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

Rémi Monasson[1,*] and Riccardo Zecchina[2,†]
[1]INFN and Dipartimento di Fisica, P. le Aldo Moro 2, I-00185 Roma, Italy
[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 giving 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 $\frac{\ln K}{\ln 2}$ and $\frac{16}{\pi}\sqrt{\ln K}$ (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] $d_{\text{VC}} \sim \sqrt{\ln K}$ and conjecture a crossover to lower generalization error for $\alpha \equiv \alpha_{\text{co}} \sim K$.

In the following, we shall consider treelike 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 σ of the network is a binary function $f(\tau_1,\ldots,\tau_K)$ of the cells $\tau_\ell = \text{sgn}\,(\sum_i J_{\ell i}\xi_{\ell i})$ in the first hidden layer. The set $\{\tau_\ell\}$ 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 $\text{sgn}\,(\sum_\ell \tau_\ell)$. The training set to be stored in the network includes $P = \alpha N$ patterns $\{\xi_{\ell i}^\mu\}$ and their corresponding outputs σ^μ ($\mu = 1,\ldots,P$). 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 $\mathcal{T} = \{\tau_\ell^\mu\}$ with a corresponding nonzero volume

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

where $\theta(\cdots)$ 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 $\overline{\ln V_G}$ is finite, where the overbar denotes the average over the patterns and their corresponding outputs [2]. Moreover, the partition of

0031-9007/95/75(12)/2432(4)$06.00

V_G into connected components may be naturally obtained using the $V_{\mathcal{T}}$'s as elementary "bricks" [8].

Once the canonical free energy

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

is known, one obtains the microcanonical entropy $\mathcal{N}(w)$ (i.e., the logarithm of the typical number) of volumes $V_{\mathcal{T}}$, whose sizes are equal to $w = \frac{1}{N}\ln V_{\mathcal{T}}$ using the Legendre relations $w_r = \frac{\partial[rg(r)]}{\partial r}$ and $\mathcal{N}(w_r) = -\frac{\partial g(r)}{\partial(1/r)}$ [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 ($\rho = 1,\dots,r$) of n replicas ($a = 1,\dots,n$). Thus the spin glass order parameters are 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 ($\ell = 1,\dots,K$) and 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 $q_\ell^{a\rho,b\lambda} = q^{a\rho,b\lambda}$ and $\hat{q}_\ell^{a\rho,b\lambda} = \hat{q}^{a\rho,b\lambda}$ independently of ℓ. Within the replica symmetric (RS) ansatz [10], we find

$$\begin{aligned} g(r) = \operatorname*{Extr}_{q,q_*}\Bigg\{ &\frac{1-r}{2r}\ln(1-q_*) - \frac{1}{2r}\ln[1-q_*+r(q_*-q)] \\ &- \frac{q}{2[1-q_*+r(q_*-q)]} - \frac{\alpha}{r}\int\prod_\ell Dx_\ell \ln\mathcal{H}(\{x_\ell\})\Bigg\}, \end{aligned} \tag{3}$$

where $\mathcal{H}(\{x_\ell\}) = \mathrm{Tr}_{\{\tau_\ell\}}\prod_\ell\int Dy_\ell H[(y_\ell\sqrt{q_*-q} + \tau_\ell x_\ell\sqrt{q})/\sqrt{1-q_*}]^r$. 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 [2,9]. 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 Eq. (3), the sum $\mathrm{Tr}_{\{\tau_\ell\}}$ 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 volume sizes is available through $g(r)$. 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 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 the implementable internal representations $\mathcal{T}$ about the weights,

$$I = -\sum_{\mathcal{T}}\frac{V_{\mathcal{T}}}{V_G}\ln\frac{V_{\mathcal{T}}}{V_G}, \tag{4}$$

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 $\mathcal{N}_R$ 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 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 [11]. 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} [12], where the replica symmetry breaking occurs for both $\mathcal{N}_D$ and V_G [9]. When $\alpha < \alpha_{\mathrm{RSB}}$, 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. 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 behavior of q_* vs α 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 $V_{\mathcal{T}}$ 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 $q \simeq 1 - 128/\pi^2\alpha^2$ for the parity and the committee machines, respectively, and that $q_* \simeq 1 - \pi^2\Gamma^2/2\alpha^2K^2$ in both cases with $\Gamma = -1/[\sqrt{\pi}\int du\,H(u)\ln H(u)] \simeq 0.62$. The corresponding entropies $\mathcal{N}_D^{(\mathrm{par})} \simeq \ln K - \alpha\ln 2$ and $\mathcal{N}_D^{(\mathrm{com})} \simeq \ln K - \pi^2\alpha^2/256$ vanish at $\alpha_D^{(\mathrm{par})} \simeq \frac{\ln K}{\ln 2}$ and $\alpha_D^{(\mathrm{com})} \simeq \frac{16}{\pi}\sqrt{\ln K}$.

