = 1$ (marked by bold dots on figure 2). Thus, $c(k_1)$ measures the number of domains contributing to V_{sol} . As a consequence, $c(k_1) \geq 0$ and the value α_{ZE} where $c(k_1)$ equals zero is an upper bound for the size of the training set α_{RSB} where the symmetry breaking occurs. The existence of such a criterion is new for neural networks with continuous couplings (and strongly reminiscent of the perceptron with binary weights [9]). One can compute α_{RSB} as the value of α for which the saddle point of eqn. (8) ceases to be replica symmetric. There exists a tricritical value of $\gamma_T \simeq 2$ above which α_{RSB} coincides with the de Almeida-Thouless line α_{AT} [10] and below which the breaking of symmetry is of first order (i.e. the intra pure state overlap q_1 jumps discontinuously at α_{RSB}). For γ small, we found that the difference between α_{RSB} and α_{ZE} was very small e.g. when $\gamma = 1$, $\alpha_{RSB} \simeq \alpha_{ZE} = 3.0 \pm 0.05$ while $\alpha_{AT} = 6.4$ and the one-step capacity $\alpha_c = 4.8$. However, this difference increased with increasing γ (they obviously differ from each other when $\gamma > \gamma_T$ since the replica symmetric saddle point giving $c(k_1)$ is unstable against replicon modes).

When $\alpha < \alpha_{RSB}$, one is therefore in the following situation. The replica symmetric assumption is correct and, at the same time, $c(k_1) > 0$, showing that the number of domains contributing to V_{sol} is exponential. This proves that disconnectivity of the space of solutions (or, in other terms, breaking of ergodicity) and replica symmetry breaking are not equivalent. Such a phenomenon has already been found for other disordered systems using a different approach [11]. It may be interpreted as follows. Although there is a huge number of domains, when α is small (i.e. less than α_{RSB}), none of them make a finite contribution to V_{sol} : they all have vanishing weights in the thermodynamical limit and the replica approach cannot distinguish them. Our calculation of the distribution $c(k)$ therefore provides some information which is usually hidden in the computation of V_{sol} . The fact that $\alpha_{RSB} \simeq \alpha_{ZE}$ suggests that, roughly speaking, replica symmetry breaking occurs when the number of domains no longer scales exponentially with N . It probably occurs at exactly the moment when the number of pure states, which would be similar to the number of domains, no longer scales exponentially with N .

To end with, let us see how the cavity approach [3] may provide a geometrical picture of the RSB ansatz and explain how it relates to our computation of $g(R)$. After having stored patterns ξ^μ , $\mu = 1 \dots P$, the space of solutions is composed of disjoint domains V_τ centered around the mean weights $\langle \mathbf{J} \rangle_\tau$. We now present a new pattern $\{\xi^0, \sigma^0\}$ whose stability is $\Delta^0 = \sigma^0 \mathbf{J} \cdot \xi^0$. Keeping in mind that the one step calculation of the volume of solutions V_{sol} is equivalent to

the one step calculation of the free-energy of the τ system at inverse temperature $R = 1$, one can easily show from [3, 8] that the underlying geometrical hypothesis of RSB are :

- A pure state ν consists of an exponential number of domains V_τ whose typical size is k_ν . The volume of solutions inside the pure state ν is e^{-Nf_ν} where $f_\nu = k_\nu - c(k_\nu)$ is the free-energy of the same pure state for the τ spins system at inverse temperature $R = 1$. The number of domains having sizes lying in $k \rightarrow k + dk$ is $e^{-Nf_\nu} e^{Nk}$. Using the clustering properties of the weights, Δ^0 is a gaussian variable of mean $\Delta_\nu = \sigma^0 < J >_\nu$, ξ^0 and variance $Q - q_1 = < J^2 >_\nu - < J >_\nu^2$. The addition of the new pattern has thus made V_{sol} decrease by a factor $\int d\Delta^0 \mathcal{P}(\Delta^0) \Theta_\tau(\Delta^0) = H\left(\frac{\Delta_\nu}{\sqrt{Q-q_1}}\right) - H\left(\frac{\gamma+\Delta_\nu}{\sqrt{Q-q_1}}\right) + H\left(\frac{\gamma-\Delta_\nu}{\sqrt{Q-q_1}}\right)$.
- Inside the cluster Γ (which is unique to the first order of RSB), the number of pure states ν of free-energies lying in the range $f \rightarrow f + df$ (i.e. the number of set of domains of total volume e^{-Nf}) follows the exponential law $e^{Nm(f-f_\Gamma)} df$ where f_Γ is the free-energy of the whole cluster. Moreover, the stability Δ_ν fluctuates from pure state to pure state with a mean $\Delta_\xi = \sigma^0 < J >_\Gamma$, ξ^0 and a variance $q_1 - q_0 = < \Delta_\nu^2 >_\Gamma - < \Delta_\nu >_\Gamma^2$. This allows the computation of the new distribution of the free-energies of pure states storing all $P + 1$ patterns.
- The “cluster” stability Δ_ξ fluctuates from sample to sample with a zero mean and variance q_0 . Averaging $\frac{1}{N} \log V_{sol} = -f_\Gamma$ over this disorder, one finds back eqn. (8).

