Learning and Generalization Theories of Large Committee–Machines

Rémi Monasson** and Riccardo Zecchina[†]

[*] *Laboratoire de Physique Théorique de l'ENS, 24 rue Lhomond, 75231 Paris cedex 05, France*

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

Abstract

The study of the distribution of volumes associated to the internal representations of learning examples allows us to derive the critical learning capacity ($\alpha_c = \frac{16}{\pi}\sqrt{\ln K}$) of large committee machines, to verify the stability of the solution in the limit of a large number K of hidden units and to find a Bayesian generalization cross–over at $\alpha = K$.

PACS Numbers : 05.20 - 64.60 - 87.10

Typeset using REVTEX

---

*Laboratoire de Physique Théorique de l'ENS est une unitée propre du CNRS rattachée à l'université de Paris-Sud

I. INTRODUCTION

Following the approach presented in refs. [1,2], we derive the learning behaviour of non overlapping committee machines with a large number K of hidden units. Scope of the paper is to clarify some of the analytical aspects of a method which is based on the internal degrees of freedom of MultiLayer Networks (MLN) and which requires a double analytic continuation. Such an approach, beside allowing for the derivation of new results both for the learning and the generalization behaviour of MLN, makes a rigorous bridge between different fields in the theory of neural computation, such as Information Theory, VC–dimension and Bayesian rule extraction, and statistical mechanics [1,11,9,10,12]. Moreover, it sheds new light on the role of internal representations of the learning examples by relating it to the distribution of domains of solutions and pure states in the weight space of the network.

The method consists in a generalization of the well known Gardner approach [4]. While the latter studies the typical volume of couplings associated to the overall input–output map implemented by the network, here we consider the decomposition of such volume in a macroscopic number of single volumes associated to all possible internal representations compatible with the learned examples. In addition to the interaction weights, we take as dynamical variables also the internal state variables of the MLN characterizing internal representations. For the storage problem, we are therefore interested in counting the typical number $\exp(\mathcal{N}_D)$ of volumes giving the dominant contribution to Gardner's volume and to compare it with the total number $\exp(\mathcal{N}_R)$ of non–empty volumes. At the learning transition, i.e. when the Gardner's total volume shrinks to zero and no more patterns can be learned without errors, we expect both entropies $\mathcal{N}_D$ and $\mathcal{N}_R$ to vanish. The vanishing condition on $\mathcal{N}_D$ and $\mathcal{N}_R$ gives thus an alternative indication on the storage performance of the studied network.

The generalization properties of the network, i.e. the rule inference capability from a given set of deterministic input–output examples, also depend on the geometrical structure of the weight space. As we shall discuss in the sequel, the internal representation approah can be straightforwardly extended to the study of the generalization error of MLN in the Bayesian framework [12,10], allowing for a geometrical interpretation of the generalization transition together with a clarification of the role of the VC-dimension [9].

The method discussed here for the committee machine can be applied to other non overlapping MLN with arbitrary decoder functions. For instance, one may show that the stability analysis of the RS solution $\alpha_c = \ln K/\ln 2$ for parity machine is an exact result in the limit $K >> 1$. Such a result coincides with the one derived in [5] following the standard Gardner [4] approach with one step of Replica Symmetry Breaking. Moreover, one may also reproduce the known results [10] on the parity machine generalization transition [1] by means of a detailed geometrical intepretation.

The paper is organized as follows. In Sec.**II** we outline the basic points of our approach, both for learning and for the Bayesian generalization problems. In Sec.**III** and Sec.**IV** we study the entropies $\mathcal{N}_R$ and $\mathcal{N}_D$ in the $K >> 1$ limit and compute the closed expression for the critical capacity. The detailed analysis of the stability of the solution is given in Sec. **V**. Finally, in Sec.**VI**, we derive the entropy $\mathcal{M}_D$ (for $K >> 1$) of the internal representation contributing to the Bayesian entropy. This allows us to analyze the generalization transistion of the committee machine and to explain why the VC–dimension is not relevant for its typical generalization properties.

