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

July 28, 1994

PACS numbers : 05.20 - 64.60 - 87.10

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, $i = 1...N$, with N binary inputs ξ_i, $i = 1...N$, and a single binary output σ. A set of $P = \alpha N$ patterns $\{\boldsymbol{\xi}^\mu, \sigma^\mu\}$ drawn independently and randomly from an unbiased distribution are correctly stored if for all $\mu = 1...P$, the network $\mathbf{J}$ obeys

$$\Delta^\mu < -\gamma \qquad \text{or} \qquad 0 < \Delta^\mu < \gamma \qquad\qquad (\Delta^\mu = \sigma^\mu \mathbf{J}.\boldsymbol{\xi}^\mu) \tag{1}$$

[3] previous 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)

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

with $\gamma > 0$. 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

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

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).

As soon as we store one pattern $\boldsymbol{\xi}^1$, the space of solutions is indeed disconnected. Figure 1a demonstrates that 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 $\boldsymbol{\xi}^2$, 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 $\boldsymbol{\tau} = \{\tau^1, ..., \tau^P\}$ where $\tau^\mu = \text{sign}(\Delta^\mu)$. Whilst 2^P is therefore the upper bound on the number of domains, 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 $V_{\boldsymbol{\tau}}$ of the domain labelled by internal representation vector $\boldsymbol{\tau}$

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

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

For $\alpha \simeq 0$, 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

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

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 $V_{\boldsymbol{\tau}}$ in this interval is simply $\binom{P}{\frac{1}{2}\left(P+\sum_\mu \tau^\mu\right)}$. This characterises the entire distribution of the sizes.

When α becomes finite, the correlations between different domains are no longer negligible. It is then natural to define the "inverse size" k of a volume $V_{\boldsymbol{\tau}}$ by $V_{\boldsymbol{\tau}} = \exp(-Nk)$ [6]. The number of domains whose sizes lie in the range $k \to k + dk$ is defined as $\exp(Nc(k)).dk$. $c(k)$ may be seen as the microcanonical entropy of the P Ising spin system $\boldsymbol{\tau}$ whose Hamiltonian is $-\log(V_{\boldsymbol{\tau}})$. In the thermodynamical limit $N \to \infty$, we can calculate $c(k)$ from a knowledge of the "free-energy" at inverse "temperature" R

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

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 $\overline{\log x} = \lim_{\,n\to 0}(\overline{x^n} - 1)/n$. 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 ($a = 1...n$) of R replicas ($\rho = 1...R$), 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 becomes $q_{\rho\nu}^{ab} = Q$ if $a = b$ and $\rho = \nu$, q_1 if $a = b$ and $\rho \neq \nu$, q_0 if $a \neq b$. 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

$$\begin{aligned} g(R) &= \operatorname*{Extr}_{Q,q_1,q_0}\left\{-\frac{1}{2}(1+\log 2\pi) - \frac{q_0}{2(Q-q_1+R(q_1-q_0))} - \frac{R-1}{2R}\log(Q-q_1)\right. \\ &- \left.\frac{1}{2R}\log(Q-q_1+R(q_1-q_0)) - \frac{\alpha}{R}\int Dx\ \log\left[\int Dy\ \left(H_+^R + H_-^R\right)\right]\right\} \end{aligned} \tag{6}$$

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

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

from which we have computed the distribution curves shown in figure 2 for different values of α in the case $\gamma = 1$ [8]. 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 $R \to 0$ (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. For $\alpha < \alpha_s$, $c(k_0) = \alpha\log 2$ signifying that $\mathcal{N} = 2^P$ (discarding polynomial factors). All possible domains exist – have a finite volume. When $\alpha > \alpha_s$, $c(k_0)/\alpha$ decreases with increasing α and becomes zero for some value α_d – domains begin to disappear, i.e. for some $\boldsymbol{\tau}$, $V_{\boldsymbol{\tau}} = 0$. 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 $R \to 0$ to the one-step replica symmetry broken expression for the logarithm of the Gardner volume

$$\begin{aligned} \frac{1}{N}\overline{\log V_{sol}} &= \operatorname*{Extr}_{Q,q_1,q_0,m}\left\{\frac{1}{2}(1+\log 2\pi) + \frac{q_0}{2(Q-q_1+m(q_1-q_0))} + \frac{m-1}{2m}\log(Q-q_1)\right. \\ &+ \left.\frac{1}{2m}\log(Q-q_1+m(q_1-q_0)) + \frac{\alpha}{m}\int Dx\ \log\left[\int Dy\ \left(H_+ + H_-\right)^m\right]\right\} \end{aligned} \tag{7}$$

in the limit $\alpha \to \alpha_c$. We will see later how the cavity approach [3, 9] can explain in geometrical terms the reason for the similarity between expressions (6) and (7).

