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

Rémi Monasson\textsuperscript{1,*} and Riccardo Zecchina\textsuperscript{2,†}

\textsuperscript{1}INFN and Dipartimento di Fisica, P.le Aldo Moro 2, I-00185 Roma, Italy
\textsuperscript{2}INFN and Dipartimento di Fisica, Politecnico di Torino, C.so Duca degli Abruzzi 24, I-10129 Torino, Italy

(Received 13 January 1995)

We analytically derive the geometrical structure of the weight space in multilayer neural networks in terms of the volumes of couplings associated with the internal representations of the training set. In this framework, focusing on the parity and committee machines, we show how to deduce 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 $K$ of hidden units, whereas for finite $K$ a complete geometrical interpretation of symmetry breaking is given.

PACS numbers: 87.10.+e, 05.20.–y, 64.60.–i

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 and has been successfully applied to simple models of multilayer neural networks (MLN’s) to compute storage capacities and generalization errors \([2–5]\).

However, the geometrical picture of MLN’s weight space describing the role of internal representation (i.e., the way the input data are encoded inside the networks) and thus a unique “microscopic” frame allowing for the physical interpretation of the computational behavior, is lacking so far. This issue is deeply related to information theory and thus of potential interest both for biologically motivated models and for algorithms.

In this Letter we analytically derive such geometrical structure and show how it allows one to deduce and interpret the networks learning and generalization capabilities as well as to provide a unified point of view on apparently unrelated issues, such as the Mitchison-Durbin calculation \([6]\), replica symmetry breaking (RSB), and information theory.

We extend the usual Gardner’s approach to learning and generalization, by explicitly taking into account both the internal state variables and the interaction weights and by studying the volumes of couplings associated with the possible internal representations of the learning task. Though the method we adopt may, in principle, be applied to any MLN, we concentrate on the parity and committee machines which are the MLN’s most studied using statistical mechanics.

For the storage problem, we focus upon the volumes given the dominant contribution to Gardner’s total volume, whose number exp$(\mathcal{N}_D)$ is smaller than the total number exp$(\mathcal{N}_R)$ of nonempty volumes. For the parity and committee machines with $K \gg 1$ hidden units, $\mathcal{N}_D$ and $\mathcal{N}_R$ both vanish at $\frac{16}{\pi} \sqrt{\ln K}$ (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 provide a geometrical interpretation of RSB together with numerical results in the case of $K = 3$.

The inference of a learnable rule is studied along the same lines. We first reinterpret recent results \([5]\) 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 \([7]\) $d_{VC} \sim \sqrt{\ln K}$ and conjecture a crossover to lower generalization error for $\alpha = \alpha_\infty \sim K$.

In the following, we shall consider tree-like MLN’s, composed of $K$ nonoverlapping perceptrons with real-valued weights $J_{\ell i}$ and connected to $K$ sets of independent inputs $\xi_{\ell i}$ ($\ell = 1, \ldots, K$, $i = 1, \ldots, N/K$) \([3]\). The output $\sigma$ of the network is a binary function $f(\tau_1, \ldots, \tau_K)$ of the cells $\tau_i = \text{sgn} \left( \sum_{\ell} J_{\ell i} \xi_{\ell i} \right)$ in the first hidden layer. The set $\{\tau_i\}$ will be called hereafter the 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 $\text{sgn} \left( \sum_{\ell} \tau_\ell \right)$. The training set to be stored in the network includes $P = \alpha N$ patterns $\{\xi_{\ell i}\}$ and their corresponding outputs $\sigma_\mu$ ($\mu = 1, \ldots, P$). 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 $T = \{\tau^T_\ell\}$ with a corresponding nonzero volume

$$V_T = \int \prod_{\ell} dJ_{\ell i} \prod_{\mu} \theta(\sigma^{\mu} f(\{\tau^T_\ell\})) \prod_{\mu, \ell} \theta \left( \tau^T_\ell \sum_{i} J_{\ell i} \xi_{\ell i}^\mu \right),$$

where $\theta(\cdots)$ is the Heaviside function and the integral over the weights fulfills $\int dJ_{\ell i} = 1$. Gardner’s total volume is simply $V_G = \Sigma_T V_T$, and the critical capacity of the network is the value $\alpha_c$ of the maximal size of the training set such that $\ln V_G$ is finite, where the overbar 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_T’s as elementary “bricks” [8].