II. THE INTERNAL REPRESENTATION VOLUMES APPROACH

We consider tree-like commitee machines composed of K non-overlapping perceptrons with real-valued weights $J_{\ell i}$ and connected to K sets of independent inputs $\xi_{\ell i}$ ($\ell = 1, ..., K$, $i = 1, ..., N/K$). Committee machines are characterized by an output σ which is a binary function $f(\{\tau_\ell\}) = \text{sign}(\sum_\ell \tau_\ell)$ of the cells $\tau_\ell = \text{sign}(\sum_i J_{\ell i}\xi_{\ell i})$ in the hidden layer. We refer to the set $\{\tau_\ell\}$ as the *internal representation* of the input pattern $\{\xi_{\ell i}\}$. Given a macroscopic set of $P = \alpha N$ binary unbiased patterns patterns (the training set), the learning problem consists in finding a suitable set of internal representations $\mathcal{T} = \{\tau_\ell^\mu\}$ with a corresponding non zero volume

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

where $\theta(\ldots)$ is the Heaviside function.

The total volume of the weight space available for learning, i.e. Gardner's total volume, is given by $V_G = \sum_{\mathcal{T}} V_{\mathcal{T}}$. We are interested in discussing the limit $\overline{\ln V_G} \to -\infty$, which defines the maximal possible size of the training set or the critical capacity of the model. The bar denotes the average over the patterns and their corresponding outputs [4] which, as usual, are drawn according to the binary unbiased distribution law.

As discussed in [1], the partition of V_G into connected components may be naturally obtained using the volumes $V_{\mathcal{T}}$ associated to the internal representations. This allows to give a geometrical interpretation of the learning and Bayesian generalization process in terms of the characteristics of the volumes dominating the overall distribution [1].

Following the standard statistical mechanics approach, we first compute

$$g(r) \equiv -\frac{1}{Nr} \overline{\ln \left(\sum_{\mathcal{T}} V_{\mathcal{T}}^{r} \right)} \tag{2}$$

and next derive the entropy $\mathcal{N}(w)$ of the volumes $V_{\mathcal{T}}$ whose sizes are equal to $w = \frac{1}{N} \ln V_{\mathcal{T}}$. This can be done using the Legendre relations $w_r = \frac{\partial (r g(r))}{\partial r}$ and $\mathcal{N}(w_r) = -\frac{\partial g(r)}{\partial (1/r)}$. Diversely from the standard replica calculations, here we deal with two analytic continuations: we have r blocks of n replicas and, once the we average over the quenched patterns for r and n integer has been done, we perform an analytic continuation to real values of r and n. Labeling blocks and replicas by ρ, λ and a, b respectively, the spin glass order parameters read

$$q_{\ell}^{a\rho,b\lambda} = \frac{K}{N} \sum_{i} J_{i\ell}^{a\rho} J_{i\ell}^{b\lambda} \quad . \tag{3}$$

They represent the typical overlaps between weight vectors incoming onto the same hidden unit ℓ ($\ell = 1, \ldots, K$) and belinging to blocks ρ, λ and replicas a, b. Associated to the the $q_{\ell}^{a\rho,b\lambda}$ there are also the conjugate Lagrange multipliers $\hat{q}_{\ell}^{a\rho,b\lambda}$. Since hidden units are equivalent, we assume that at the saddle point $q_\ell^{a\rho,b\lambda} = q^{a\rho,b\lambda}$ and $\hat{q}_\ell^{a\rho,b\lambda} = \hat{q}^{a\rho,b\lambda}$ independently of ℓ. Then, within the replica symmetric (RS) Ansatz [6], we find

