Storage capacity of the fully-connected committee machine

1997 J. Phys. A: Math. Gen. 30 L387

(http://iopscience.iop.org/0305-4470/30/11/007)

J. Phys. A: Math. Gen. **30** (1997) L387–L392. Printed in the UK PII: S0305-4470(97)81260-7

LETTER TO THE EDITOR

Storage capacity of the fully-connected committee machine

R Urbanczik†

Institut für theoretische Physik, Universität Würzburg, Am Hubland, D-97074 Würzburg, Germany

Received 24 January 1997

Abstract. The storage capacity, that is the number of patterns which can be stored per weight, is calculated for the fully-connected committee machine with real couplings and K hidden units from the vanishing of the entropy of the internal representations, and it is found to diverge as $(16/(\pi-2))\sqrt{\ln K}$.

In the last few years statistical mechanics has been applied with considerable success to the study of multilayer neural networks, e.g. [1–5]. However, for the committee machine, the theoretical architecture which is arguably most similar to the networks used in real-world applications, a satisfactory analysis of the capacity problem has not been available. Recently, progress has been made in this matter for a restricted architecture, the tree committee machine, by focusing on the number of implementable internal representations instead of on the Gardner volume [6, 7]. Here we extend this analysis to the fully-connected version and show that some claims made in the literature [3, 8] with regard to the capacity of this machine are incorrect.

The fully-connected committee machine consists of K hidden units, each characterized by a weight vector $J_j \in \mathbb{R}^N$ and computing $\tau_{J_j}(\xi) = \mathrm{sign}(J_j^T\xi)$ for an input $\xi \in \mathbb{R}^N$. The output $\sigma_J(\xi)$ of the committee is then determined by the majority of the hidden units, $\sigma_J(\xi) = \mathrm{sign}(\sum_{j=1}^K \tau_{J_j}(\xi))$. In the capacity problem one asks whether for a given set of αKN input/output pairs, $\mathcal{T} = \{(\xi^\mu, \sigma^\mu)\}_{\mu=1}^{\alpha KN}$, weight vectors J_j exist such that $\sigma_J(\xi^\mu) = \sigma^\mu$. In statistical mechanics one is interested in finding up to what critical size $\alpha_c(K)KN$ of $\mathcal{T}$ this problem will *typically* have a solution, that is with probability 1 in the thermodynamic limit for independent choices of (ξ^μ, σ^μ) from the normal distribution (ξ^μ) and, in the case of σ^μ, from the uniform distribution on $\{-1, 1\}$. A given set $\mathcal{T}$ will be realizable by the machine if an internal representation $\tau = \{\tau_j^\mu\}$ exists, which yields the desired outputs and is implementable. One is thus lead to consider the volume V_τ associated with such a representation:

$$V_\tau(\mathcal{T}) = \prod_{\mu=1}^{\alpha KN} \theta\left(\sigma^\mu \sum_{j=1}^K \tau_j^\mu\right) \int \mathrm{d}J\; p_0(J) \prod_{\mu=1}^{\alpha KN}\prod_{j=1}^K \theta(\tau_j^\mu J_j^T \xi^\mu). \tag{1}$$

Here θ is the Heaviside function and $p_0(J)$ is the uniform density on the set $\{J \in \mathbb{R}^{NK} : |J_j| = 1\}$. In the end we shall not be interested in the magnitude of the $V_\tau(\mathcal{T})$ but only whether

† E-mail address: urbanczik@ubaclu.unibas.ch

0305-4470/97/110387+06$19.50

they are zero or not, but the calculation of (1) is a useful intermediate step. In particular, we may consider

$$S(m) = \frac{1}{KN}\langle \ln \mathrm{Tr}_\tau V_\tau(\mathcal{T})^m \rangle_{\mathcal{T}} \tag{2}$$

where the trace runs over all internal representations. If $S(0) = \lim_{m\to 0} S(m) = 0$ for some $\alpha_{\mathrm{d}}(K)$ the entropy of the implementable internal representations (which give the correct output) is no longer extensive. For large K we expect $\alpha_{\mathrm{d}}(K) \simeq \alpha_{\mathrm{c}}(K)$ since the volume associated with any single internal representation shrinks to zero in this limit. Furthermore the calculations will yield that $S(0)$ is negative for $\alpha > \alpha_{\mathrm{d}}(K)$ and this is incompatible with the assumption that $\alpha_{\mathrm{c}}(K)$ is greater than $\alpha_{\mathrm{d}}(K)$.

