Weight space structure and the storage capacity of a fully connected committee machine

Yuansheng Xiong and Jong-Hoon Oh*
Department of Physics, Pohang Institute of Science and Technology, Hyoja San 31, Pohang, Kyongbuk 790-784, Korea

Chulan Kwon
Department of Physics, Myong Ji University, Yongin, Kyonggi 449-728, Korea
(Received 10 June 1997)

We study the storage capacity of a fully connected committee machine with a large number K of hidden nodes. The storage capacity is obtained by analyzing the geometrical structure of the weight space related to the internal representation. By examining the asymptotic behavior of order parameters in the limit of large K, the storage capacity α_c is found to be proportional to $K\sqrt{\ln K}$ up to the leading order. This result satisfies the mathematical bound given by Mitchison and Durbin [Biol. Cybern. **60**, 345 (1989)], whereas the replica-symmetric solution in a conventional Gardner approach [Europhys. Lett. **41**, 481 (1987); J. Phys. A **21**, 257 (1988)] violates this bound. [S1063-651X(97)02310-6]

PACS number(s): 87.10.+e, 05.50.+q, 64.60.Cn

Since Gardner's pioneering work on the storage capacity of a single-layer perceptron [1], there have been numerous efforts to use the statistical-mechanics formulation to study feed-forward neural networks. The storage capacity of multilayer neural networks has been of particular interest, together with the generalization problem. Barkai, Hansel, and Kanter [2] studied a parity machine with a nonoverlapping receptive field of continuous weights within a one-step replica symmetry breaking (RSB) scheme and their result agrees with a mathematical bound previously found by Mitchison and Durbin [3]. Subsequently Barkai, Hansel, and Sompolinsky [4] and Engel *et al.* [5] have studied the committee machine, which is closer to the multilayer perceptron architecture and is most frequently used in real-world applications. Though they have derived many interesting results, particularly for the case of a finite number of hidden units, it was found that their the replica-symmetric (RS) result violates the Mitchison-Durbin (MD) bound in the limit where the number of hidden units K is large.

Recently, Monasson and O'Kane [6] proposed a statistical-mechanics formalism that can analyze the weight space structure related to the internal representations of hidden units. It was applied to single-layer perceptrons [7–9] as well as multilayer networks [10–12]. Monasson and Zecchina [10] have successfully applied this formalism to the case of both committee and parity machines with nonoverlapping receptive fields (NRFs) [10]. They suggested that analysis of the RS solution under this statistical-mechanics formalism can yield results just as good as the one-step RSB solution in the conventional Gardner method.

In this paper we apply this formalism for a derivation of the storage capacity of a fully connected committee machine, which is also called a committee machine with overlapping receptive field and is believed to be a more relevant architecture. In particular, we obtain the value of the critical storage capacity in the limit of large K, which satisfies the MD bound. It also agrees with a recent one-step RSB calculation, using the conventional Gardner method, to within a small difference of a numerical prefactor [13]. Finally, we will briefly discuss the fully connected parity machine.

We consider a fully connected committee machine with N input units, K hidden units, and one output unit, where weights between the hidden units and the output unit are set to 1. The network maps input vectors $\{x_i^\mu\}$, where $\mu=1,\ldots,P$, to output y^μ as

$$y^\mu=\mathrm{sgn}\left(\sum_{j=1}^{K} h_j^\mu\right)=\mathrm{sgn}\left[\sum_{j=1}^{K}\ \mathrm{sgn}\left(\sum_{i=1}^{N} W_{ji}x_i^\mu\right)\right], \quad (1)$$

where W_{ji} is the weight between the ith input node and the jth hidden unit. $h_j^\mu\equiv\mathrm{sgn}(\Sigma_{i=1}^{N}W_{ji}x_i^\mu)$ is the jth component of the internal representation for input pattern $\{x_i^\mu\}$. We consider continuous weights with spherical constraint, $\Sigma_i^N W_{ji}=N$.