$$g(r) = \operatorname*{Extr}_{q,q_*} \left\{ \frac{1-r}{2r} \ln(1-q_*) - \frac{1}{2r} \ln(1-q_*+r(q_*-q)) - \frac{q}{2(1-q_*+r(q_*-q))} \right. \\ \left. - \frac{\alpha}{r} \int \prod_{\ell=1}^{K} Dx_\ell \; \ln \mathcal{H}(\{x_\ell\}) \right\} \quad , \tag{4}$$

where

$$\mathcal{H}(\{x_\ell\}) = \operatorname*{Tr}_{\{\tau_\ell\}} \prod_{\ell=1}^{K} \int Dy_\ell H \left[\frac{y_\ell \sqrt{q_*-q} + \tau_\ell x_\ell \sqrt{q}}{\sqrt{1-q_*}} \right]^r \quad . \tag{5}$$

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 [4,2]. The Gaussian measure is denoted by $Dx = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ whereas the function H is defined as $H(y) = \int_y^\infty Dx$. Since with no loss of genereality the outputs σ^μ can be set equal to 1, in eqn.(4) the sum $\mathrm{Tr}_{\{\tau_\ell\}}$ runs over the internal representations $\{\tau_\ell\}$ giving a positive output $f(\{\tau_\ell\}) = +1$ only.

As discussed in [1], when $N \to \infty$, $\frac{1}{N}\overline{\ln(V_G)} = -g(r=1)$ is dominated by volumes of size $w_{r=1}$ whose corresponding entropy (i.e. the logarithm of their number divided by N) is $\mathcal{N}_D = \mathcal{N}(w_{r=1})$. At the same time the most numerous volumes are those of smaller size $w_{r=0}$, since in the limit $r \to 0$ all the $\mathcal{T}$ are counted irrespectively of their relative volumes. Their corresponding entropy $\mathcal{N}_R = \mathcal{N}(w_{r=0})$ is the (normalized) logarithm of the total number of implementable internal representations. The quantities $\mathcal{N}_D$ and $\mathcal{N}_R$ are easily obtained from the RS free–energy eqn.(4) using Legendre identities. In particular, $q(r=1)$ is the usual saddle point overlap of the Gardner volume $g(1)$ [4,5]. The vanishing condition for the entropies is related to the zero volume condition for V_G and thus to the storage capacity of the models.

In the above discussion, we have focused on the storage problem. However, the generalization properties also depend on the internal structure of the coupling space. Let us for instance consider the case of a learnable rule, defined by a teacher network. When a student with the same architecture is given more and more examples of the rule to infer, its version space [12] shrinks. In the perceptron case, the version space is simply connected and the typical generalization error done by the student on a new example goes to zero as its overlap with the teacher increases. The situation is much more involved in multilayer neural networks since the presence of separated components of the version space makes the alignment of the student along the teacher direction more difficult. The approach we have exposed above for the learning problem may be extended to acquire a better understanding of the generalization process in multilayer networks.

We shall restrict to the Bayesian framework where all teacher are sorted according to their a priori probabilities. The generalization properties are derived through the Bayesian entropy

$$S_G = -\frac{1}{N} \sum_{\{\sigma^\mu\}} \overline{V_G \ln V_G} \quad , \tag{6}$$

where the sum runs over all 2^P sets of possible outputs. If we know intend to look at the distibution of the sizes fo the internal representation volumes $V_{\mathcal{T}}$, we have to consider the generating free–energy [8]

$$s(r) \equiv -\frac{1}{Nr} \sum_{\{\sigma^\mu\}} \overline{V_G \ln \left(\sum_{\mathcal{T}} V_{\mathcal{T}}^r \right)} \quad . \tag{7}$$