$S(m)$ may be calculated by averaging $(\mathrm{Tr}_\tau V_\tau(\mathcal{T})^m)^n$ for integer n and m and using an analytical continuation in these parameters. Replicating (1) in this manner replaces the K integrals over the J_j by Knm integrals over J_j^{ab}, $a = 1, \ldots, m$ and $b = 1, \ldots, n$. The average over the sets $\mathcal{T}$ then leads to order parameters $q_{jk}^{aba'b'} = J_j^{ab\,T} J_k^{a'b'}$. We assume a site and replica symmetric (RS) parametrization of this matrix that is

$$\begin{aligned} q_{jk}^{aba'b'} = {} & K^{-1}[(1-\delta_{aa'})p_0 + \delta_{aa'}(1-\delta_{bb'})p_1 + \delta_{aa'}\delta_{bb'}p_2] \\ & + \delta_{jk}[(1-\delta_{aa'})q_0 + \delta_{aa'}(1-\delta_{bb'})q_1 + \delta_{aa'}\delta_{bb'}q_2] \end{aligned} \tag{3}$$

where by normalization $q_2 = 1 - p_2/K$. The RS assumption has been checked in [7] for the tree architecture: the RS saddlepoint was found unstable for finite K but marginally stable in the limit $K \to \infty$. Furthermore, (3) incorporates the scaling $p_l = \mathcal{O}(1)$ as $K \to \infty$, an assumption we shall recover in a self-consistent way later.

Using this parametrization, for large N the value of $S(m)$ is given by

$$S(m) = \operatorname*{extr}_{\{p_l\},\{q_l\}} \alpha G_r(\{p_l\},\{q_l\},m) + G_s(\{p_l\},\{q_l\},m). \tag{4}$$

By arguments similar to [9] the energy term G_r may be written as

$$G_r = \left\langle \ln \left\langle \mathrm{Tr}_\tau\, \theta\Big(\sum_{j=1}^K \tau_j\Big) \left\langle \prod_{j=1}^K \theta[\tau_j(ux_j + \bar{u}\bar{x} + vy_j + \bar{v}\bar{y} + wz_j + \bar{w}\bar{z})] \right\rangle_{\{z_j\}}^m \right\rangle_{\{y_j\}} \right\rangle_{\{x_j\}} \tag{5}$$

where the x_j, y_j, z_j are independent normally distributed real random variables and $\bar{x} = K^{-1}\sum_{j=1}^K x_j$, $\bar{y} = K^{-1}\sum_{j=1}^K y_j$, $\bar{z} = K^{-1}\sum_{j=1}^K z_j$. The prefactors controlling their contribution to the sum in the theta function are related to the order parameters by

$$\begin{aligned} u^2 &= q_0 & (u+\bar{u})^2 &= p_0 + q_0 \\ v^2 &= q_1 - q_0 & (v+\bar{v})^2 &= q_1 + p_1 - q_0 - p_0 \\ w^2 &= 1 - p_2/K - q_1 & (w+\bar{w})^2 &= 1 - p_2/K + p_2 - q_1 - p_1. \end{aligned} \tag{6}$$

The entropy term G_s in (4) can be written as

$$\begin{aligned} G_s &= \frac{K-1}{2K} F(u^2, v^2, w^2, m) + \frac{1}{2K} F((u+\bar{u})^2, (v+\bar{v})^2, (w+\bar{w})^2, m) \\ F(a,b,c,m) &= (m-1)\ln c + \ln(c+mb) + \frac{ma}{c+mb}. \end{aligned} \tag{7}$$

It is interesting to note that (4) is closely related to the calculation of the Gardner volume $V(\mathcal{T}) = \mathrm{Tr}_\tau V_\tau(\mathcal{T})$ with one step of replica symmetry breaking. In particular, one finds

$$\frac{1}{KN}\langle \ln V(\mathcal{T}) \rangle_{\mathcal{T}} = \operatorname*{extr}_{\{p_l\},\{q_l\},m} \frac{1}{m}[\alpha G_r^*(\{p_l\},\{q_l\},m) + G_s(\{p_l\},\{q_l\},m)]. \tag{8}$$

