Weight Space Structure and Internal Representations : a Direct Approach to Learning and Generalization in Multilayer Neural Networks

Rémi Monasson* and Riccardo Zecchina†

* *INFN and Dipartimento di Fisica, P.le Aldo Moro 2, I-00185 Roma, Italy*

† *INFN and Dip. di Fisica, Politecnico di Torino, C.so Duca degli Abruzzi 24, I-10129 Torino, Italy*

Abstract

We analytically derive the geometrical structure of the weight space in multilayer neural networks (MLN), in terms of the volumes of couplings associated to the internal representations of the training set. Focusing on the parity and committee machines, we deduce their learning and generalization capabilities both reinterpreting some known properties and finding new exact results. The relationship between our approach and information theory as well as the Mitchison–Durbin calculation is established. Our results are exact in the limit of a large number of hidden units, showing that MLN are a class of exactly solvable models with a simple interpretation of replica symmetry breaking.

PACS Numbers : 05.20 - 64.60 - 87.10

Typeset using REVTEX

Memorization, rule inference or information processing by a neural network may be seen as a complicated selection of one part of its whole weight space [1,2]. Statistical mechanics has permitted a quantitative study of this selection process for the simple perceptron by providing a measure on the weight space resulting from learning [2]. In particular, the purely geometrical meaning of the spin-glass order parameter [3] has been shown to emerge naturally in this context. These techniques have been successfully applied to simple models of multilayer neural networks (MLN) to compute their storage capacities and generalization errors [4–6]. However, a geometrical picture of MLN's weight space and thus a unique "conceptual" frame allowing for the interpretation of the physical and computational behaviour is lacking so far.

In this letter we analytically derive such geometrical structure for MLN and show how it is hidden in the usual Gardner's approach. The study of the distribution of volumes of couplings associated to the internal representations of the training set, leads to a simple geometrical interpretation of replica symmetry breaking (RSB) and allows to deduce the networks learning and generalization properties. Moreover, we show the key importance of the issue for analyzing the encoding of information provided by the *internal representations* [4] in the intermediate layers of MLN by establishing a correspondence with information theory and the Mitchison–Durbin calculation [7]. For the storage problem, we focus upon the volumes giving the dominant contribution to Gardner's total volume, whose number $\mathcal{N}_D$ is smaller than the total number $\mathcal{N}_R$ of non–empty volumes. For the parity and committee machines with $K(\gg 1)$ hidden units, $\mathcal{N}_D$ and $\mathcal{N}_R$ both vanish at $\frac{\log K}{\log 2}$ and $\frac{16}{\pi}\sqrt{\log K}$ (so far unknown) respectively. Our results are shown to be exact in this limit and are likely to coincide with the storage capacities of both machines. For finite K, we give a general geometrical interpretation of RSB together with numerical results in the case $K = 3$. The inference of a learnable rule is studied along the same lines. We first reinterpret recent results [6] concerning the Bayesian learning of a rule by a parity machine. We then explain the smoothness of the generalization curve of the committee machine near its Vapnik–Chervonenkis (VC) dimension [9] $d_{vc} \sim \sqrt{\log K}$ and conjecture a cross-over to lower generalization error for

$\alpha \sim K$.

In the following, we shall consider tree-like MLN, composed of K non-overlapping perceptrons with real-valued weights $J_{\ell i}$ and connected to K sets of independent inputs $\xi_{\ell i}$ ($\ell = 1, ..., K$, $i = 1, ..., N/K$) [4]. The output σ of the network is a binary function $f(\tau_1, ..., \tau_K)$ of the cells $\tau_\ell = \mathrm{sign}(\sum_i J_{\ell i}\xi_{\ell i})$ in the first hidden layer. The set $\{\tau_\ell\}$ will be called hereafter *internal representation* of the input pattern $\{\xi_{\ell i}\}$. For the parity and committee machines, the decoder functions f are respectively $\prod_\ell \tau_\ell$ and $\mathrm{sign}(\sum_\ell \tau_\ell)$. The training set to be stored in the network includes $P = \alpha N$ patterns $\{\xi_{\ell i}^\mu\}$ and their corresponding outputs σ^μ ($\mu = 1, ..., P$). For simplicity, both patterns and outputs are drawn according to the binary unbiased distribution law. In order to store the patterns, one must find a suitable set of internal representations $\mathcal{T} = \{\tau_\ell^\mu\}$ with a corresponding non zero volume