To compute $s(r)$ with the replica method, we have to introduce $1 + nr$ replicas and send $n \to 0$ at the end of the computation. The order parameters entering the computation are the overlaps $p^{a\rho}$ between the teacher and the nr students and the overlaps $q^{a\rho,b\lambda}$ between two differents students. Within the RS Ansatz, we assume that $p^{a\rho} = p$ and $q^{a\rho,b\lambda} = q$ if $a \neq b$, q^* otherwise. The result we obtained is

$$s(r) = \operatorname*{Extr}_{p,q,q_*} \left\{ \frac{1-r}{2r} \ln(1-q_*) - \frac{1}{2r} \ln(1 - q_* + r(q_* - q)) - \frac{q - p^2}{2(1 - q_* + r(q_* - q))} \right.$$
$$\left. - \frac{2\alpha}{r} \int \prod_{\ell=1}^{K} Dx_\ell \left[\operatorname*{Tr}_{\{\tau_\ell\}} \prod_\ell H\left(\frac{\tau_\ell x_\ell p}{\sqrt{q_0 - p^2}} \right) \right] \ln \mathcal{H}(\{x_\ell\}) \right\} \quad , \tag{8}$$

In the following , we shall focus on the logarithm (divided by N) of the number of internal representations contributing to $S_G = -s(1)$, that is

$$\mathcal{M}_D = \frac{\partial s}{\partial r}(r = 1) \quad . \tag{9}$$

It can be easily verified that $p = q$ is always a saddle–point when $r = 1$. In the Bayesian framework, the typical overlap between two student is equal to the scalar product between the teacher and any student.

III. ANALYSIS OF $\mathcal{N}_R$ IN THE $K >> 1$ LIMIT

We first focus on the $r \to 0$ case to compute the typical logarithm $\mathcal{N}_R$ of the total number of internal representations. One can check on the saddle–point equations for q, q^* that the correct scalings of the order parameters are $q = O(1)$ and $q^* = 1 - O(r)$. In the following, we shall call $\mu = \lim_{r\to 0}[r/(1-q^*)]$. Upon keeping the leading terms in K, the trace over $\mathcal{T}$ in equation (5) becomes [5]

$$H\left(\frac{Q_1}{\sqrt{1-Q_2}}\right) \prod_{\ell=1}^{K} A(x_\ell) \tag{10}$$

where

$$Q_1 = \frac{1}{\sqrt{K}} \sum_{\ell=1}^{K} \frac{B(x_\ell) + B(-x_\ell)}{A(x_\ell)} \quad , \tag{11}$$

and

$$Q_2 = \frac{1}{K} \sum_{\ell=1}^{K} \left(\frac{B(x_\ell) + B(-x_\ell)}{A(x_\ell)}\right)^2 \quad . \tag{12}$$

In the above expressions we have adopted the definitions

$$A(x) \equiv 1 + \frac{\exp\left(-x^2 \frac{\mu q}{2(1+\mu(1-q))}\right)}{\sqrt{1+\mu(1-q)}} \tag{13}$$

and

$$B(x) \equiv H\left(x\sqrt{\frac{q}{1-q}}\right) + \frac{\exp\left(-x^2 \frac{\mu q}{2(1+\mu(1-q))}\right)}{\sqrt{1+\mu(1-q)}} H\left(\frac{-x\sqrt{q/(1-q)}}{\sqrt{1+\mu(1-q)}}\right) \quad . \tag{14}$$

In the $K \to \infty$ limit, Q_1 becomes a gaussian variable with zero mean and variance $Q_2 = \int Dx(B(x) + B(-x))^2/A(x)^2$. The free–energy for $r \to 0$ then reads $-G(q,\mu)/r + O(\ln r)$ where

$$G(q,\mu) = \frac{1}{2}\ln(1+\mu(1-q)) + \frac{1}{2}\frac{\mu q}{1+\mu(1-q)} + \alpha\int Dx \ln H\left(\frac{x\sqrt{Q_2}}{\sqrt{1-Q_2}}\right) +$$
$$\alpha K \int Dx \ln\left(1 + \frac{\exp\left(-x^2\frac{\mu q}{2(1+\mu(1-q))}\right)}{\sqrt{1+\mu(1-q)}}\right) + O\left(\frac{1}{\sqrt{K}}\right) \tag{15}$$

