Storage capacity of a fully-connected parity machine with continuous weights

1998 J. Phys. A: Math. Gen. 31 7043

(http://iopscience.iop.org/0305-4470/31/34/007)

J. Phys. A: Math. Gen. **31** (1998) 7043–7049. Printed in the UK PII: S0305-4470(98)92454-4

Storage capacity of a fully-connected parity machine with continuous weights

Yuan Sheng Xiong†§, Chulan Kwon‡‖ and Jong-Hoon Oh†

† Department of Physics, Pohang University of Science and Technology, Hyoja San 31, Pohang, Kyongbuk 790-784, Korea

‡ Department of Physics, Myong Ji University, Yongin, Namdong San 38-2, Kyonggi-Do 449-728, Korea

Received 12 March 1998

Abstract. We study a fully-connected parity machine with K hidden units for continuous weights. The geometrical structure of the weight space of this model is analysed in terms of the volumes associated with the internal representations of the training set. By examining the asymptotic behaviour of order parameters in the large K limit, we find the maximum number α_c, the storage capacity, of patterns per input unit to be $K \ln K / \ln 2$ up to leading order, which saturates the mathematical bound given by Mitchison and Durbin. Unlike the committee machine, the storage capacity per weight remains unchanged compared with the corresponding tree-like architecture.

1. Introduction

Since Gardner's pioneering work [1], statistical mechanics has proved to be a useful tool to study feed-forward neural networks. In particular, it has allowed the derivation of the storage capacity and the generalization error of neural networks inferring a rule by examples. Therefore, the investigation via statistical mechanics in various feed-forward neural networks constitutes a subject of current interest [2, 3].

For storage capacity, Barkai *et al* [4] obtained the storage capacity of a tree-like parity machine with non-overlapping receptive fields (NRF), for continuous weights. Their value is exact within the one-step replica-symmetry-breaking (RSB) scheme [5] and satisfies the mathematical bound obtained by Mitchison and Durbin [6]. Later, Barkai *et al* [7] and Engel *et al* [8] considered committee machines and found many interesting results. They found that, for a fully-connected machine, permutation symmetry, that is the invariance of output under permutation of hidden units, breaks as the number of input patterns increases. This permutation symmetry breaking was also observed for generalization problems [9] and is known to be characteristic of fully-connected architectures.

Monasson and O'Kane [10] proposed a new statistical mechanics formalism which can analyse the weight space structure related to the internal representations of hidden units. It was applied to single-layer perceptrons [11–13] as well as multi-layer networks [14–16]. Monasson and Zecchina have successfully applied this formalism to both the NRF committee and parity machines [14]. They suggested that replica symmetric (RS) solutions under this

§ Present address: Neural Computing Research Group, Department of Computer Science and Applied Mathematics, Aston University, Birmingham B4 7ET, UK.

‖ Author to whom correspondence should be addressed.

new formalism can yield reliable results equivalent to the one-step RSB solution in the conventional Gardner method. Recently, Xiong *et al* [18] and Urbanzik [17] independently obtained the same storage capacity for a fully-connected committee machine via this new approach. Using the Gardner approach, Kwon *et al* [19] also obtained a similar value which is different by only a constant factor, showing the same scaling behaviour in large K.

In the present paper, we present the analysis of weight space associated with internal representations for a parity machine with fully-connected architecture and continuous weights. Most attention will be paid to the limit of large K hidden units. The scaling of the order parameters is studied analytically and the storage capacity is obtained up to leading order.

2. Analytical formalism

We consider the fully-connected parity machine with N input units, K hidden units and one output unit, where the weights between hidden units and the output unit are set to 1. Every pair of input and hidden units is connected by weight J_{li}, where $i = 1, \ldots, N$ label the input units and $l = 1, \ldots, K$ label the hidden units. We consider the case of continuous weights with spherical constraint, $\sum_i^N J_{li}^2 = N$. Input patterns are represented by ξ_i^μ where $\mu = 1, \ldots, P$ are indices for the patterns. The output in response to an input pattern μ is given by $\prod_l \tau_l^\mu$ with $\tau_l^\mu = \mathrm{sgn}(\boldsymbol{J}_l \cdot \boldsymbol{\xi}^\mu)$ representing the internal state of hidden unit l.