Given $P=\alpha N$ patterns, the learning process in a layered neural network can be interpreted as the selection of cells in the weight space corresponding to a set of suitable internal representations $\boldsymbol{h}=\{h_j^\mu\}$, each of which has a nonzero elementary volume defined by

$$V_{\boldsymbol{h}}=\mathrm{Tr}_{\{W_{ji}\}}\prod_\mu\ \Theta\left(y^\mu\sum_j\ h_j^\mu\right)\prod_{\mu,j}\ \Theta\left(h_j^\mu\sum_i\ W_{ji}x_i^\mu\right), \quad (2)$$

where $\Theta(x)$ is the Heaviside step function. Gardner's volume V_G, that is, the volume of the weight space that satisfies the given input-output relations, can be written as the sum of the cells over all internal representations

$$V_G=\sum_{\boldsymbol{h}}\ V_{\boldsymbol{h}}. \quad (3)$$

The method developed by Monasson and his collaborators [6,10] is based on the analysis of the detailed internal structure, that is, how Gardner's volume V_G is decomposed into elementary volumes $V_{\boldsymbol{h}}$ associated with a possible internal

*Present address: Bell Laboratories, Lucent Technologies, 600 Mountain Avenue, Murray Hill, NJ 07974.

representation. The distribution of the elementary volumes can be derived from the free energy

$$g(r)=-\frac{1}{Nr}\left\langle\left\langle \ln\left(\sum_{\boldsymbol{h}} V_{\boldsymbol{h}}^{r}\right)\right\rangle\right\rangle, \tag{4}$$

where $\langle\langle\ \rangle\rangle$ denotes the average over patterns. The entropy $\mathcal{N}[w(r)]$ of the volumes whose average sizes are equal to $w(r)=-1/N\ \ln\langle\langle V_{\boldsymbol{h}}\rangle\rangle$ can be given by the Legendre relations

$$\mathcal{N}[w(r)]=-\frac{\partial g(r)}{\partial(1/r)},\quad w(r)=\frac{\partial[rg(r)]}{\partial r}, \tag{5}$$

respectively.

The entropies $\mathcal{N}_D=\mathcal{N}[w(r=1)]$ and $\mathcal{N}_R=\mathcal{N}[w(r=0)]$ are of most importance and will be discussed below. In the thermodynamic limit, $1/N\langle\langle\ln(V_G)\rangle\rangle=-g(r=1)$ is dominated by elementary volumes of size $w(r=1)$, of which there are $\exp(N\mathcal{N}_D)$. Furthermore, the most numerous elementary volumes have the size $w(r=0)$ and number $\exp(N\mathcal{N}_R)$. The vanishing condition for the entropies is related to the zero-volume condition for V_G and thus gives the storage capacity. We focus on the entropy $\mathcal{N}_D$ of elementary volumes dominating the weight space V_G.

The replicated partition function for the fully connected committee machine reads

$$\left\langle\left\langle\left(\sum_{\boldsymbol{h}} V_{\boldsymbol{h}}^{r}\right)^{n}\right\rangle\right\rangle=\left\langle\left\langle \mathrm{Tr}_{h_j^{\mu\alpha}}\mathrm{Tr}_{W_j^{\mu\alpha}}\prod_{\mu,\alpha}\Theta\left(\sum_j h_j^{\mu\alpha}\right)\prod_{\mu,j,\alpha,a}\Theta\left(h_j^{\mu\alpha}\sum_i W_{ji}^{\alpha a}x_i^{\mu}\right)\right\rangle\right\rangle, \tag{6}$$

with $a=1,\ldots,r$ and $\alpha=1,\ldots,n$. Unlike Gardner's conventional approach, we need two sets of replica indices for the weights.

For a fully connected machine, the overlaps between different hidden units should be taken into account, which makes this problem much more difficult than the treelike (NRF) architecture studied in Ref. [10]. We introduce the order parameters

$$Q_{jk}^{\alpha\beta ab}=\frac{1}{N}\sum_i W_{ji}^{\alpha a}W_{ki}^{\beta b}, \tag{7}$$

where the indices a,b originate from the integer power r of elementary volumes and α,β are the standard replica indices. The replica symmetry ansatz leads to five order parameters as

$$Q_{jk}^{\alpha\beta ab}=\begin{cases} q^* & (j=k,\ \alpha=\beta,\ a\neq b)\\ q & (j=k,\ \alpha\neq\beta)\\ c & (j\neq k,\ \alpha=\beta,\ a=b)\\ d^* & (j\neq k,\ \alpha=\beta,\ a\neq b)\\ d & (j\neq k,\ \alpha\neq\beta), \end{cases} \tag{8}$$

where q^* and q are, respectively, the overlaps between the weight vectors connected to the same hidden unit of the same ($\alpha=\beta$) and different ($\alpha\neq\beta$) replicas corresponding to the two different internal representations. The order parameters c, d^*, and d describe the overlaps between weights that are connected to different hidden units, of which c and d^* are the overlaps within the same replica, whereas d correlates different replicas.