and the typical logarithm $\mathcal{N}_R$ of the total number of internal representations is simply the maximum of $G(q,\mu)$ over q and μ. Taking the scaling relation $\mu = mK^2$ (which can be inferred from the equation $\frac{\partial G}{\partial \mu} = 0$), and $q = O(1)$, one finds

$$Q_2 = \frac{2}{\pi}\arcsin(q) \quad . \tag{16}$$

Finally, defining $q = 1-\epsilon$ and taking the saddle point equation with respect to m (which implies $m = \alpha^2$) and ϵ, one finds the following result

$$\mathcal{N}_R = \ln(K) - \frac{\pi^2\alpha^2}{256} + O(\ln(\alpha)) \quad , \tag{17}$$

which vanishes at

$$\alpha_R = \frac{16}{\pi}\sqrt{\ln(K)} \quad . \tag{18}$$

IV. ANALYSIS OF $\mathcal{N}_D$ IN THE $K >> 1$ LIMIT

We shall now concentrate on the $r \to 1$ case, which corresponds to the internal representations giving the dominant contribution to the Gardner volume V_G. The typical logarithm of such internal representations is $\mathcal{N}_D$. Before taking the limit $K \to \infty$, the Legendre transform of expression (4) gives

$$\mathcal{N}_D = \frac{1}{N}\overline{\ln V_G} + \frac{2qq^* - q^* - q^2}{2(1-q)^2} - \frac{1}{2}\ln(1-q^*) - \alpha K \times$$
$$\int \prod_\ell Dx_\ell \frac{\mathrm{Tr}_{\{\tau_\ell\}} \prod_{l=2}^K H\left(\tau_\ell x_\ell \sqrt{\frac{q}{1-q}}\right) \int Dy H\left(\frac{y\sqrt{q^*-q}+x_1\tau_1\sqrt{q}}{\sqrt{1-q^*}}\right) \ln H\left(\frac{y\sqrt{q^*-q}+x_1\tau_1\sqrt{q}}{\sqrt{1-q^*}}\right)}{\mathrm{Tr}_{\{\tau_\ell\}} \prod_{l=1}^K H\left(\tau_\ell x_\ell \sqrt{\frac{q}{1-q}}\right)} \tag{19}$$

where $\overline{\ln V_G}$ is the replica symmetric expression of the Gardner volume. When K is large, the trace over all allowed internal representations may be evaluated as in the previous section [5]. We find the following scalings for the order parameters

$$q \simeq 1 - \frac{128}{\pi^2\alpha^2}$$
$$Q_2 \simeq 1 - \frac{32}{\pi^2\alpha}$$
$$q^* \simeq 1 - \frac{\Gamma^2}{2\pi^2 K^2\alpha^2} \tag{20}$$

where

$$\frac{1}{\Gamma} = -\sqrt{\pi} \int_{-\infty}^{\infty} duH(u)\ln H(u) \tag{21}$$

for large K and α. Therefore, the asymptotic expression of the entropy of contributing internal representations is

$$\mathcal{N}_D = \ln(K) - \frac{\pi^2\alpha^2}{256} + O(\ln(\alpha)) \quad , \tag{22}$$

which vanishes at

$$\alpha_D = \frac{16}{\pi}\sqrt{\ln(K)} \quad . \tag{23}$$

For large K, α_D coincide with α_R. Reasonnably, we expect α_c to be equal to these critical numbers and to scale as $\frac{16}{\pi}\sqrt{\ln(K)}$ too.