$$V_\mathcal{T} = \int \prod_{\ell,i} dJ_{\ell i} \prod_\mu \theta\left(\sigma^\mu f(\{\tau_\ell^\mu\})\right) \prod_{\mu,\ell} \theta\left(\tau_\ell^\mu \sum_i J_{\ell i}\xi_{\ell i}^\mu\right) \tag{1}$$

where $\theta(\ldots)$ is the Heaviside function and the integral over the weights fulfills $\int \prod_{\ell,i} dJ_{\ell i} = 1$. Gardner's total volume is simply $V_G = \sum_\mathcal{T} V_\mathcal{T}$ and the critical capacity of the network is the value α_c of the maximal size of the training set such that $\overline{\log V_G}$ is finite, where the bar denotes the average over the patterns and their corresponding outputs [2]. Moreover, the partition of V_G into connected components may be naturally obtained using the $V_\mathcal{T}$'s as elementary "bricks". Indeed, from definition (1), the set of weights $\{J_{\ell i}\}$ contributing to a given $V_\mathcal{T}$ is convex (or empty). For the parity machine, two volumes corresponding to two adjacent set of internal representations (i.e. differing for one single τ_ℓ^μ) cannot coexist (they would give opposite outputs for the pattern μ) and one of them at least must be empty. Thus each connected component of V_G coincides with one and only one volume associated to an internal representation. For the committee machine, a connected component of V_G may include several volumes $V_\mathcal{T}$. The labelling of the different subsets of V_G using the internal representations of the training set $\mathcal{T}$ may therefore be redundant depending on the particular decoder under study. It is nevertheless a convenient starting point from the analytical point of view and, as shown below, it does capture the main features of the geometry of the coupling space.

The formalism recently introduced for a toy-model of MLN [8] can be used to compute the distribution of the "sizes" of the volumes associated to the internal representations $\mathcal{T}$. Once the canonical free-energy $g(r) = -\frac{1}{Nr}\overline{\log(\sum_{\mathcal{T}} V_{\mathcal{T}}{}^{r})}$ is known, one obtains the micro-canonical entropy $\mathcal{N}(k)$ (i.e. the logarithm of the typical number) of volumes $V_{\mathcal{T}}$ whose sizes are equal to $k = \frac{1}{N}\log V_{\mathcal{T}}$ using the Legendre relations $k_r = \frac{\partial(rg(r))}{\partial r}$ and $\mathcal{N}(k_r) = -\frac{\partial g(r)}{\partial(1/r)}$ [8]. The average over the patterns is performed using the replica trick for r integer expecting that the final results remains valid for any real value of r. There are r blocks ($\rho = 1, \ldots, r$) of n replicas ($a = 1, \ldots, n$). Thus the spin glass order parameters are the matrices $\mathcal{Q}_\ell$ and $\hat{\mathcal{Q}}_\ell$ of the typical overlaps $q_\ell^{a\rho,b\lambda} = \frac{K}{N}\sum_i J_{i\ell}^{a\rho} J_{i\ell}^{b\lambda}$ between two weight vectors incoming onto the same hidden unit ℓ ($\ell = 1, \ldots, K$) and of their conjugate Lagrange multipliers $\hat{q}_\ell^{a\rho,b\lambda}$. Since all the hidden units are indistinguishable, we assume that at the saddle point $\mathcal{Q}_\ell = \mathcal{Q}$ and $\hat{\mathcal{Q}}_\ell = \hat{\mathcal{Q}}$ independently of ℓ. Within the replica symmetric (RS) Ansatz [3], we find