Here G_r^* is obtained from G_r by commuting the trace and the exponentiation with m. More explicitly let

$$f(\{Y_j\},\{\tau_j\}) = \left\langle \prod_j \theta[\tau_j(Y_j + wz_j + \bar{w}\bar{z})] \right\rangle_{\{z_j\}} \qquad \text{and} \qquad Y_j = ux_j + \bar{u}\bar{x} + vy_j + \bar{v}\bar{y} \tag{9}$$

then

$$G_r^* = \left\langle \ln \left\langle \left[\mathrm{Tr}_\tau \theta\left(\sum_j \tau_j\right) f(\{Y_j\},\{\tau_j\}) \right]^m \right\rangle_{\{y_j\}} \right\rangle_{\{x_j\}}. \tag{10}$$

In contrast to the internal representations approach, where we need to consider $m \to 0$, the extremal value of m must be obtained in (8). However, even for the latter, $m \to 0$ as the critical capacity is approached, and so this difference is only minor. Furthermore, as $\alpha \to \alpha_c(K)$ one expects $w, \bar{w} \to 0$, so consequently

$$\begin{aligned} \left[\mathrm{Tr}_\tau \theta\left(\sum_j \tau_j\right) f(\{Y_j\},\{\tau_j\}) \right]^m &\simeq \max_\tau \theta\left(\sum_j \tau_j\right) f(\{Y_j\},\{\tau_j\})^m \\ &\leqslant \mathrm{Tr}_\tau \theta\left(\sum_j \tau_j\right) f(\{Y_j\},\{\tau_j\})^m \end{aligned} \tag{11}$$

and thus $G_r^* \leqslant G_r$. So calculating $\alpha_d(K)$ using $S(0) = 0$ and (4) and $\alpha_c(K)$ from the divergence of (8) will yield $\alpha_c(K) \leqslant \alpha_d(K)$. This contradicts the definitions of these capacities and shows that for finite K the RS parametrization (3) is insufficient. However, the scaling of the order parameters we shall find below suggests that the difference between the left- and right-hand side of (11) is immaterial for large K, and that (4) and (8) should to leading order yield the same result in this limit. These observations are quite compatible with what was found by the stability analysis for the tree.

To calculate the capacity, our main task is to bring the energy term (5) into a more manageable form. We only need to do this for $m \to 0$ and the only reasonable behaviour of w in this limit is $w \to 0$. Consequently the average over z_j in the expression for G_r will be dominated by the value of the maximum, and after a gauge transformation we find that for $m, w \to 0$

$$f(\{Y_j\},\{\tau_j\})^m \simeq \max_{\{z_j\}} \mathrm{e}^{-\frac{1}{2}mw^{-2}\sum_j z_j^2} \prod_j \theta\left(\tau_j Y_j + z_j + \frac{\bar{w}}{w}\tau_j \frac{1}{K}\sum_k \tau_k z_k\right). \tag{12}$$

The above can be seen as a quadratic optimization problem with linear inequality constraints. Denote by $\{z_j^*\}$ the argument of the maximum and let us say that a site j is in the interior if $\tau_j Y_j + z_j^* + (\bar{w}/w)\tau_j \frac{1}{K}\sum_k \tau_k z_k^*$ is positive. It is then easily seen that, for j and k two interior sites, $\tau_j z_j^* = \tau_k z_k^*$, and we denote this common value by s^*. The main obstacle to simplifying (12) is that we do not know which of the sites lie in the interior. Consequently, we focus on the large-K limit and, taking a hint from the analysis of the tree committee in [7], assume mw^{-2} to be of the order of K^2. Furthermore, we assume that the ratio of $\bar{w}$ and w does not diverge with increasing K. Let us call a site j embedded if $\tau_j Y_j$ is positive. With the above scaling, one sees that (12) will be negligibly small for large K unless $Y_j = \mathcal{O}(1/K)$ for any site which is not embedded and unless the number of non-embedded sites is small. Generically the Y_j are on the order of 1 and so we may assume this for the sites which are embedded. This scaling of the Y_j implies that for large K a site will lie in the interior exactly if it is embedded. Assuming s^* is known, this enables us to

calculate the value of z_j^* for the sites which do not lie in the interior as a function of s^*. Optimizing s^* yields, for large K and the above scaling of the Y_j,

$$s^* \simeq -\frac{\bar{w}}{w}\frac{1}{K}\sum_j \tau_j Y_j \theta(-\tau_j Y_j) \tag{13}$$

and

$$f(\{Y_j\},\{\tau_j\})^m \simeq \exp\left[-\frac{K^2}{2c}\sum_j Y_j^2\theta(-\tau_j Y_j)+\frac{K}{2c}\left(\frac{\bar{c}}{c}-1\right)\left(\sum_j Y_j\theta(-\tau_j Y_j)\right)^2\right]. \tag{14}$$