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 points on figure 2). Thus, $c(k_1)$ is the logarithm of the number of domains contributing to V_{sol} and must be positive. 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. One can compute α_{RSB} as the value of α for which the saddle point of eqn. (7) ceases to be replica symmetric. We found that for small γ, the difference between α_{ZE} and α_{RSB} 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 $\alpha_c = 4.8$ and that this difference increased with increasing γ, indicating that the exact calculation of V_{sol} probably requires more than one step of breaking.

When $\alpha < \alpha_{RSB}$, 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 [11]. Although there is a huge number of domains, when α is 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}.

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 $\boldsymbol{\xi}^\mu$, $\mu = 1...P$, the space of solutions is composed of disjoint domains $V_{\boldsymbol{\tau}}$ centered around the mean weights $< \boldsymbol{J} >_{\boldsymbol{\tau}}$. We now present a new pattern $\{\boldsymbol{\xi}^0, \sigma^0\}$ whose stability is $\Delta^0 = \sigma^0 \boldsymbol{J}.\boldsymbol{\xi}^0$. 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 $\boldsymbol{\tau}$ system at inverse temperature $R = 1$, one can easily show from [3, 9] that the underlying geometrical hypothesis of RSB are :

- A pure state ν consists of an exponential number of domains $V_{\boldsymbol{\tau}}$ 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 $\boldsymbol{\tau}$ spins system at inverse temperature $R = 1$. The number of domains having size lying in $k \to k + dk$ is $e^{-Nf_\nu} e^{Nk}$. Using the clustering properties of the weights, Δ^0 is a Gaussian variable of mean $\Delta_\nu = \sigma^0 < \boldsymbol{J} >_\nu .\boldsymbol{\xi}^0$ and variance $Q - q_1$. The addition of the new pattern has thus made V_{sol} decrease by a factor $\int d\Delta^0 \mathcal{P}(\Delta^0) \Theta_\gamma(\Delta^0) = H\left(\frac{\Delta_\nu}{\sqrt{Q-q_1}}\right) - H\left(\frac{\gamma+\Delta_\nu}{\sqrt{Q-q_1}}\right) + H\left(\frac{\gamma-\Delta_\nu}{\sqrt{Q-q_1}}\right)$.

- The number of pure states ν of free-energies lying in the range $f \to f + df$ (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 cluster Γ of pure states [3]. Moreover, Δ_ν fluctuates from pure state to pure state with a mean $\Delta_{\boldsymbol{\xi}} = \sigma^0 < \boldsymbol{J} >_\Gamma .\boldsymbol{\xi}^0$ and a variance $q_1 - q_0$.

- The "cluster" stability $\Delta_{\boldsymbol{\xi}}$ 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. (7).

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 $V_{\boldsymbol{\tau}}$, weights may vary around $< \boldsymbol{J} >_{\boldsymbol{\tau}}$, inducing a Gaussian distribution $\mathcal{P}$ for Δ^0 of mean $\Delta_{\boldsymbol{\tau}} = \sigma^0 < \boldsymbol{J} >_{\boldsymbol{\tau}} .\boldsymbol{\xi}^0$ and variance $Q - q_1$. The two possible internal representations $\{\boldsymbol{\tau}, -\}$ and $\{\boldsymbol{\tau}, +\}$ of the whole set of patterns correspond thus to the volumes $V_{\boldsymbol{\tau}} \times \int_{-\infty}^{-\gamma} d\Delta^0 \mathcal{P}(\Delta^0)$ and $V_{\boldsymbol{\tau}} \times \int_0^{\gamma} d\Delta^0 \mathcal{P}(\Delta^0)$ respectively. The Rth power of the volume $V_{\boldsymbol{\tau}}^R$ is multiplied by the factor $\left[H\left(\frac{\Delta_{\boldsymbol{\tau}}}{\sqrt{Q-q_1}}\right) - H\left(\frac{\gamma+\Delta_{\boldsymbol{\tau}}}{\sqrt{Q-q_1}}\right)\right]^R + \left[H\left(\frac{\gamma-\Delta_{\boldsymbol{\tau}}}{\sqrt{Q-q_1}}\right)\right]^R$.