Once the canonical free energy
\[
g(r) = -\frac{1}{N} \ln \left( \sum_\mathcal{T} V'_\mathcal{T} \right) \tag{2}
\]
is known, one obtains the microcanonical entropy \( \mathcal{N}(w) \) (i.e., the logarithm of the typical number) of volumes \( V_T \), whose sizes are equal to \( w = \frac{1}{V} \ln V_T \) using the Legendre relations \( w_r = \frac{\delta g(r)}{\delta r} \) and \( \mathcal{N}(w_r) = -\delta g(r) \) [9]. The average over the patterns is performed using the replica trick for \( r \) integer, expecting that the final results remain valid for any real value of \( r \). There are \( r \) blocks (\( p = 1, \ldots, r \)) of \( n \) replicas (\( a = 1, \ldots, n \)). Thus the spin glass order parameters are the typical overlaps \( q_{\mathcal{T}\mathcal{B}}^{ap,b} = \frac{1}{N} \sum_{\ell \in \mathcal{T}} j_{\ell}^{ap} j_{\ell}^{b} \) between two weight vectors incoming onto the same hidden unit \( \ell (\ell = 1, \ldots, K) \) and their conjugate Lagrange multipliers \( q_{\ell}^{ap,b} \). Since all the hidden units are indistinguishable, we assume that at the saddle point \( q_{\mathcal{T}\mathcal{B}}^{ap,b} = q_{\mathcal{T}\mathcal{B}}^{ap,b} \) and \( \tilde{q}_{\ell}^{ap,b} = \tilde{q}_{\ell}^{ap,b} \) independently of \( \ell \). Within the replica symmetric (RS) ansatz [10], we find

\[
g(r) = \mathbb{E}_{\mathcal{Q}^a} \left[ \frac{1}{2r} \ln (1 - q_{\ast}) - \frac{1}{2r} \ln [1 - q_{\ast} + r(q_{\ast} - q)] - \frac{\alpha}{r} \int_{\mathcal{T}} D\mathcal{T} \ln \mathcal{H}((\mathcal{T})) \right], \tag{3}
\]

where \( \mathcal{H}((\mathcal{T})) = \text{Tr}[\mathcal{T}] \prod_{\ell} \int D\gamma \mathcal{H}(\{\gamma_{\ell}\mathcal{T} \} = +1) \) only, since the outputs \( \gamma_{\ell} \) can always be set equal to +1 at the cost of redefining the input patterns.

The whole distribution of volume sizes is available through \( g(r) \). When \( N \to \infty \), \( -\frac{1}{N} \ln \mathcal{N}(V_G) = -g(r = 0) \) is dominated by volumes of size \( w_{r=0} \) whose corresponding entropy (i.e., the logarithm of their number divided by \( N \)) is \( \mathcal{N}_D = \mathcal{N}(w_{r=0}) \). At the same time the most numerous volumes are those of smaller size \( w_{r>0} < \infty \) in the limit \( r \to 0 \) all the \( \mathcal{T} \) are counted irrespective 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 Eq. (3) using the above Legendre identities. In particular, \( q(r = 1) \) is the usual saddle-point overlap of the Gardner volume \( g(1) \) [2,3]. The vanishing condition for the entropies coincides with the zero volume condition for \( V_G \) and thus gives 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 \( I \) carried by the distribution of implementable internal representations \( \mathcal{T} \) about the weights,

\[
I = -\sum_{\mathcal{T}} V_{\mathcal{T}} \ln \frac{V_{\mathcal{T}}}{V_G}, \tag{4}
\]
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 [6] geometrical calculation is simply an upper (and decoder-independent) bound on \( \mathcal{N}_D \) and that our approach allows us to compute systematically the corrections to this bound.

Figure 1 displays the RS entropy \( \mathcal{N}_D \) as a function of \( \alpha \) for both the parity and committee machines with \( K = 3 \) hidden units. This entropy vanishes at a critical value \( \alpha_D \) of the size of the training set. Numerically, we find \( \alpha_D \approx 3.8 \) and 2.9 for the parity and the committee machines, respectively [11]. Being the entropy of a discrete system, \( \mathcal{N}_D \) cannot be negative and therefore \( \alpha_D \) is an upper bound of the size of the training set \( \alpha_{\text{RSB}} \) [12], where the replica symmetry breaking occurs for both \( \mathcal{N}_D \) and \( V_G \) [9]. When \( \alpha < \alpha_{\text{RSB}} \), the RS assumption is exact, whereas \( \mathcal{N}_D \) is positive, showing that the number of internal representation volumes contributing to \( V_G \) is exponentially large with \( N \). \( \alpha \) 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 \( V_T \). The behavior of \( q_{\ast} \) vs \( \alpha \) is shown in the inset of Fig. 1. When choosing randomly two weight vectors storing the training set, the probability that they belong to the same set \( V_T \) 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 \( \alpha_D \) may be obtained analytically. We find that \( q = 0 \) and \( \alpha = 1 - 128/\pi^2 \alpha^2 \) for the parity and the committee machines, respectively, and that \( q_{\ast} = 1 - \pi^2 \alpha^2/2 \alpha^2 K^2 \) in both cases with \( \Gamma = -1/[\sqrt{\pi} \int du \mathcal{H}(u) \ln \mathcal{H}(u)] \approx 0.62 \).