Here we have introduced the new parameters c and $\bar{c}$ reflecting the scaling of m, w and $\bar{w}$ via

$$\frac{m}{w^2}=\frac{K^2}{c} \qquad \text{and} \qquad \frac{m}{(w+\bar{w})^2}=\frac{K^2}{\bar{c}}. \tag{15}$$

Later we shall find that $c=\bar{c}$ holds for the stationary values and hence $\bar{w}/w=0$. So we could have arrived at the statement that the interior sites are the embedded sites, and thus at (14), by the argument that this will be true in the limit of small values of $\bar{w}/w$.

To perform the trace and the average over y_j in (5) we now rewrite the argument of the logarithm in G_r as

$$\int \mathrm{d}\lambda\,\mathrm{d}\mu\,\theta(\mu)\left\langle \mathrm{Tr}_\tau\,\delta\left(\mu-\frac{1}{\sqrt{K}}\sum_j\tau_j\right)\delta(\lambda-\sqrt{K}(\bar{u}\bar{x}+\bar{v}\bar{y}))\right.$$
$$\left.\times f\left(\left\{ux_j+vy_j+\frac{\lambda}{\sqrt{K}}\right\},\{\tau_j\}\right)^m\right\rangle_{\{y_j\}} \tag{16}$$

and introduce a Fourier representation of the δ-functions with conjugate variables $\hat{\lambda}$ and $\hat{\mu}$. Using (14) for f and linearizing the square of the sum in (14) by a Gaussian integral $(D\rho)$ factorizes the trace and the average over y_j. Performing the trace and the average allows us to rewrite (16) to leading order in K as

$$\int\frac{\mathrm{d}\lambda\,\mathrm{d}\hat{\lambda}\,\mathrm{d}\mu\,\mathrm{d}\hat{\mu}}{4\pi^2}\theta(\mu)\int D\rho\,\exp\left[-\mathrm{i}\mu\hat{\mu}-\mathrm{i}\lambda\hat{\lambda}+\mathrm{i}\hat{\lambda}\bar{u}\sqrt{K}\bar{x}+\mathrm{i}\hat{\mu}Z(\lambda)-\frac{1}{2}\hat{\mu}^2(1-A(\lambda))\right.$$
$$\left.-\hat{\mu}\hat{\lambda}\bar{v}B(\lambda)-\frac{1}{2}\hat{\lambda}^2\bar{v}^2+\sqrt{\frac{\pi}{2}}\frac{\sqrt{c}}{v}B(\lambda)\left(1-\mathrm{i}\sqrt{\frac{2}{\pi}}\mu\rho d/K+\rho^2d^2/(2K)\right)\right]. \tag{17}$$

Here $d^2=(\bar{c}/c)-1$ and

$$Z(\lambda)=\frac{1}{\sqrt{K}}\sum_j\left(1-2H\left(\frac{ux_j+\lambda/\sqrt{K}}{v}\right)\right)$$
$$A(\lambda)=\frac{1}{K}\sum_j\left(1-2H\left(\frac{ux_j+\lambda/\sqrt{K}}{v}\right)\right)^2 \qquad B(\lambda)=-\frac{2}{K}\sum_j H'\left(\frac{ux_j+\lambda/\sqrt{K}}{v}\right). \tag{18}$$

It is straightforward to perform the integral over ρ in (17). For large K one may linearize the dependence on λ and find that $Z(\lambda)\simeq Z(0)+\lambda B(0)/v$ as well as $A(\lambda)\simeq A(0)$, $B(\lambda)\simeq$