$$g(r) = \underset{q,q_*}{Extr}\left\{\frac{1-r}{2r}\log(1-q_*) - \frac{1}{2r}\log(1-q_*+r(q_*-q)) - \frac{q}{2(1-q_*+r(q_*-q))}\right.$$
$$\left. -\frac{\alpha}{r}\int\prod_\ell Dx_\ell \;\log\mathcal{H}(\{x_\ell\})\right\} \tag{2}$$

where $\mathcal{H}(\{x_\ell\}) = Tr_{\{\tau_\ell\}}\prod_\ell\int Dy_\ell H[(y_\ell\sqrt{q_*-q} + \tau_\ell x_\ell\sqrt{q})/\sqrt{1-q_*}]^r$. Here, $q_*(r) = q^{a\rho,a\lambda}$ and $q(r) = q^{a\rho,b\lambda}$ are the typical overlaps between two weight vectors corresponding to the same (a,$\rho \neq \lambda$) and to different ($a \neq b$) internal representations $\mathcal{T}$ respectively [2,8]. The Gaussian measure is denoted by $Dx = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ whereas the function H is defined as $H(y) = \int_y^\infty Dx$. In eqn.(2), the sum $Tr_{\{\tau_\ell\}}$ runs over the internal representations $\{\tau_\ell\}$ giving a positive output $f(\{\tau_\ell\}) = +1$ only, since the outputs σ^μ can always be set equal to $+1$ at the cost of redefining the input patterns.

The whole distribution of sizes is available through $g(r)$. When $N \to \infty$, $\frac{1}{N}\overline{\log(V_G)} = -g(r=1)$ is dominated by volumes of size $k_{r=1}$ whose corresponding entropy (i.e the logarithm of their number divided by N) is $\mathcal{N}_D = \mathcal{N}(k_{r=1})$. At the same time the most numerous volumes are those of smaller size $k_{r=0}$, since in the limit $r \to 0$ all the $\mathcal{T}$ are counted irrespectively of their relative volumes. Their corresponding entropy $\mathcal{N}_R = \mathcal{N}(k_{r=0})$ is the normalized logarithm of the total number of implementable internal representations. The quantities $\mathcal{N}_D$ and $\mathcal{N}_R$ (that for lack of space we do not write explicitly) are easily obtained from the RS free–energy eqn.(2) using the above Legendre identities. In particular, $q(r=1)$ is the usual saddle point overlap of the Gardner volume $g(1)$ [2,4]. The vanishing condition for the entropies should coincide with the zero volume condition for V_G and thus should give the storage capacity of the models.

Both $\mathcal{N}_D$ and $\mathcal{N}_R$ have a straightforward interpretation in the context of information theory. One can easily verify that the quantity of information $\mathcal{I}$ carried by the distribution of the implementable internal representations $\mathcal{T}$ about the weights, $\mathcal{I} = -\sum_{\mathcal{T}} \frac{V_{\mathcal{T}}}{V_G} \log \frac{V_{\mathcal{T}}}{V_G}$, is equal to $\mathcal{N}_D$. The information capacity, i.e. the maximal quantity of information one can extract from the internal representations, is achieved when all internal representations $\mathcal{T}$ are equiprobable and thus equals $\mathcal{N}_R$. One should notice that the Mitchison–Durbin [7] geometrical calculation is simply an upper (and decoder–independent) bound on $\mathcal{N}_R$.