The corresponding entropies \( \mathcal{N}_{D_{\text{par}}} = \ln K - \alpha \ln 2 \) and \( \mathcal{N}_{D_{\text{com}}} = \ln K - \pi^2 \alpha^2/256 \) vanish at \( \alpha_{\text{par}} = \ln K \) and \( \alpha_{\text{com}} \approx \frac{16}{\pi} \sqrt{\ln K} \).
The parity machine with $K \geq 3$, a locally stable saddle-point solution leads to $N_R^{(\text{par})} = \alpha K \ln(\alpha K) - (\alpha K - 1)\ln(\alpha K - 1) - \alpha \ln 2$, which exactly saturates the upper bound derived by Mitchison and Durbin [6]. In the case of the committee machine, a simple analytical expression for $N_R^{(\text{com})}$ is not available for finite $K$. Once more in Fig. 1, we report the numerical results concerning the RS calculations of $N_R$ for both machines with $K = 3$. The value $\alpha_R$ at which $N_R$ vanishes should satisfy the obvious inequality $\alpha_D \leq \alpha_R \leq \alpha_c$; the RS approximation, however, overestimates $\alpha_R$ leading to an expression which is slightly larger than the one step value of $\alpha_c$ [14]. This is evidence for the necessity of RSB to compute $N_R$ exactly for finite $K$.

When $K \gg 1$, $N_R(\alpha_R)$ is asymptotically equal to $N_D(\alpha_D)$. In the case of the parity machine $\alpha_D$ and $\alpha_R$ also coincide with the known value of $\alpha_c = \frac{\ln K}{\ln 2}$ [3]. We expect the same equality ($\alpha_D = \alpha_R = \alpha_c = \frac{\ln K}{\ln 2}$) to hold in the case of the committee machine.

In order to show that the RS solution of $N_R$ is asymptotically correct, we have checked its local stability with respect to fluctuations of the order parameter matrices. This stability calculation will be displayed in detail in [13]. We find that there is no need to break the symmetry inside a single volume described by the RS order parameter $q$, whereas the RS ansatz $q = q^{\text{sp}}$ between two distinct volumes is unstable for any finite $K$. However, this instability decreases with increasing $K$. For both machines, our RS solution is marginally stable when $K \to \infty$ and should therefore become exact in this limit.

Not only storage but also generalization abilities strongly depend on the weight space structure. When a “student” network infers a learnable rule (i.e., generated by a “teacher” network endowed with an identical architecture), its weight space (or version space [15]) progressively shrinks when increasing the number of examples. In the simple perceptron case, the version space is connected and the typical generalization error done by the student goes to zero as its overlap with the teacher increases. The situation is more complex in MLN [16], where the alignment of the student along the teacher becomes more difficult due to the existence of the separated components they belong to. To improve our understanding of this competition between the shrinking of the volumes and the occurrence of a large number of them, we have modified our approach to the case of deterministic input-output mappings.

In the Bayesian framework—where all target rules are weighted with their a priori probabilities—the generalization properties are derived through the knowledge of the entropy $S_D = -\frac{1}{N} \sum_{\tau} V^\tau \ln \left( \sum_{\tau} V^\tau \right)$ [15]. Therefore the free energy generating the distribution of the “sizes” of the internal representation volumes $V^\tau$ must now be replaced by

$$s(r) = -\frac{1}{Nr} \sum_{\tau} V^\tau \ln \left( \sum_{\tau} V^\tau \right),$$

(5)
and the logarithm $\mathcal{M}_D$ of the number of the internal representation volumes contributing to the Bayesian entropy $S_G = s(1)$ is given by $\mathcal{M}_D = \frac{q}{2} \ln r - \frac{1}{2}$ [17]. In the case of the parity machine, it has been found [5] that there exists a critical value $\alpha_0$ of the size of the training set separating a high generalization error ($\epsilon_q = \frac{1}{2}$ when $\alpha < \alpha_0$) regime from a low generalization error phase ($\epsilon_q = \frac{1}{2}$ for large $\alpha$) [5]. This transition may be geometrically understood along the lines developed above. Computing $s(r)$ within the RS ansatz, we find

$$s(r) = \text{Exr} \left[ \frac{1 - r}{2} \ln (1 - q_s) - \frac{1}{2r} \right]$$

$$\times \ln[1 + q_s + r(q_s - q)] - \frac{q}{2[1 + (r - 1)q_s] - 2\alpha \ln f(q_s)}$$