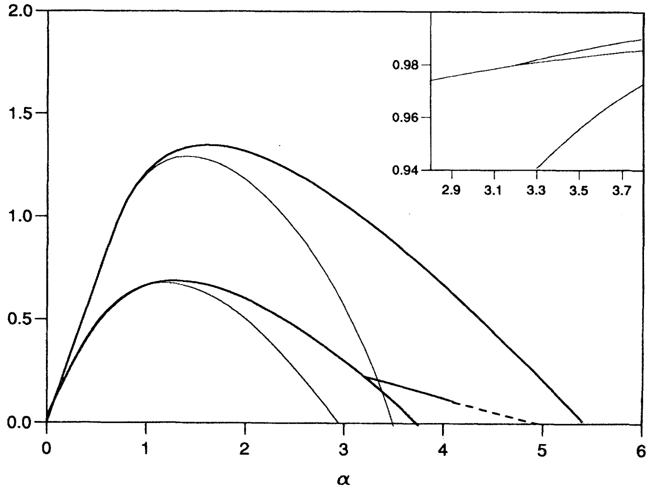

FIG. 1. $\mathcal{N}_R$ (upper curves) and $\mathcal{N}_D$ (lower curves) for the parity machine (bold) and the committee machine (light), with $K = 3$ hidden units. Inset: q_1, q_*, q_*' (lower, middle, and upper curves, respectively) vs α for the parity machine (q_1 starts at $\alpha = \alpha_{\text{RSB}}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

When $\alpha > \alpha_{\text{RSB}}$, the computation of $\mathcal{N}_D$ requires the introduction, at the first stage of RSB, of four order parameters q_*', q_0, q_1, and m: q_*' is the internal overlap of the internal representation volumes and q_0, q_1, and m are simply the usual parameters arising in the one step Gardner's computation [13]. Above α_{RSB}, there exist a finite number of big regions with mutual overlap q_0. Each region contains an exponential number of volumes of internal overlap q_*' and typically are separated by an overlap q_1. The number of such regions may be roughly estimated by $\frac{1}{1-m}$ [10]. We have checked this geometrical scenario numerically 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 $\alpha > \alpha_{\text{RSB}}$.

We conjecture that on 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 subregions characterized by $q(x, \alpha)$ [10]. 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 [3].

The RS calculation of $\mathcal{N}_R$ for both machines leads the to following general result. 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. For the parity machine with $K \geq 3$, a locally stable saddle-point solution leads to $\mathcal{N}_R^{(\text{par})} = \alpha K \ln(\alpha K) - (\alpha K - 1)\ln(\alpha K - 1) - \alpha \ln 2$, which exactly saturates the upper bound derived by Mitchison and Durbin [6]. In the case of the committee machine, a simple analytical expression for $\mathcal{N}_R^{(\text{com})}$ is not available for finite K. Once more in Fig. 1, we report the numerical results concerning the RS calculations of $\mathcal{N}_R$ 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 [14]. This is evidence for the necessity of RSB to compute $\mathcal{N}_R$ exactly for finite K.

When $K \gg 1$, $\mathcal{N}_R$ (α_R) is asymptotically equal to $\mathcal{N}_D$ (α_D). In the case of the parity machine α_D and α_R also coincide with the known value of $\alpha_c = \frac{\ln K}{\ln 2}$ [3]. We expect the same equality ($\alpha_D = \alpha_R = \alpha_c = \frac{16}{\pi}\sqrt{\ln K}$) 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. 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 $q = q^{a\rho,b\lambda}$ 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 $K \longrightarrow \infty$ 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 $S_G = -\frac{1}{N}\overline{V_G \ln V_G}$ [15]. Therefore the freeenergy generating the distribution of the "sizes" of the internal representation volumes $V_{\mathcal{T}}$ must now be replaced by

$$s(r) = -\frac{1}{Nr}\overline{\sum_{\mathcal{T}} V_{\mathcal{T}}^r \ln\Big(\sum_{\mathcal{T}} V_{\mathcal{T}}^r\Big)}, \qquad (5)$$

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 $\mathcal{M}_D = \frac{\partial s}{\partial r}|_{r=1}$ [17].

In the case of the parity machine, it has been found [5] that there exists a critical value α_0 of the size of the training set separating a high generalization error ($\epsilon_g = \frac{1}{2}$ when $\alpha < \alpha_0$) regime from a low generalization error phase ($\epsilon_g = \frac{\Gamma}{\alpha}$ for large α) [5]. This transition may be geometrically understood along the lines developed above. Computing $s(r)$ within the RS ansatz, we find

$$s(r) = \operatorname*{Extr}_{q}\left\{\frac{1-r}{2r}\ln(1-q_*) - \frac{1}{2r} \times \ln[1 - q_* + r(q_* - q)] - \frac{q}{2[1+(r-1)q_*]} - \frac{2\alpha}{f(q_*)}\int \prod_{\ell} Dx_\ell \mathcal{H}(\{x_\ell\})\ln\mathcal{H}(\{x_\ell\})\right\}, \tag{6}$$