Now, let us consider the physical significance of the replica symmetric ansatz used to compute the distribution of the sizes of the domains :

- Inside one domain V_τ , weights may vary around $< J >_\tau$, inducing a gaussian distribution $\mathcal{P}$ for Δ^0 of mean $\Delta_\tau = \sigma^0 < J >_\tau$, ξ^0 and variance $Q - q_1 = < J^2 >_\tau - < J >_\tau^2$. The two possible internal representations $\{\tau, -\}$ and $\{\tau, +\}$ of the whole set of patterns correspond thus to the volumes $V_\tau \times \int_{-\infty}^{\gamma} d\Delta^0 \mathcal{P}(\Delta^0)$ and $V_\tau \times \int_{\gamma}^{\infty} d\Delta^0 \mathcal{P}(\Delta^0)$ respectively. As a consequence of the addition of a new pattern, the R th power of the volume V_τ^R is multiplied by the factor $\left[H\left(\frac{\Delta_\tau}{\sqrt{Q-q_1}}\right) - H\left(\frac{\gamma+\Delta_\tau}{\sqrt{Q-q_1}}\right)\right]^R + \left[H\left(\frac{\gamma-\Delta_\tau}{\sqrt{Q-q_1}}\right)\right]^R$.
- Because of the broad scattering of domains throughout the whole of the sphere $J^2 \leq 1$, the stability Δ_τ fluctuates from domain to domain. In the thermodynamic limit and for a given τ , it becomes a gaussian variable. Its mean Δ_ξ depends only on the disorder and its variance is $q_1 - q_0$ where q_0 is the typical overlap between two coupling vectors belonging to two different domains. The previous shrinkage ratio must be averaged over these domain to domain fluctuations.
- Last of all, Δ_ξ fluctuates from sample to sample with a zero mean and variance q_0 . Averaging $\log \sum_\tau V_\tau^R$ over this disorder, one finds back eqn. (7).

For completeness sake, we have checked the AT [10] stability of our RS Ansatz (6). The replicon modes give four distinct eigenvalues whose multiplicities are $1, R-1, R-1$ and $(R-1)^2$. As expected, our solution is always stable at low α . Increasing α , we find that an unstable region exists at $R < 0$ and progressively includes the whole curve. However, this does not alter

qualitatively our conclusions since these are concerned more with understanding the relationship between two different replica approaches than the derivation of exact results.

In this context, the importance of the cavity method is that it provides access to the concept of the pure state. Using it to compare these two approaches, we have been able to demonstrate explicitly that a pure state corresponds to a region of configuration space composed of an exponential number of domains of one particular dominant size. Despite the fact that the remaining domains are neglected, this simplification of solution space does not prevent us from deriving exact results. Indeed, as we have seen in this model at low α , one has a non-ergodic configuration space which has an exact RS solution. This is encouraging conclusion since it means that one may still be able to compute exact results even when one knows that the space of solutions is disconnected.

To conclude, the model studied here has provided us with a geometrical picture of the replica approach. In doing so, it has clarified the important difference between the uneasy concept of a pure state and the more natural notion of a domain of solutions.

We suggest that the approach described here may be extended to provide insight into the behaviour of other less “toy-like” multilayer neural networks such as the parity and committee machines.

We wish to thank M. Mézard for his valuable insights and encouragement. We also thank S. Franz, R. Penney, D. Sherrington and M. Virasoro for stimulating discussions. This study was partly funded by the Oxford-Paris exchange program, the EEC, DENI and the INFN.

References

- [1] E. Barkai, D. Hansel, I. Kanter *Phys. Rev. Lett.* **65**, 18 (1990)
E. Barkai, D. Hansel, H. Sompolinsky *Phys. Rev. A* **45**, 4146 (1992)
A. Engel, H.M. Kohler, F. Tschepe, H. Vollmayr, A. Zippelius *Phys. Rev. A* **45**, 7590 (1992)
- [2] E. Gardner *J. Phys. A* **21**, 257 (1988)
E. Gardner, B. Derrida *J. Phys. A* **21**, 271 (1988)
- [3] M. Mézard, G. Parisi, M.A. Virasoro *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987)
- [4] K. Kobayashi *Network* **2**, 237 (1991)
- [5] G. Boffetta, R. Monasson, R. Zecchina *J. Phys. A* **26**, L507 (1993)
- [6] B. Derrida, R.B. Griffiths, A. Prugel-Bennet *J. Phys. A* **24**, 4907 (1991)
- [7] G. Parisi, M.A. Virasoro *J. Phys. France* **50**, 3317 (1989)
- [8] M. Mézard *J. Phys. A* **22**, 2181 (1989)
- [9] W. Krauth, M. Mézard *J. Phys. France* **50**, 3057 (1989)

[10] J. de Almeida, D. Thouless *J. Phys. A* 11, 983 (1978)

[11] J. Kurchan, G. Parisi, M.A. Virasoro *J. Phys. I France* 3, 1819 (1993)

Figure Captions

figure 1 : how the networks stores one (fig. 1a) and two (fig. 1b) patterns. The shaded regions denote the regions of J space where both patterns can be correctly stored. For $P > 1$, they are disconnected. Each domain has been labelled by its corresponding internal representation vector (see text).

figure 2 : the entropy $c(k)$ of domains having size k for $\gamma = 1$ and $\alpha = 0, 0.1, 0.5, 1, 2, 3, 4$. The horizontal lines represent the asymptotic limits $c(k_0)$ of $c(k)$ as $R \rightarrow 0$ for the different values of α . The bold marks are the points $(k_1, c(k_1))$ of slope $R = 1$. Note that the curve corresponding to $\alpha = 0$ should be rescaled according to $k \rightarrow k/\alpha$ to find back the distribution (4) (see text).

figure 3 : the logarithm of the total number of domains $c(k_0)/\alpha$ for $\gamma = 1, 2$ as a function of α . Below α_s , all the 2^P domains exist while above α_s , only an exponentially small fraction of them remain. The total number of domains ceases to be exponential at a pattern loading $\alpha \simeq \alpha_c$ (computed with the one-step RSB ansatz).

Fig. 1

Fig. 2

Fig 3