Using a standard replica trick, we obtain

$$\begin{aligned} g(r)=\mathrm{Extr}_{q,q^*,c,d,d^*}\Bigg\{ &-\frac{1}{2}\Bigg[\frac{(K-1)(q-d)}{1-q^*+r(q^*-q)-[c-d^*+r(d^*-d)]}+\frac{K-1}{r}\ln\{1-q^*+r(q^*-q)-[c-d^*+r(d^*-d)]\}\\ &+\frac{q+(K-1)d}{1-q^*+r(q^*-q)+(K-1)[c-d^*+r(d^*-d)]}+\frac{1}{r}\ln\{1-q^*+r(q^*-q)+(K-1)[c-d^*+r(d^*-d)]\}+(K-1)\\ &\times\left(1-\frac{1}{r}\right)\ln(1-q^*-c+d^*)+\left(1-\frac{1}{r}\right)\ln[1-q^*+(K-1)(c-d^*)]\Bigg]\\ &-\frac{\alpha}{r}\int Dt_5\int\prod_j Dt_3^j\ln\left[\mathrm{Tr}_{h_j}\Theta\left(\sum_j h_j\right)\int Dt_4\int\prod_j Dt_1^j\left(\int Dt_2\prod_j H(\Omega_j)\right)^r\right]\Bigg\}, \end{aligned} \tag{9}$$

where we have posed $Dx=\exp(-x^2/2)dx/\sqrt{2\pi}$, $H(y)=\int_y^\infty Dx$, and

$$\Omega_j=\frac{\sqrt{q^*-q-d^*+d}\ t_1^j+(\sqrt{c-d^*}\ t_2+\sqrt{q-d}\ t_3^j+\sqrt{d^*-d}\ t_4+\sqrt{d}\ t_5)h_j}{\sqrt{1-q^*+d^*-c}}. \tag{10}$$

One may notice that the free energy evaluated at $r=1$ is reduced to the RS results obtained by the conventional method on the committee machine [4,5], which is independent of q^* and d^*. This means that the internal structure of the weight space is overlooked by conventional calculation of Gardner's volume. When we take the limit $r\to 1$, the free energy in Eq. (9) can be expanded as

$$g(r,q^*,q,c,d^*,d)=g(1,q,c,d)+(r-1)\left.\frac{\partial g(r,q^*,q,c,d^*,d)}{\partial r}\right|_{r=1}. \tag{11}$$

As noticed, $g(r,q^*,q,c,d^*,d)$ is the same as the RS free energy in Gardner's method. From the relation

$$\mathcal{N}_D=-\left.\frac{\partial g(r)}{\partial(1/r)}\right|_{r=1}=\left.\frac{\partial g(r)}{\partial r}\right|_{r=1}, \tag{12}$$

we obtain the explicit form of $\mathcal{N}_D$ as

$$\begin{aligned}\mathcal{N}_D=\frac{1}{2}\Bigg[&\frac{(K-1)(q-d)[q^*-q-(d^*-d)]}{(1-q-c+d)^2}+(K-1)\ln[1-q-(c-d)]-\frac{(K-1)[q^*-q-(d^*-d)]}{1-q-(c-d)}\\&+\frac{[q+(K-1)d][q^*-q+(K-1)(d^*-d)]}{[1-q+(K-1)(c-d)]^2}-(K-1)\ln(1-q^*-c+d^*)\\&-\ln[1-q^*+(K-1)(c-d^*)]+\ln[1-q+(K-1)(c-d)]-\frac{q^*-q+(K-1)(d^*-d)}{1-q+(K-1)(c-d)}\Bigg]\\&-\alpha\int Dt_5\int\prod_j Dt_3^j\frac{\mathrm{Tr}_{h_j}\Theta\left(\sum_j h_j\right)\int Dt_4\int\prod_j Dt_1^j\int Dt_2\prod_j H(\Omega_j)\ln\left[\int Dt_2\prod_j H(\Omega_j)\right]}{\mathrm{Tr}_{h_j}\Theta\left(\sum_j h_j\right)\int Dt_4\prod_j H(\Omega_j')}\\&+\alpha\int Dt_5\int\prod_j Dt_3^j\ln\left[\mathrm{Tr}_{h_j}\Theta\left(\sum_j h_j\right)\int Dt_4\prod_j H(\Omega_j')\right],\end{aligned} \tag{13}$$

with

$$\Omega_j'=\frac{\sqrt{c-d}\;t_4+\sqrt{q-d}\;t_3^j+\sqrt{d}t_5}{\sqrt{1-q+d-c}}. \tag{14}$$