with $f(q_*) = [2\int Dz H(z\sqrt{q_*}/\sqrt{1-q_*})^r]^K$ and $q_*(r)$ is the saddle-point overlap of $s_0(r) = (1-r)\ln(1-q_*) - \ln[1+(r-1)q_*] - 2\alpha\ln f(q_*)$ [18]. In the large K limit, we find that $q = 0$, $\mathcal{M}_D^{(\mathrm{par})} \simeq \ln K - \alpha\ln 2$ for $\alpha < \alpha_0 = \frac{\ln K}{\ln 2}$, and $q = q_*$, $\mathcal{M}_D^{(\mathrm{par})} = 0$ 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 $\alpha < \alpha_0$, $S_G^{(\mathrm{par})} = \alpha\ln 2$, 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 depends upon RSB effects for finite K [18], while the asymptotic RS expression of ϵ_g 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 $K \gg \alpha \gg 1$. The typical teacher-student overlap q decreases as $1 - \pi^6\Gamma^4/2\alpha^4$ giving an entropy $S_G^{(\mathrm{com})} \simeq 2\ln\alpha$ 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 $S_G^{(\mathrm{com})}$ and explains why the generalization curve is smooth for $\alpha \simeq \sqrt{\ln K}$ (which is the order of magnitude of d_{VC}). We find $\mathcal{M}_D^{(\mathrm{com})} \simeq \ln K - \ln\alpha$, confirming that α_D (and thus d_{VC}) is not relevant to the computation of the typical generalization error. At $\alpha_{\mathrm{co}} \sim K$, 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 [4]. Note that the order of magnitude of α_{co} is corroborated by the condition $q = q_*$ one has to fulfill once a unique $V_{\mathcal{T}}$ remains nonempty [20]. A rigorous proof of the presence of this crossover (from $\epsilon_g = 2\frac{\Gamma}{\alpha}$ to $\epsilon_g = \frac{\Gamma}{\alpha}$) at α_{co} would, however, require us 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. This work has been supported by the EEC, Contract No. CHRX-CT-920063 and by an Elsag-ISI grant.

*Electronic address: monasson@roma1.infn.it
†Electronic address: zecchina@to.infn.it

[1] T.M. Cover, IEEE Trans. Electron. Comput. **14**, 326 (1965).
[2] E. Gardner, J. Phys. A **21**, 257 (1988); E. Gardner and B. Derrida, J. Phys. A **21**, 271 (1988).
[3] E. Barkai, D. Hansel, and I. Kanter, Phys. Rev. Lett. **65**, 2312 (1990); E. Barkai, D. Hansel, and H. Sompolinsky, Phys. Rev. A **45**, 4146 (1992); A. Engel, H.M. Kohler, F. Tschepke, H. Vollmayr, and A. Zippelius, Phys. Rev. A **45**, 7590 (1992).
[4] H. Schwarze and J. Hertz, Europhys. Lett. **20**, 375 (1992).
[5] M. Opper, Phys. Rev. Lett. **72**, 2113 (1994).
[6] G.J. Mitchison and R.M. Durbin, Biol. Cybernet. **60**, 345 (1989).
[7] V.N. Vapnik, *Estimation of Dependences Based on Empirical Data* (Springer-Verlag, New York, 1982).
[8] Indeed, from definition (1), the set of weights $\{J_{\ell i}\}$ contributing to a given $V_{\mathcal{T}}$ is convex (or empty). Though the number of distinct $V_{\mathcal{T}}$ 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 V_G with the internal representation does capture the main features of the geometry of the coupling space.
[9] R. Monasson and D. O'Kane, Europhys. Lett. **27**, 85 (1994).
[10] M. Mézard, G. Parisi, and M.A. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).
[11] For comparison, the storage capacities obtained with the one step RSB ansatz are $\alpha_c \simeq 5$ and 3, respectively [3].
[12] It is indeed known that $\alpha_{\mathrm{RSB}} = 3.2$ and 1.8 for the parity and the committee machines, respectively [3].
[13] R. Monasson and R. Zecchina (to be published).
[14] For the parity and committee machines with $K = 3$ we find $\alpha_R = 5.4$ and 3.5, respectively.
[15] T. Watkin, A. Rau, and M. Biehl, Rev. Mod. Phys. **65**, 499 (1993).
[16] M. Opper, Phys. Rev. E **51**, 3613 (1995).
[17] The replica calculation of $s(r)$ technically differs from the computation of $g(r)$ only by taking the limit $n \longrightarrow 1$ instead of $n \longrightarrow 0$ [5,13].
[18] α_0 is indeed rigorously equal to α_{RSB} [3,5,13].
[19] A. Engel and L. Reimers, Europhys. Lett. **28**, 531 (1994).
[20] $q^*(r)$, the internal overlap of the domains, is decoder independent and decreases as $q_* \simeq 1 - \pi^2\Gamma^2/2\alpha^2K^2$ for large α, K.