V. STABILITY ANALYSIS

In order to show that our RS calculation of $\mathcal{N}_R$ is asymptotically correct when the number of hidden units K is large, 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 [7], which are usually the most "dangerous" modes [6]. For a free–energy functional depending only on one order parameter matrix $q^{a\rho,b\lambda}$, the corresponding eigenvalues are

$$\Lambda_{011} = \frac{\partial^2\mathcal{F}}{\partial q^{a\rho,b\lambda}\partial q^{a\rho,b\lambda}} - 2r\frac{\partial^2\mathcal{F}}{\partial q^{a\rho,b\lambda}\partial q^{a\rho,c\mu}} + 2(r-1)\frac{\partial^2\mathcal{F}}{\partial q^{a\rho,b\lambda}\partial q^{a\rho,b\mu}} + (r-1)^2\frac{\partial^2\mathcal{F}}{\partial q^{a\rho,b\lambda}\partial q^{a\mu,b\nu}} - 2r(r-1)\frac{\partial^2\mathcal{F}}{\partial q^{a\rho,b\lambda}\partial q^{a\mu,c\nu}} + r^2\frac{\partial^2\mathcal{F}}{\partial q^{a\rho,b\lambda}\partial q^{c\mu,d\nu}} \tag{24}$$

and

$$\Lambda_{122} = \frac{\partial^2 \mathcal{F}}{\partial q^{a\rho,a\lambda}\partial q^{a\rho,a\lambda}} - 2\frac{\partial^2 \mathcal{F}}{\partial q^{a\rho,a\lambda}\partial q^{a\rho,a\mu}} + \frac{\partial^2 \mathcal{F}}{\partial q^{a\rho,a\lambda}\partial q^{a\mu,b\nu}} \tag{25}$$

are given by formula (41) in ref. [7]. In our case, however, the free-energy depends upon the $2K$ matrices $\{\mathcal{Q}_l, \hat{\mathcal{Q}}_l\}$. According to [4,5], the stability condition for each mode reads

$$\Delta(\alpha, K) = \hat{\Lambda}\ (\ \Lambda + (K-1)\overline{\Lambda}\ ) - \frac{1}{K^2} < 0 \tag{26}$$

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. Since we are interested in the stability of the saddle–point giving $\mathcal{N}_R$, we focus on the limit $r \to 0$. In this case, the correct scalings of the order parameters are $q^* = 1 - r/\mu + O(r^2)$, $q = O(1)$. For the (011) mode, we find

$$\begin{aligned}
\hat{\Lambda}_{011} &= \frac{(1-q)^2}{K} r^2 \\
\Lambda_{011} &= \frac{\alpha\mu^2}{r^2} \int \prod_{\ell=1}^{K} Dx_\ell \left[\frac{N_1}{D} - \mu \left(\frac{N_2}{D}\right)^2 \right]^2 \\
\overline{\Lambda}_{011} &= \frac{\alpha\mu^4}{r^2} \int \prod_{\ell=1}^{K} Dx_\ell \left[\frac{N_3}{D} - \frac{N_4}{D^2} \right]^2
\end{aligned} \tag{27}$$

at leading order when $r \ll 1$. The quantities defined in (27) are

$$D = \mathop{\mathrm{Tr}}_{\{\tau_\ell\}} \prod_{\ell=1}^{K} B(\tau_\ell x_\ell) \tag{28}$$

$$\begin{aligned}
N_1 &= \mathop{\mathrm{Tr}}_{\{\tau_\ell\}} \left[\prod_{\ell=2}^{K} B(\tau_\ell x_\ell) \right] \frac{\exp\left(-x_1^2 \frac{\mu q}{2X)}\right)}{X^{3/2}} \left[\left(1 - \frac{\mu q x_1^2}{X}\right) \times \right. \\
&\quad \left. H\left(\frac{-\tau_1 x_1 \sqrt{q/(1-q)}}{\sqrt{X}} \right) - \frac{\mu\sqrt{q(1-q)}}{\sqrt{2\pi}\sqrt{X}} \exp\left(- \frac{x_1^2 \mu q}{2X)(1-q)} \right) \right] \\
N_2 &= \mathop{\mathrm{Tr}}_{\{\tau_\ell\}} \left[\prod_{\ell=2}^{K} B(\tau_\ell x_\ell) \right] \frac{\sqrt{1-q}}{X} Y_1 \\
N_3 &= \mathop{\mathrm{Tr}}_{\{\tau_\ell\}} \left[\prod_{\ell=3}^{K} B(\tau_\ell x_\ell) \right] \frac{1-q}{X^2} Y_1 Y_2 \\
N_4 &= \left\{ \mathop{\mathrm{Tr}}_{\{\tau_\ell\}} \left[\prod_{\ell\neq 1} B(\tau_\ell x_\ell) \right] \frac{\sqrt{1-q}}{X} Y_1 \right\} \times \\
&\quad \left\{ \mathop{\mathrm{Tr}}_{\{\tau_\ell\}} \left[\prod_{\ell\neq 2} B(\tau_\ell x_\ell) \right] \frac{\sqrt{1-q}}{X} Y_2 \right\} \quad ,
\end{aligned} \tag{29}$$