Given P patterns with output codes σ^μ, the learning process in a layered neural network can be interpreted as selecting cells in the weight space corresponding to a set of suitable internal representations $\tau = \{\tau_l^\mu\}$, each of which has a non-zero elementary volume defined by

$$V_\tau = \mathrm{Tr}_{\{J_{li}\}} \prod_\mu \Theta\left(\sigma^\mu \prod_l \tau_l^\mu\right) \prod_{\mu,l} \Theta\left(\tau_l^\mu \sum_i J_{li} \xi_i^\mu\right) \tag{1}$$

where $\Theta(x)$ is the Heaviside step function. The Gardner volume V_{G}, i.e. the volume of the weight space which satisfies the given input–output relations $\{\xi_i^\mu\} \to \sigma^\mu$, can be written as the sum of the cells over all internal representations:

$$V_{\mathrm{G}} = \sum_\tau V_\tau. \tag{2}$$

As mentioned in [10, 14], the method adopted here is based on the analysis of the detailed decomposition of the Gardner volume V_{G} in elementary volumes V_τ associated with a possible internal representation. The distribution of elementary volumes can be derived from the free energy

$$g(r) = -\frac{1}{Nr} \overline{\ln\left(\sum_\tau V_\tau^r\right)} \tag{3}$$

where the overbar denotes the average over ξ_i^μ and σ^μ. The entropy $\mathcal{N}[w(r)]$ and the typical size $w(r) = -(1/N)\ln V_\tau$ of volume V_τ can be obtained through the Legendre transformations:

$$\mathcal{N}[w(r)] = -\frac{\partial g(r)}{\partial(1/r)} \qquad w(r) = \frac{\partial[rg(r)]}{\partial r}. \tag{4}$$

The entropies $\mathcal{N}_D = \mathcal{N}[w(r=1)]$ and $\mathcal{N}_R = \mathcal{N}[w(r=0)]$ are of most importance. In the thermodynamic limit, $(1/N)\overline{\ln(V_G)} = -g(r=1)$ is dominated by elementary volumes of size $w(r=1)$ with the number $\exp(N\mathcal{N}_D)$ of internal representations. Furthermore, the most numerous elementary volumes have size $w(r=0)$ and number $\exp(N\mathcal{N}_R)$ of internal representations. The vanishing condition for the entropies is related to the zero volume condition for V_G and thus gives the storage capacity.

The replicated partition function for the fully-connected parity machine reads

$$\overline{\left(\sum_\tau V_\tau^r\right)^n} = \sum_{\{\tau_l^{\mu\alpha}\}} \mathrm{Tr}_{\{J_{li}^{\alpha a}\}} \overline{\prod_{\mu\alpha} \Theta\left(\prod_l \tau_l^{\mu\alpha}\right) \prod_{\mu l\alpha a} \Theta\left(\tau_l^{\mu\alpha} \sum_i J_{li}^{\alpha a} \xi_i^\mu\right)} \tag{5}$$

with $a = 1, \ldots, r$ and $\alpha = 1, \ldots, n$. One can set σ^μ to 1 for a symmetric distribution of input patterns. Differing from the conventional Gardner approach, two kinds of replica indices for the weights are introduced. The first comes from the integer power r of internal volumes appearing in the partition function and the second from the conventional replica trick. Introducing the order parameters as

$$Q_{lk}^{\alpha\beta ab} = \frac{1}{N} \sum_i J_{li}^{\alpha a} J_{ki}^{\beta b} \tag{6}$$

as observed in our previous calculation [18], the RS ansatz implies the five sets of order parameters

$$Q_{lk}^{\alpha\beta ab} = \begin{cases} q^* & (l = k, \alpha = \beta, a \neq b) \\ q & (l = k, \alpha \neq \beta) \\ c & (l \neq k, \alpha = \beta) \\ d^* & (l \neq k, \alpha = \beta, a \neq b) \\ d & (l \neq k, \alpha \neq \beta). \end{cases} \tag{7}$$

q^* and q are known to be the typical overlaps between two weight vectors corresponding to the same and to different internal representations τ, respectively [14]. For the fully-connected architecture, additional order parameters c, d^* and d have been introduced, which describe the correlations between weights that leave the same input unit and arrive at different hidden units. c and d^* characterize those correlations within one internal representation, whereas d correlates different internal representations.