$$= \frac{2\alpha}{f(q_s)} \int \frac{\partial \mathcal{G}}{\partial \mathcal{W}} \int d\mathcal{W} \mathcal{H}([x_\mathcal{W}] \ln \mathcal{H}([x_\mathcal{W}])$$

$$= f(q_s)[1 - \frac{\alpha}{2}]$$

with $f(q_s) = [2 \int d\mathcal{G} \mathcal{H}(z, \sqrt{\frac{q_s}{1 - q_s^2}})]^K$ and $q_s(r) = (1 - r) \ln (1 - q_s) - \ln (1 + q_s - q)$. In the large $K$ limit, we find that $q = 0$, $\mathcal{M}_D(\text{par}) = \ln K - \alpha \ln 2$ for $\alpha < \alpha_0 = \frac{\ln 2}{2\alpha}$, and $q = q_s$, $\mathcal{M}_D(\text{par}) = 0$ for $\alpha > \alpha_0$. Thus, below $\alpha_0$, 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: $\epsilon_q = \frac{1}{2}$. Above $\alpha_0$, since only one internal representation survives, the student has fallen down into the teacher volume: $q = q_s$, and $\epsilon_q = \frac{1}{2}$. When $\alpha < \alpha_0$, $S_G(\text{par}) = \alpha \ln 2$, meaning that all the sets of outputs are equiprobable. Choosing them with a probability $V_G([\sigma])$ is then equivalent to drawing them randomly. This is the reason why $\alpha_D$ 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 $\alpha_0$ depends upon RSB effects for finite $K$ [18], while the asymptotic RS expression of $\epsilon_q$ ought to be exact (see also [19]).

Turning to the committee machine, a calculation of the Bayesian entropy $S_G$ similar to [4] leads to the following results when $K \gg 1$. The typical teacher-student overlap $q$ decreases as $1 - \pi^2 \Gamma^4/(2\alpha^4)$ giving an entropy $S_G(\text{com}) = 2 \ln \pi$ and $\epsilon_q = \frac{2}{4\pi^2}$. This shows that, at variance with the parity machine case, only a small fraction among the $2^p$ possible sets of outputs contribute to $S_G(\text{com})$ and explains why the generalization curve is smooth for $\alpha = \sqrt{\ln K}$ (which is the order of magnitude of $d_{VC}$). We find $\mathcal{M}_D(\text{com}) = \ln K - \ln \pi$, confirming that $\alpha_D$ and thus $d_{VC}$ is not relevant to the computation of the typical generalization error. At $\alpha_0 \sim K$, only a single internal representation subsists and beyond this critical stage of the training set the generalization error should equal $\epsilon_q = \frac{1}{2}$ as is for finite $K$ and large $\alpha$ [4]. Note that the order of magnitude of $\alpha_0$ is corroborated by the condition $q = q_s$. One has to fulfill once a unique $V_T$ remains nonempty [20]. A rigorous proof of the presence of this crossover (from $\epsilon_q = 0$ to $\epsilon_q = \frac{1}{2}$) at $\alpha_0$ would, however, require us to extend the validity of our calculation to the regime $1 \ll \alpha \ll K$.

We are grateful to N. Brunel, M. Budinich, M. Ferrero, and D. O’Kane for discussions. This work has been supported by the EEC, Contract No. CHRX-CT-920063 and by an Elsevier-ISIS grant.

*Electronic address: monasson@roma1.infn.it

*Electronic address: zechcina@to.infn.it

[7] Indeed, from definition (1), the set of weights $\{J_{il}\}$ contributing to a given $V_T$ is convex (or empty). Though the number of distinct $V_T$ contained in each connected components of $V_G$ depends on the decoder under study (e.g., exactly one for the parity machine), the labeling of the different subsets of $\mathcal{G}$ with the internal representation does capture the main features of the geometry of the coupling space.
[10] For comparison, the storage capacities obtained with the one step RSB ansatz are $\alpha_e = 5$ and 3, respectively [3].
[12] It is indeed known that $\alpha_{RSB} = 3.2$ and 1.8 for the parity and the committee machines, respectively [3].
[14] For the parity and committee machines with $K = 3$ we find $\alpha_0 = 5.4$ and 5.5, respectively.
[17] The replica calculation of $s(r)$ technically differs from the computation of $g(r)$ only by taking the limit $n \rightarrow 1$ instead of $n \rightarrow 0$ [5,13].
[18] $\alpha_0$ is indeed rigorously equal to $\alpha_{RSB}$ [3,5,13].
[20] $g^*(r)$, the internal overlap of the domains, is decoder independent and decreases as $q_s = 1 - \pi^2 \Gamma^4/(2\alpha^4)$ for large $\alpha$ and $\alpha < \alpha_0$.