in which we have posed $X = X(\mu, q) \equiv 1 + \mu(1-q)$ and where

$$Y_i \equiv \exp\left(-\frac{x_i^2 \mu q}{2X}\right)\left[\frac{\tau_i x_i \sqrt{q}}{\sqrt{X}\sqrt{1-q}} H\left(\frac{-\tau_i x_i \sqrt{q}}{\sqrt{X}\sqrt{1-q}}\right) + \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x_i^2 \mu q}{2X(1-q)}\right)\right] \quad . \tag{30}$$

In the large K limit, the asymptotic expressions of the order parameters are $\mu \simeq \alpha^2 K^2$ and $q \simeq 1 - 128/\pi^2/\alpha^2$. Using the previous expressions of $\hat{\Lambda}_{011}, \Lambda_{011}, \overline{\Lambda}_{011}$, we have

$$\Delta_{011}(\alpha, K) \simeq\simeq \frac{\sqrt{2}}{\pi^3 K} \tag{31}$$

when $K \gg 1$ and $\alpha \gg 1$. Therefore, our RS solution is unstable against 011 replicon fluctuations. However, in the large K limit, Δ_{011} vanishes and the RS Ansatz becomes marginally stable.

Let us now analyse the (122) mode. Similar calculations lead to

$$\begin{aligned}
\hat{\Lambda}_{122} &= \frac{1}{\mu^2 K} r^2 \\
\Lambda_{122} &= \frac{\alpha \mu^2}{r^2} \int \prod_{\ell=1}^{K} D x_\ell \frac{N_5}{D} \\
\overline{\Lambda}_{122} &= 0
\end{aligned} \tag{32}$$

where

$$N_5 = \operatorname*{Tr}_{\{\tau_\ell\}} \left[\prod_{\ell=2}^{K} B(\tau_\ell x_\ell)\right] \frac{\exp\left(-x_1^2 \frac{\mu q}{2(1+\mu(1-q))}\right)}{\sqrt{1+\mu(1-q)}} H\left(\frac{-\tau_1 x_1 \sqrt{q}}{\sqrt{1+\mu(1-q)}\sqrt{1-q}}\right) \quad . \tag{33}$$

Therefore, we obtain

$$\Delta_{122}^{(Com)}(\alpha, K) \simeq -\frac{1}{2K^2} \tag{34}$$

when $K \gg 1$ and $\alpha = O(1)$. 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 $\mathcal{T}$.

VI. ANALYSIS OF $\mathcal{M}_D$ IN THE $K >> 1$ LIMIT

Let us now turn to the generalization problem. The Bayesian entropy $-s(r=1)$ is given by

$$S_G = \operatorname*{Extr}_q \left\{ \frac{q}{2} + \frac{1}{2}\ln(1-q) + 2\alpha \int \prod_\ell Dx_\ell \mathcal{H}(\{x_\ell\}) \ln \mathcal{H}(\{x_\ell\}) \right\} \tag{35}$$

where, as $r = 1$, $\mathcal{H}(\{x_\ell\})$ depends on q only (5). In the large K limit, the scaling of q is