Let us see now the physical and geometrical interpretation of $\mathcal{N}_D$. Fig. 1 displays the RS entropy $\mathcal{N}_D$ as a function of α for both the parity and committee machines with $K = 3$ hidden units. This entropy vanishes at a critical value α_D of the size of the training set. Numerically, we find $\alpha_D \simeq 3.8$ and 2.9 for the parity and the committee machines respectively. For comparison, the storage capacities obtained with the one step RSB Ansatz are $\alpha_c \simeq 5$ and 3 respectively [4]. Being the entropy of a discrete system, $\mathcal{N}_D$ cannot be negative and therefore α_D is an upper bound of the size of the training set α_{RSB} where the replica symmetry breaking occurs for both $\mathcal{N}_D$ and V_G [8] . It is indeed known that $\alpha_{RSB} = 3.2$ and 1.8 for the parity and the committee machines respectively [4]. When $\alpha < \alpha_{RSB}$, the RS assumption is exact whereas $\mathcal{N}_D$ is positive, showing that the number of internal representations volumes contributing to V_G is exponentially large with N. q_* measures the typical overlap inside one of these volumes, while the usual overlap q arising in the RS computation of V_G tells us how far away are two different volumes $V_{\mathcal{T}}$. The behaviour of q_* versus α is shown in the inset of figure 1. When choosing randomly two weights vectors storing the training set, the probability that they belong to the same V_T vanishes as $\exp(-N\mathcal{N}_D)$ and their overlap distribution cannot be told from a Dirac peak in q, as must be for the RS solution to be exact. As a consequence, the blind computation of V_G, though it gives correct results, hides the geometrical structure of the weight space. In the limit of a large number K of hidden units, the asymptotic expressions of the overlaps and of α_D may be obtained analytically. We find that $q = 0$ and $q \simeq 1 - \frac{128}{\pi^2 \alpha^2}$ for the parity and the committee machines respectively and that $q_* \simeq 1 - \frac{\pi^2 \Gamma^2}{2\alpha^2 K^2}$ in both cases with $\Gamma = -1/(\sqrt{\pi} \int du H(u) \log H(u)) \simeq 0.62$. The corresponding entropies $\mathcal{N}_D^{(Par)} \simeq \log K - \alpha \log 2$ and $\mathcal{N}_D^{(Com)} \simeq \log K - \frac{\pi^2 \alpha^2}{256}$ vanish at $\alpha_D^{(Par)} \simeq \frac{\log K}{\log 2}$ and $\alpha_D^{(Com)} \simeq \frac{16}{\pi} \sqrt{\log K}$.

When $\alpha > \alpha_{RSB}$, the computation of $\mathcal{N}_D$ requires the introduction, at the first stage of RSB, of four order parameters q'_*, q_0, q_1, m : q'_* is the internal overlap of the internal representations volumes and q_0, q_1, m are simply the usual parameters arising in the one step Gardner's computation [11]. For brevity we only present below our numerical results together with their geometrical interpretation. Above α_{RSB}, there exist a finite number of big regions with mutual overlap q_0. Each region ρ contains an exponential number of volumes M_ρ of internal overlap q'_* and typically separated by an overlap q_1. The number of such regions may be roughly estimated by $\frac{1}{1-m}$, since $m = 1 - \sum_\rho (M_\rho / \sum_{\rho'} M_{\rho'})^2$, whereas in the RS phase $m = 1$. We have checked numerically this geometrical scenario for the parity machine with $K = 3$ hidden units (numerically much simpler than the committee machine case since $q_0 = 0$ at the saddle point). The internal overlap q'_* is continuous at the RSB transition – see the inset of fig. 1 – with $q_* < q'_*$ for $\alpha > \alpha_{RSB}$. We conjecture that increasing α a whole continuous breaking of RSB occurs. The geometrical process should then be thought of as a progressive shrinking and disappearance of volumes with internal overlap $q_*(\alpha)$ inside sub-regions characterized by $q(x, \alpha)$ [3]. In fig. 1, we have reported the curve of $\mathcal{N}_D$ computed with this one step Ansatz for the parity machine $K = 3$. α_D increases from $\simeq 3.8$ (RS value) to a value close to 5 and thus to the one step RSB value of α_c [4] (since q'_* and q_1 are close to 1, our numerical results become less precise for α larger than ~ 4.1 and $\alpha_D \simeq 5$ is obtained through the linear extrapolation corresponding to the dashed part of the curve).