By using the standard saddle-point method in the $N \to \infty$ limit, we find

$$\begin{aligned} g(r) = -\frac{1}{2}\Bigg[& \frac{(K-1)(q-d)}{1 - q^* + r(q^* - q) - [c - d^* + r(d^* - d)]} \\ & + \frac{K-1}{r} \ln\{1 - q^* + r(q^* - q) - [c - d^* + r(d^* - d)]\} \\ & + \frac{q + (K-1)d}{1 - q^* + r(q^* - q) + (K-1)[c - d^* + r(d^* - d)]} \\ & + \frac{1}{r} \ln\{1 - q^* + r(q^* - q) + (K-1)[c - d^* + r(d^* - d)]\} \\ & + (K-1)\left(1 - \frac{1}{r}\right) \ln(1 - q^* - c + d^*) \\ & + \left(1 - \frac{1}{r}\right) \ln[1 - q^* + (K-1)(c - d^*)]\Bigg] - \frac{\alpha}{r} \int Dx \int \prod_l Dy_l \\ & \times \ln\left[\mathrm{Tr}_{\tau_l} \Theta\left(\prod_l \tau_l\right) \int Dz \int \prod_l Du_l \left(\int Dv \prod_l H(\Omega_l)\right)^r\right] \end{aligned} \tag{8}$$

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

$$\Omega_l = \frac{\sqrt{q^* - q - d^* + d}\, u_l + (\sqrt{c - d^*}\, v + \sqrt{q - d}\, y_l + \sqrt{d^* - d}\, z + \sqrt{d}\, x)\tau_l}{\sqrt{1 - q^* + d^* - c}}. \tag{9}$$

Here α denotes P/N, the number of patterns per input unit.

The order parameters can now be found from the stationary condition of $g(r)$. Then one can estimate the values $\alpha = \alpha_D$ for $\mathcal{N}_D = 0$ and $\alpha = \alpha_R$ for $\mathcal{N}_R = 0$. α_D determines when the volumes of the most dominant size comprising the Gardner volume V_{G} vanish and α_R determines when the most numerous volumes vanish. In general, the storage capacity α_{c} should satisfy the inequality $\alpha_D \leqslant \alpha_R \leqslant \alpha_{\mathrm{c}}$. For the NRF parity machine, it was found asymptotically in the large K limit that $\alpha_D = \alpha_R = \alpha_{\mathrm{c}} = \ln K/\ln 2$ [14], agreeing with the result found from the Gardner approach [7]. For committee machines, the values of α_D were found for both the NRF [14] and ORF [17, 18] cases, also showing good agreement, other than a constant factor $\sqrt{2}$, with the result from the Gardner approach [19]. This discrepancy observed in the committee machines might come from the different levels of RSB scheme used in the two approaches: an RS calculation in the new approach based on internal representations and an 1RSB in the conventional Gardner approach. In the following we will consider the ORF parity machine and find the asymptotic values of α_D and α_R.

3. Storage capacity in large K limit

We investigate the asymptotic expression for the storage capacity in the limit of large K hidden units. There is a particular phenomenon, permutation symmetry breaking (PSB), which is only observed in fully-connected machines. Permutation symmetry (PS) is invariance of the output under the permutation of the hidden units. From the usual view of a valley landscape, which is useful in spin-glass-like systems with broken ergodicity, the Gardner volume for a smaller number of patterns is obtained within a single valley. Given a weight vector inside this single valley, the other weight vectors transformed by permuting hidden units then also lie in the same valley. PS is preserved in this case. As a result, there is no preferred alignment of hidden units so that on-site overlap between different sites are not distinguishable and found to be at most $\mathcal{O}(K^{-1})$ vanishing in the leading order. For an increased number of patterns, the weight space constituting the Gardner volume is decomposed into many valleys. Weight vectors with different alignment of hidden units then belong to different valleys, so q^*, that is the largest on-site overlap, has a non-zero value. This stage is called the PSB phase. In this phase, as the number of patterns increase, $q^* \to 1$ and the machine reaches its maximal storing capability. The shift from the PS to PSB phase is expected to be driven by a phase transition, as observed for the committee machine [17, 19]. We concentrate here only on the PSB phase in order to find the storage capacity.

For fully-connected machines, there appear order parameters between different hidden units: c, d and d^*. One can observe from the expression of the free energy that they contribute in rescaled forms: $(K-1)c$, $(K-1)d$ and $(K-1)d^*$. For the committee machine these rescaled order parameters were found to be of $\mathcal{O}(1)$ to leading order. Corrections were also found [18], but they were not necessary in order to find only the dominant term of the storage capacity. This is also true in this study of the parity machine.