$$q \simeq 1 - \frac{\pi^6 \Gamma^4}{2\alpha^4} \tag{36}$$

where Γ has been defined in (21). Therefore, the Bayesian entropy asymptotically equals $S_G = 2\ln\alpha$ and the generalization error decreases as $e_g = 2\Gamma/\alpha$. This proves that, contrary to the parity machine case [12,10], only a small fraction among the 2^P possible sets of outputs contribute to S_G and explains why the generalization curve is smooth around $\alpha_D \sim \sqrt{\ln K}$ [3], defined by an average over all sets of outputs. The typical entropy of internal representations is given by

$$\mathcal{M}_D = \operatorname*{Extr}_{q^*} \left\{ -\frac{1}{2}\ln(1-q^*) + \frac{1}{2}\ln(1-q) + \frac{1}{2}(q-q^*) + 2\alpha \int \prod_\ell \mathcal{H}(\{x_\ell\}) \ln \mathcal{H}(\{x_\ell\}) - 2\alpha K \int DxH\left(x\sqrt{\frac{q^*}{1-q^*}}\right) \ln H\left(x\sqrt{\frac{q^*}{1-q^*}}\right) \right\} \tag{37}$$

In the limit $1 \ll \alpha \ll K$, the internal overlap q^* scales as

$$q^* \simeq 1 - \frac{\pi^2 \Gamma^2}{2\alpha^2 K^2} \tag{38}$$

and the entropy of contributing internal representations reads

$$\mathcal{M}_D = \ln K - \ln\alpha \tag{39}$$

Therefore, α_D defined for the storage problem, and more generally the Vapnik–Chervonenkis dimension [9], are not relevant for the typical generalization properties of a large committee machine inferring a learnable rule. We can moreover note that above $\alpha_G \simeq K$, one single domain survives and the generalization error asymptotically decreases as $e_g = \Gamma/\alpha$ as is for finite K and large α [3,1]. To end with, the condition $q = q^*$ signaling that a unique volume is non empty gives back the estimated value of α_G.

VII. CONCLUSION

In this paper we have developed a complete analysis of the learning and generalization properties of large committee machines. Our approach – in which the weight space is partitioned according to the internal representations of the learning examples – allows us to derive the relevant entropies $\mathcal{N}_R$, $\mathcal{N}_D$ and $\mathcal{M}_D$ and successively to find the storage capacity of the model, to verify the stability of the solution and to study the rule inference capability. In ref. [1] we have discussed the physical and geometrical issues arising in the application of such a method to the learning and generalization theory of MLN. Here the chief results are the explicit derivation of the asymptotic storage capacity endowed with a detailed analysis of the stability of the solution and the derivation of the generalization cross–over at $\alpha = K$. From a methodological point of view, it is interesting to note that the RS computation of the distribution of volumes is very close to the one-step calculation of the Gardner volume. However, it is technically simpler and allows for instance the derivation of the asymptotic storage capacity of large committee machines while the same quantity seemed out of reach using the standard RSB computation [5].

REFERENCES

* Email: monasson@physique.ens.fr

† Email: zecchina@to.infn.it

[1] R. Monasson and R. Zecchina, *Phys. Rev. Lett.* **75**, 2432 (1995)

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

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

[4] E. Gardner, *J. Phys.* **A** 21, 257 (1988)
E. Gardner, B. Derrida, *J. Phys.* **A** 21, 271 (1988)

[5] E. Barkai, D. Hansel, I. Kanter, *Phys. Rev. Lett.* **65**, 2312 (1990)
E. Barkai, D. Hansel, H. Sompolinsky, *Phys. Rev. A* **45**, 4146 (1992)

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

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

[8] Formulae (5-6) in reference [1] are not correct : the power r must appear in the argument of the logarithm only. However, all asymptotic results derived in [1] are true when $K \gg 1$.

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

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

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

[12] T. Watkin, A. Rau, M. Biehl, *Rev. Mod. Phys.* **65**, 499 (1993)