The RS calculation of $\mathcal{N}_R$ for both machines, leads the following general results. When $\alpha < \frac{2}{K}$, one finds that all the $2^{(K-1)P}$ internal representations may be implemented. This obviously coincides with the storage capacity of the hidden perceptrons seeing only N/K input units. For $\alpha > \frac{2}{K}$, we find that at the saddle-point $q_* = 1$, meaning that the most numerous volumes $V_{\mathcal{T}}$ are almost empty and are therefore the smallest ones at the same time. The resolution of the saddle-point equations requires the introduction of a new order parameter $\mu = \lim_{r\to 0} r/(1-q_*)$, describing how quickly the typical size of the volumes decreases with respect to the inverse "temperature" r [8]. For the parity machine with $K \geq 3$, $q = 0$, $\mu = \alpha K(\alpha K - 2)$ is always a locally stable saddle-point giving $\mathcal{N}_R^{(Par)} = \alpha K \log(\alpha K) - (\alpha K - 1)\log(\alpha K - 1) - \alpha \log 2$ which exactly saturates the upper bound derived by Mitchison–Durbin [7]. In the case of the committee machine, a simple analytical expression for $\mathcal{N}_R^{(Com)}$ is not available for finite K. Once more in fig. 1, we report the numerical results concerning the RS calculations of $\mathcal{N}_R$ for both machines with $K = 3$. The value α_R at which $\mathcal{N}_R$ vanishes should satisfy the obvious inequality $\alpha_D \leq \alpha_R \leq \alpha_c$; the RS approximation however overestimates α_R leading to an expression which is slightly larger than the one step value of α_c. For the parity and committee machines with $K = 3$ we find $\alpha_R = 5.4$ and 3.5 respectively. This is an evidence for the necessity of RSB to compute exactly $\mathcal{N}_R$ for finite K.

When $K \gg 1$, $\mathcal{N}_R$ (resp. α_R) is asymptotically equal to $\mathcal{N}_D$ (resp. α_D). In the case of the parity machine α_D and α_R also coincide with the known value of $\alpha_c = \frac{\log K}{\log 2}$ [4]. We expect the same equality ($\alpha_D = \alpha_R = \alpha_c = \frac{16}{\pi}\sqrt{\log K}$) to hold in the case of the committee machine. In order to show that the RS solution of $\mathcal{N}_R$ is asymptotically correct, we have checked its local stability with respect to fluctuations of the order parameter matrices. Although it would require a complete analysis of the eigenvalues of the Hessian matrix, we have focused only on the replicons 011 and 122 in the notations of [10], which are usually the most "dangerous" modes [3]. For a free–energy functional depending only on one order parameter matrix $q^{a\rho,b\lambda}$, the corresponding eigenvalues are Λ_{011} and Λ_{122} given by formula

(41) in ref. [10]. In our case, however, the free-energy depends on $2K$ matrices $\{\mathcal{Q}_l, \hat{\mathcal{Q}}_l\}$ and the stability condition for each mode reads $\Delta(\alpha, K) = \hat{\Lambda}\ (\ \Lambda + (K-1)\overline{\Lambda}\ ) - \frac{1}{K^2} < 0$ where $\hat{\Lambda}, \Lambda, \overline{\Lambda}$ are the eigenvalues computed for the fluctuations with respect to $\hat{\mathcal{Q}}_\ell \hat{\mathcal{Q}}_\ell$, $\mathcal{Q}_\ell \mathcal{Q}_\ell$ and $\mathcal{Q}_\ell \mathcal{Q}_m$ $(\ell \neq m)$ respectively [2,4]. A tedious calculation leads to the final expressions Δ_{011} and Δ_{122} [11]. For the parity machine, we find $\Delta_{011}^{(Par)}(\alpha, K) = \frac{\alpha}{K}(\frac{2}{\pi} + \frac{1}{\alpha K}(1 - \frac{4}{\pi}))^2 - \frac{1}{K^2}$ and $\Delta_{122}^{(Par)}(\alpha, K) = 0$ which are valid for $K \geq 3$ and $\alpha \geq \frac{2}{K}$. The RS solution is unstable against 011 replicon mode for $\alpha \geq \frac{3.27}{K}$ (i.e. of the same order as the storage capacity of each single input perceptron). However, in the large K limit, Δ_{011} vanishes. For the committee machine, one finds $\Delta_{011}^{(Com)}(\alpha, K) \simeq \frac{\sqrt{2}}{\pi^3 K}$ and $\Delta_{122}^{(Com)}(\alpha, K) \simeq -\frac{1}{2K^2}$ for $K \gg 1$ and $\alpha \gg 1$. We notice that the 122 mode is always stable and a unique order parameter q_* is thus sufficient to describe the volume associated to a set of internal representations $\mathcal{T}$. For both machines, our RS solution is marginally stable when $K \to \infty$ and should therefore become exact in this limit.