First, we concentrate on the $r \to 1$ limit, which corresponds to the internal representations giving the dominant contribution to the Gardner volume V_G. In this limit,

$$g(r) \simeq g(1) + (r-1)\mathcal{N}_D. \tag{10}$$

The free energy $g(1)$ is independent of q^* and d^*, and is given by

$$\begin{aligned} g(1) = -\frac{1}{2}\bigg[&\frac{(K-1)(q-d)}{1-q-(c-d)} + (K-1)\ln[1-q-(c-d)] \\ &+ \frac{q+(K-1)d}{1-q+(K-1)(c-d)} + \ln[1-q+(K-1)(c-d)]\bigg] \\ &-\alpha\int Dx\int\prod_l Dy_l \ln\left[\mathrm{Tr}_{\tau_l}\Theta\left(\prod_l \tau_l\right)\int Dz\prod_l H(Q_l)\right] \end{aligned} \tag{11}$$

where

$$Q_l = \frac{\sqrt{c-d}\,z + \sqrt{q-d}\,y_l + \sqrt{d}\,x}{\sqrt{1-q+d-c}}\tau_l. \tag{12}$$

The explicit form of $\mathcal{N}_D$ is

$$\begin{aligned} \mathcal{N}_D = \frac{1}{2}\bigg[&\frac{(K-1)(q-d)[q^*-q-(d^*-d)]}{(1-q-c+d)^2} \\ &+(K-1)\ln[1-q-(c-d)] - \frac{(K-1)[q^*-q-(d^*-d)]}{1-q-(c-d)} \\ &+\frac{[q+(K-1)d][q^*-q+(K-1)(d^*-d)]}{[1-q+(K-1)(c-d)]^2} \\ &-(K-1)\ln(1-q^*-c+d^*) - \ln[1-q^*+(K-1)(c-d^*)] \\ &+\ln[1-q+(K-1)(c-d)] - \frac{q^*-q+(K-1)(d^*-d)}{1-q+(K-1)(c-d)}\bigg] \\ &-\alpha\int Dx\int\prod_l Dy_l \\ &\times\frac{\mathrm{Tr}_{\tau_l}\Theta(\prod_l \tau_l)\int Dz\int\prod_l Du_l\int Dv\prod_l H(\Omega_l)\ln[\int Dv\prod_l H(\Omega_l)]}{\mathrm{Tr}_{\tau_l}\Theta(\prod_l\tau_l)\int Dz\prod_l H(Q_l)} \\ &+\alpha\int Dx\int\prod_l Dy_l\ln\left[\mathrm{Tr}_{\tau_l}\Theta\left(\prod_l\tau_l\right)\int Dz\prod_l H(Q_l)\right]. \end{aligned} \tag{13}$$

Notice that $g(1)$ has the same expression as the RS free energy found from the Gardner approach, which is generally true for both the parity and committee machines, irrespective of NRF and ORF architectures. $g(1)$ is a function of q, $(K-1)c$ and $(K-1)d$ only, which can be found from the saddle-point condition for $g(1)$. For the committee machine, it was shown that $q \to 1$, $(K-1)c \simeq -1$ and $(K-1)d \simeq -1$. q as well as q^* showed PSB. However, we observe a quite different behaviour in these order parameters for the parity machine. We find that q preserves PS; $q = 0$. Maybe the most different property is that $(K-1)c = 0$ and $(K-1)d = 0$ to leading order. Corrections are not necessary for the dominant term of the storage capacity. This simple structure makes further calculations easy. q^* and $(K-1)d^*$ appear only in $g'(1)$, i.e. $\mathcal{N}_D$. We find

$$1-q^* \simeq \frac{\pi^2\Gamma^2}{2\alpha^2} \qquad (K-1)d^* \sim 1-q* \tag{14}$$

with $\Gamma^{-1} = -\sqrt{\pi} \int \mathrm{d}u H(u) \ln H(u) \simeq 0.62$. This shows that PS is broken inside the same internal representation. The detailed expression for $(K-1)d^*$ is not needed. Then the asymptotic expression for the entropy $\mathcal{N}_D$ is

$$\mathcal{N}_D \simeq \frac{K}{2} + K \ln \alpha - \alpha \ln 2 \tag{15}$$

which vanishes at $\alpha = \alpha_D \simeq K \ln K / \ln 2$.