- Because of the broad scattering of domains throughout the whole of the sphere $\boldsymbol{J}^2 \leq 1$, the stability $\Delta_{\boldsymbol{\tau}}$ fluctuates with a mean $\Delta_{\boldsymbol{\xi}}$ and a variance $q_1 - q_0$ 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, $\Delta_{\boldsymbol{\xi}}$ fluctuates from sample to sample with a zero mean and variance q_0. Averaging $\log \sum_{\boldsymbol{\tau}} V_{\boldsymbol{\tau}}^R$ over this disorder, one finds back eqn. (6).

We explicitly see that a pure state corresponds to a region of configuration space composed of an exponential number of domains of one particular dominant size. Though remaining domains are neglected, this simplification 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. It is an 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.

The model studied here has therefore 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 encouragment. 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. Tschepke, 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] Checking the local stability of our solution [3], we find that it is always stable at low α becoming unstable at large α beginning with the region $R \ll 0$. However, as our conclusions are qualitative, concerning an understanding of the relationship between two replica approaches, this is not a serious limitation.

[9] M. Mézard *J. Phys. A* **22**, 2181 (1989)

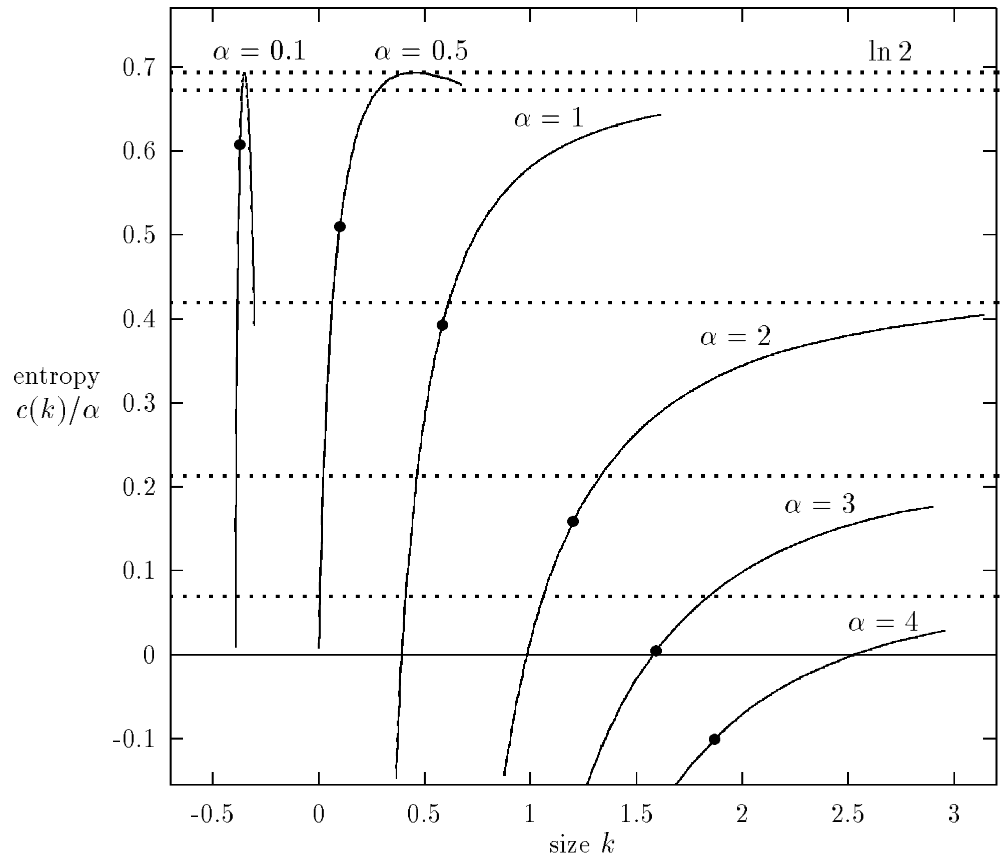
[10] D.Thirumalai, T.R. Kirkpatrick *Phys. Rev. A* **37**, 4439 (1988)
J. Kurchan, G. Parisi, M.A. Virasoro *J. Phys. I France* **3**, 1819 (1993)