In order to understand what are the consequences of the weight space structure on the generalization ability of MLN, we now modify our approach to the case of deterministic input–output mappings.

The case of the parity machine trained on a learnable rule (i.e. generated by a "teacher" network endowed with an identical architecture) has been recently studied [6] in the Bayesian framework where the generalization properties are derived through the knowledge of the entropy $S_G = -\frac{1}{N}\overline{V_G \log V_G}$. The transition from high generalization error $\epsilon_g = \frac{1}{2}$ to low $\epsilon_g (= \frac{\Gamma}{\alpha}$ for large α) [6] may be geometrically understood along the lines developed above. The free-energy $s(r)$ generating the distribution of the "sizes" of the internal representation volumes $V_\mathcal{T}$ becomes $s(r) = -\frac{1}{Nr}\overline{\sum_\mathcal{T} V_\mathcal{T}^{\ r} \log(\sum_\mathcal{T} V_\mathcal{T}^{\ r})}$ where we obviously recover $S_G = s(1)$. The replica calculation of $s(r)$ technically differs from the computation of $g(r)$ by taking the limit $n \to 1$ instead of $n \to 0$ [11]. Within the RS Ansatz, we find

$$s(r) = \underset{q}{Extr}\left\{ \frac{1-r}{2r}\log(1-q_*) - \frac{1}{2r}\log(1 - q_* + r(q_* - q)) - \frac{q}{2(1+(r-1)q_*)} \right.$$

$$-\frac{2\alpha}{f(q_*)}\int\prod_\ell Dx_\ell \mathcal{H}(\{x_\ell\})\log\mathcal{H}(\{x_\ell\})\Bigg\} \tag{3}$$

with $f(q_*) = [2\int DzH(z\sqrt{q_*}/\sqrt{1-q_*})^r]^K$ and $q_*(r)$ is the saddle-point overlap of $s_0(r) = (1-r)\log(1-q_*) - \log(1+(r-1)q_*) - 2\alpha\log f(q_*)$. The logarithm $\mathcal{M}_D$ of the number of the internal representation volumes contributing to the Bayesian entropy S_G is given by $\mathcal{M}_D = \frac{\partial s}{\partial r}(r=1)$. We find that for $\alpha K \gg 1$, $q_* \simeq 1 - \frac{\pi^2\Gamma^2}{2\alpha^2K^2}$. In the case of the parity machine $q=0$, $\mathcal{M}_D^{(Par)} \simeq \log K - \alpha\log 2$ for $\alpha < \alpha_0 = \frac{\log K}{\log 2}$ and $q = q_*$, $\mathcal{M}_D = 0$ for $\alpha > \alpha_0$. Thus, below α_0, the weight space is composed of an exponentially large number of volumes and the typical overlap q between the volume occupied by the teacher and any other one is zero : $\epsilon_g = \frac{1}{2}$. Above α_0, since only one internal representation survives, the student has fallen down into the teacher volume : $q = q_*$ and $\epsilon_g \simeq \frac{\Gamma}{\alpha}$. When $\alpha < \alpha_0$, $S_G^{(Par)} = \alpha\log 2$, meaning that all the sets of P outputs are equiprobable. Choosing them with a probability $V_G(\{\sigma\})$ is then equivalent to drawing them randomly. This is the reason why α_D defined for the storage problem (and more generally d_{vc}) appears on the generalization curve of the parity machine. Our calculation also indicates that the computation of α_0 should include RSB effects for finite K, while the asymptotic RS expression of ϵ_g ought to be exact, as has been found for the non–monotonic perceptron [12].