In the case of the NRF committee machine, where each of the hidden units is connected to different input units, we do not observe a phase transition. A single solution is applicable for the whole range of α. In contrast, the phase-space structure of the fully connected committee machine is more complicated than that of the NRF committee machine. When a small number of input patterns are given, the system is in the permutation-symmetric (PS) phase [4,5,14,15], where the role of each hidden unit is not specialized. In the PS phase, Gardner's volume is a single connected region. The order parameters associated with different hidden units are equal to the corresponding ones associated with the same hidden unit. When a critical number of patterns is given, Gardner's volume is divided into many islands, each one of which can be transformed into other by permutation of hidden units. In the case of generalization problem [14], this is accompanied by the specialization of the hidden nodes and their receptive fields. This phenomenon is called permutation symmetry breaking (PSB). It induces a first-order phase transition and discontinuity of the learning curve. In the storage capacity problem, specialization of the role of each hidden unit is less obvious. However, separation of Gardner's volume characterizes the onset of specialization among hidden nodes, which leads to a better storage capacity. It has already been pointed out that the critical storage capacity is attained in the PSB phase [4,5] and our recent one-step replica symmetry breaking calculation confirmed this picture [13]. Therefore, we will focus on the analysis of the PSB solution near the storage capacity, in which $q^*,q\to 1$ and c, d^*, and d are of order $1/K$.

In particular, $q(r=1)$, $c(r=1)$, and $d(r=1)$ are reduced to the usual saddle-point solutions of the replica-symmetric expression of Gardner's volume $g(r=1)$ [4,5]. When K is large, the trace over all allowed internal representations can be evaluated similarly to Ref. [4]. The saddle-point equations for q^* and d^* are derived from the derivative of the free energy in the limit $r\to 1$, as in Eq. (11). The details of the self-consistent equations are not shown for space consideration. In the following, we only summarize the asymptotic behavior of the order parameters for large α:

$$1-q+d-c\sim\frac{128}{(\pi-2)^2}\frac{K^2}{\alpha^2}, \tag{15}$$

$$1-q+(K-1)(c-d)\sim\frac{32}{\pi-2}\frac{K}{\alpha^2}, \tag{16}$$

$$q+(K-1)d\sim\frac{\pi-2}{\alpha}, \tag{17}$$

$$1-q^*+d^*-c\sim\frac{\pi^2\Gamma^2}{2\alpha^2}, \tag{18}$$

$$1-q^*+(K-1)(c-d^*)\sim\frac{\pi^2\Gamma^2}{2\alpha^2}, \tag{19}$$

where $\Gamma=-[\sqrt{\pi}\int duH(u)\ln H(u)]^{-1}\simeq 0.62$.

It is found that all the overlaps between weights connecting different hidden units have scalings of $-1/K$, whereas the typical overlaps between weights connecting the same hidden unit approach one. The order parameters c, d, and d^* are negative, showing antiferromagnetic correlations between different hidden units, which implies that each hidden unit attempts to store patterns different from those of the others [4,5].

Finally, the asymptotic behavior of the entropy $\mathcal{N}_D$ in the large-K limit can be derived using the scalings given above. Near the storage capacity, $\mathcal{N}_D$ can be written, up to the leading order, as

$$\begin{aligned}\mathcal{N}_D&\simeq\frac{1}{2}\left[\frac{K}{1-q}+K\ln(1-q)-K\ln(1-q^*-c+d^*)\right]+\alpha\int Dt_5\int\prod_j Dt_3^j\ln\left[\mathrm{Tr}_{h_j}\Theta\left(\sum_j h_j\right)\int Dt_4\prod_j H(\Omega_j')\right]\\&\simeq\frac{1}{2}\left[\frac{K}{1-q}+2K\ln\frac{K}{\alpha}-2K\ln\frac{1}{\alpha}\right]-\frac{(\pi-2)^2}{128}\frac{\alpha^2}{K}\\&\simeq K\ln K-\frac{(\pi-2)^2\alpha^2}{256K}.\end{aligned} \tag{20}$$

Being the entropy of a discrete system, $\mathcal{N}_D$ cannot be negative. Therefore, $\mathcal{N}_D=0$ gives an indication of the upper bound of storage capacity, that is, $\alpha_c\sim[16/(\pi-2)]K\sqrt{\ln K}$. The storage capacity per synapse $16/(\pi-2)\sqrt{\ln K}$, satisfies the rigorous bound $\sim\ln K$ derived by Mitchison and Durbin [3], whereas the conventional RS result [4,5], which scales as $\sqrt{K}$, violates the MD bound.