Figure Captions

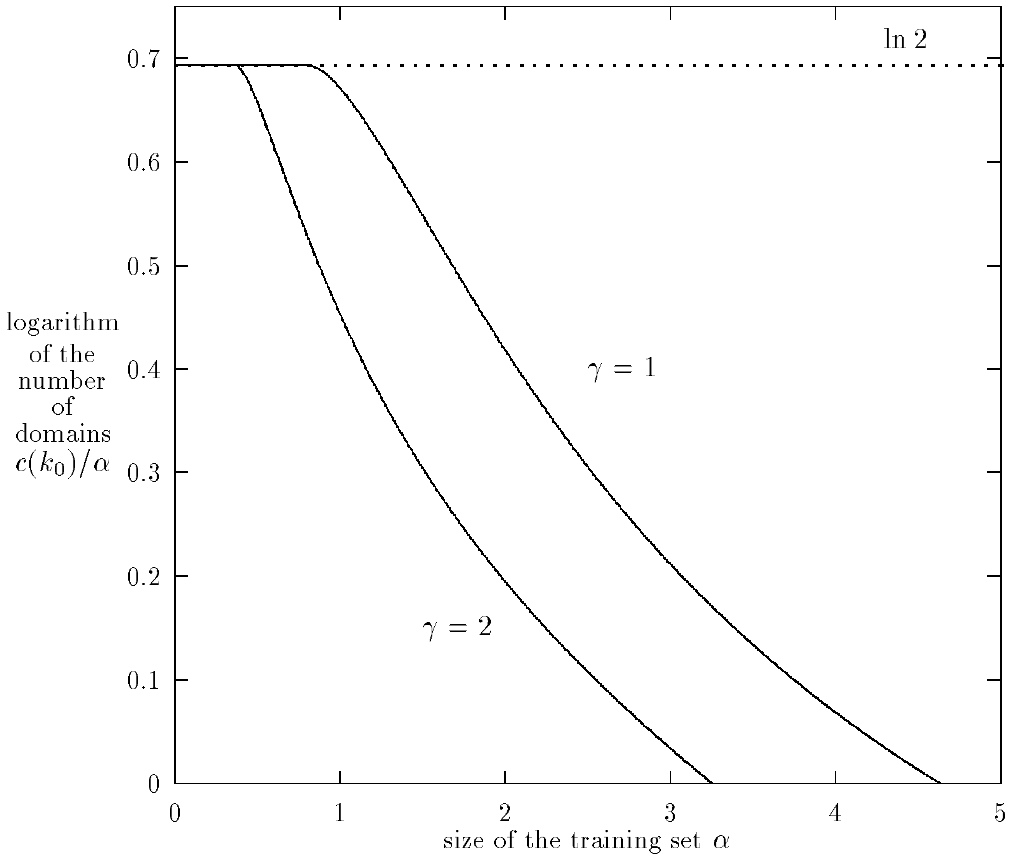
figure 1 : how the networks stores one (fig. 1a) and two (fig. 1b) patterns. The shaded regions denote the regions of $\boldsymbol{J}$ space where patterns are correctly stored. Each suitable domain has been labelled by its corresponding internal representation vector (see text).

figure 2 : the entropy $c(k)/\alpha$ 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 $R \to 0$ for the different values of α. The bold points are the points $(k_1, c(k_1))$ of slope $R = 1$.

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



$\alpha = 0.1$
$\alpha = 0.5$
$\ln 2$
$\alpha = 1$
$\alpha = 2$
$\alpha = 3$
$\alpha = 4$
entropy $c(k)/\alpha$
size k



$\ln 2$
$\gamma = 1$
$\gamma = 2$
logarithm of the number of domains $c(k_0)/\alpha$
size of the training set α