Turning to the committee machine, a calculation of the Bayesian entropy S_G similar to [5] leads to the following results when $K \gg \alpha \gg 1$. The typical teacher–student overlap q decreases as $1 - \frac{\pi^6\Gamma^4}{2\alpha^4}$ giving an entropy $S_G^{(Com)} \simeq 2\log\alpha$ and $\epsilon_g \simeq \frac{2\Gamma}{\alpha}$. This shows that, at variance with the parity machine case, only a small fraction among the 2^P possible sets of outputs contribute to $S_G^{(Com)}$ and explains why the generalization curve is smooth for $\alpha \simeq \sqrt{\log K}$ (which is the order of magnitude of d_{vc}). We find $\mathcal{M}_D^{(Com)} \simeq \log K - \log\alpha$, confirming that α_D (and thus d_{vc}) is not relevant to the computation of the typical generalization error. At $\alpha_{c.o.} \sim K$, only a single internal representation subsists and beyond this critical size of the training set the generalization error should equal $\epsilon_g = \frac{\Gamma}{\alpha}$ as is for finite K and large α [5]. Note that the order of magnitude of $\alpha_{c.o.}$ is corroborated by the condition $q = q_*$ one has to fulfill once a unique $V_{\mathcal{T}}$ remains non–empty. A rigorous proof of the presence of this cross–over (from $\epsilon_g = 2\frac{\Gamma}{\alpha}$ to $\epsilon_g = \frac{\Gamma}{\alpha}$) at $\alpha_{c.o.}$ would however require to extend the validity of our calculation to the regime $1 \ll \alpha \sim K$.

We are grateful to N. Brunel, M. Budinich, M. Ferrero and D. O'Kane for discussions.

REFERENCES

* Email: monasson@roma1.infn.it

† Email: zecchina@to.infn.it

[1] T.M. Cover, *IEEE Trans. Electron. Comput.* **EC**-14, 326 (1965)

[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] E. Barkai, D. Hansel, I. Kanter, *Phys. Rev. Lett.* **65**, 2312 (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)

[5] H. Schwarze, J. Hertz, *Europhys. Lett.* **20**, 375 (1992)

[6] M. Opper, *Phys. Rev. Lett.* **72**, 2113 (1994)

[7] G.J. Mitchison, R.M. Durbin, *Bio. Cybern.* **60**, 345 (1989)

[8] R. Monasson, D. O'Kane, *Europhys. Lett.* **27**, 85 (1994)

[9] V.N. Vapnik, *Estimation of Dependences Based on Empirical Data* (Springer-Verlag, New York, 1982)

[10] T. Temesvari, C. De Dominicis, I. Kondor, *J. Phys.* **A** 27, 7569 (1994)

[11] R. Monasson, R. Zecchina, in preparation

[12] A. Engel, L. Reimers, *Europhys. Lett.* **28**, 531 (1994)

FIG. 1. $\mathcal{N}_R$ (upper curves) and $\mathcal{N}_D$ (lower curves) for the parity machine (bold) and the committee machine (light), with $K = 3$ hidden units. Inset: q_1, q_*, q'_* (lower, middle and upper curves respectively) versus α for the parity machine (q_1 starts at $\alpha = \alpha_{RSB} \simeq 3.2$ with a value close to 0.93).

Monasson-Zecchina

Fig.1