Recently, we have studied this problem using a conventional Gardner approach in the one-step RSB scheme [13]. The result yields the same scaling with respect to K, but a coefficient smaller by a factor $\sqrt{2}$. In the present paper, we are dealing with the fine structure of version space related to internal representations. On the other hand, the RSB calculation seems to handle this fine structure in association with symmetry breaking between replicas. Although the physics of the two approaches seems to be somehow related, it is not clear which of the two can yield a better estimate of the storage capacity. It is possible that the present RS calculation does not properly handle the RSB picture of the system. Monasson and his co-workers reported that the Almeida-Thouless instability of the RS solutions decreases with increasing K, in the NRF case [10,11]. A similar analysis for the fully connected case certainly deserves further research. On the other hand, the one-step RSB scheme also introduces an approximation and possibly cannot fully explain the weight space structure associated with internal representations.

It is interesting to compare our result with that of the NRF committee machine along the same lines [10]. Based on the conventional RS calculation, Engel *et al.* suggested that the same storage capacity per synapse for both fully connected and NRF committee machines will be similar, as the overlap between the hidden nodes approaches zero [5]. While the asymptotic scaling with respect to K is the same, the storage capacity in the fully connected committee machine is larger than in the NRF one. It is also consistent with our result from one-step RSB calculation [13]. This implies that the small but nonezero negative correlation between the weights associated with different hidden units enhances the storage capacity. This may be good news for those people using a fully connected multilayer perceptron in applications.

From the fact that the storage capacity of the NRF parity machine is $\ln K/\ln 2$ [2,10], which saturates the MD bound, one may guess that the storage capacity of a fully connected parity machine is also proportional to $K\ln K$. It will be interesting to check whether the storage capacity per synapse of the fully connected parity machine is also enhanced compared to the NRF machine. Our recent calculation of the fully connected parity machine seems to support this expectation [16].

We thank I. Kanter, R. Monasson, A. Engel, M. Opper, M. Biehl, and J. Hertz for stimulating discussions and helpful comments. This work was partially supported by the Basic Science Special Program of POSTECH and the Korea Ministry of Education through the POSTECH Basic Science Research Institute. It was also supported by a nondirected fund from Korea Research Foundation (1997) and by KOSEF Grant No. 971-0202-010-2.

[1] E. Gardner, Europhys. Lett. **4**, 481 (1987); E. Gardner, J. Phys. A **21**, 257 (1988); E. Gardner and B. Derrida, *ibid.* **21**, 271 (1988).
[2] E. Barkai, D. Hansel, and I. Kanter, Phys. Rev. Lett. **65**, 2312 (1990).
[3] G. J. Mitchison and R. M. Durbin, Biol. Cybern. **60**, 345 (1989).
[4] E. Barkai, D. Hansel, and H. Sompolinsky, Phys. Rev. E **45**, 4146 (1992).
[5] A. Engel, H. M. Köhler, F. Tschepke, H. Vollmayr, and A. Zippeelius, Phys. Rev. E **45**, 7590 (1992).
[6] R. Monasson and D. O'Kane, Europhys. Lett. **27**, 85 (1994).
[7] B. Derrida, R. B. Griffiths, and A. Prugel-Bennett, J. Phys. A **24**, 4907 (1991).
[8] M. Biehl and M. Opper, in *Neural Networks: The Statistical Mechanics Perspective*, edited by Jong-Hoon Oh, Chulan Kwon, and Sungzoon Cho (World Scientific, Singapore, 1995).
[9] A. Engel and M. Weigt, Phys. Rev. E **53**, R2064 (1996).
[10] R. Monasson and R. Zecchina, Phys. Rev. Lett. **75**, 2432 (1995); **76**, 2205 (1996).
[11] R. Monasson and R. Zecchina, Mod. Phys. B **9**, 1887 (1996).
[12] S. Cocco, R. Monasson, and R. Zecchina, Phys. Rev. E **54**, 717 (1996).
[13] C. Kwon and J. H. Oh, J. Phys. A (to be published).
[14] K. Kang, J. H. Oh, C. Kwon, and Y. Park, Phys. Rev. E **48**, 4805 (1993).
[15] H. Schwarze and J. Hertz, Europhys. Lett. **21**, 785 (1993).
[16] Y. Xiong, C. Kwon, and J.-H. Oh (unpublished).