$B(0)$. This makes it possible to carry out the remaining integrations and rewrite (17) as

$$\left(1-\sqrt{\frac{\pi}{2}}\frac{B(0)}{Kv}\frac{1}{\sqrt{c}}(\bar{c}-c)\right)^{-\frac{1}{2}} \mathrm{e}^{\sqrt{\frac{\pi}{2}}\sqrt{c}B(0)/v} H\left(-\frac{Z(0)+\bar{u}\sqrt{K}\bar{x}B(0)/v}{\sqrt{1-A(0)+B(0)^2(\bar{v}^2/v^2+2\bar{v}/v)}}\right). \tag{19}$$

To arrive at the large K expansion of G_r, we now need to average the logarithm of this expression over x_j. This is readily done by applying the central limit theorem to show that in this limit $Z(0)$ and $\sqrt{K}\bar{x}$ are correlated Gaussian random variables, whereas $A(0)$ and $B(0)$ may be equated with their average over the x_j. Thus for large K

$$G_r = \sqrt{c}+\frac{1}{2K\sqrt{c}}(\bar{c}-c)+\int Dx \ln H\left(-\sqrt{\frac{(2/\pi)(p_0+\arcsin q_0)}{1-(2/\pi)(-p_2+p_0+\arcsin q_0)}}x\right). \tag{20}$$

Similarly to the tree committee, the last summand is just the energy term arising in the RS calculation of the Gardner volume for large K. By arguments analogous to [5] the extremal problem (4) may now be found to yield

$$\begin{aligned} &c = a_c(K)^{-2} \qquad \bar{c}=c \qquad p_2=-1 \\ &1-q_0 = \frac{128}{(\pi-2)^2}\alpha_c(K)^{-2} \qquad p_0=-q_0 \end{aligned} \tag{21}$$

and

$$\alpha_c(K) = \frac{16}{\pi-2}\sqrt{\ln K}. \tag{22}$$

In contrast to the claim in [3] the storage capacity is higher than for the tree committee† where the prefactor is $16/\pi$. But it does not diverge like $\ln K$ as claimed in [8].‡

It is a pleasure to acknowledge helpful discussions with M Biehl. This work was supported by grants from the Stiftung Emilia Guggenheim-Schnurr der naturforschenden Gesellschaft in Basel and the Freiwillige Akademische Gesellschaft (Basel).

References

[1] Barkai E, Hansel D and Kanter 1990 Statistical mechanics of a multilayered neural network *Phys. Rev. Lett.* **65** 2312–15

[2] Barkai E, Hansel D and Sompolinsky H 1992 Broken symmetries in multilayered perceptrons *Phys. Rev.* A **45** 4146–61

[3] Engel A, Köhler H M, Tschepke F, Vollmayr H and Zippelius A 1992 Storage capacity and learning algorithms for two layer neural networks. *Phys. Rev.* A **45** 7590–607

[4] Opper M 1994 Learning and generalization in a two-layer neural network: the role of the Vapnik-Chervonenkis dimension *Phys. Rev. Lett.* **72** 2113–16

[5] Schwarze H 1993 Learning a rule in a multilayer neural network *J. Phys. A: Math. Gen.* **26** 5781–94

[6] Monasson R and Zecchina R 1995 Weight space structure and internal representations: a direct approach to learning and generalization in multilayer neural networks *Phys. Rev. Lett.* **75** 2432–5

[7] Monasson R and Zecchina R 1996 Learning and generalization theories of large committee machines *Mod. Phys. Lett.* B **9** 1887–97

† The above shows that the anticorrelation of the hidden unit still plays a rôle for large K even if the maximal possible anticorrelation in the site symmetric parametrization is $-1/K$.

‡ In [8] the parameter q_0 (in the Gardner theory with one step of replica symmetry breaking) is assumed to be zero at the critical capacity for large K. However, in [10] it is shown that the global minimum in q_0 of (8) has positive q_0 for $\alpha > 15.4$ and large K.

[8] Kwon C, Park Y and Oh J 1995 Storage capacity of a fully connected committee machine *Neural Networks: The Statistical Mechanics Perspective* ed J Oh, C Kwon and S Cho *et al* (Singapore: World Scientific) pp 191–7

[9] Urbanczik R 1995 A fully connected committee machine learning unrealizable rules *J. Phys. A: Math. Gen.* **28** 7097–104

[10] Urbanczik R 1996 A large committee machine learning noisy rules *Neural Comput.* **8** 1267–76