Second, we consider the limit $r \to 0$. In this limit q^* behaves as $1 - q^* = r/\mu$, where $\mu \to \infty$. As in the $r \to 1$ limit, one finds that $(K-1)c = 0$, $(K-1)d = 0$ and $(K-1)d^* \sim 1 - q^*$ to leading order. Then the free energy is given by

$$\begin{aligned} rg(r) \simeq & -\frac{K}{2}[\ln(1 - q^* + rq^*) - \ln(1 - q^*)] \\ & -\alpha \ln \left[\mathrm{Tr}_{\tau_l} \Theta\left(\prod_l \tau_l \right) \prod_l \int Du_l H^r \left(\sqrt{\frac{q^*}{1 - q^*}} u_l \right) \right] \\ \simeq & -\frac{K}{2} \ln(1 + \mu) + \alpha \ln 2 - K\alpha \ln \left[1 + \frac{1}{\sqrt{1 + \mu}} \right]. \end{aligned} \tag{16}$$

The saddle-point equation with respect to μ gives $\sqrt{\mu} \simeq \alpha$. Finally, we obtain the typical logarithm $\mathcal{N}_R$ of the total number of internal representations,

$$\mathcal{N}_R \simeq K + K \ln \alpha - \alpha \ln 2 \tag{17}$$

which vanishes at $\alpha = \alpha_R \simeq K \ln K / \ln 2$. We have found that $\alpha_D = \alpha_R$ to leading order. Therefore, we expect that the dominant term of the asymptotic value of the storage capacity is given by

$$\alpha_D = \alpha_R = \alpha_c = \frac{K \ln K}{\ln 2}. \tag{18}$$

The storage capacity per weight of the fully-connected parity machine is $\ln K / \ln 2$, which satisfies the mathematical bound $\sim \ln K$ by Mitchison and Durbin [6].

It is interesting to compare this result with the storage capacity of the NRF parity machine [4, 14]. The storage capacity per weight remains unchanged for the fully-connected case. This phenomena is different from the case of the committee machines, where the storage capacity per weight for the ORF case is larger than that for the NRF case [14, 17–19].

We argue that this feature comes from the unique structure of the fully-connected parity machine, which is characterized by global symmetry [4] as well as permutation symmetry. For the fully-connected committee machine near saturated input patterns, PS is broken both inside the same internal representation and between different internal representations, i.e. $q^* \neq d^*$ and $q \neq d$. However, in the fully-connected parity machine, PS is broken only inside the same internal representation, whereas PS between different internal representations is preserved due to the global symmetry of the system, i.e. $q^* \neq d^*$ and $q = d$. As a result, the storage capacity per weight remains unchanged for the fully-connected parity machine.

4. Summary

In this paper, we have presented a theoretical investigation of the storage capacity of the fully-connected parity machine with continuous weights. The geometrical structure of the weight space has been analysed using the new method proposed by Monasson and O'Kane [10] and developed by Monasson and Zecchina [14, 15]. By examining the asymptotic behaviour of order parameters in the large K limit, we find a reliable estimate of the storage capacity α_c to be $K \ln K/\ln 2$ without using the RSB scheme.

In the Gardner approach for the parity and committee machines, the RS calculation leads to a wrong result. However, in this approach based on the analysis of internal representations, the RS calculation seems to give a reliable result. A small discrepancy of a factor $\sqrt{2}$ between the two approaches was observed for the committee machine [19]. For the parity machine, there is no such discrepancy found in the NRF case [4, 15]. One can expect a similar coincidence between the two approaches in the ORF case, as has been shown in this study. As mentioned in [14, 15], in the NRF case the instability of the RS solution for finite K decreases with increasing K. Then a similar stability analysis in the ORF case deserves further research. It is also interesting to consider this problem via the Gardner approach.

Acknowledgments

We appreciate Daniel D Lee for critical reading of the manuscript. This work was supported in part by KOSEF under grant No 971-0202-010-2 and in part by the Korea Ministry of Education through the POSTECH Basic Science Research Institute (under grant BSRI-97-2438). CK acknowledges the support from KOSEF through the Center for Theoretical Physics, Seoul National University.

References

[1] Gardner E 1987 *Europhys. Lett.* **4** 481
Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257
Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271
[2] Watkin T, Rau A and Biehl M 1993 *Rev. Mod. Phys.* **65** 499
[3] Opper M and Kinzel W 1996 Statistical mechanics of generalization *Physics of Neural Networks* ed J L van Hemmen, E Domany and K Shulten (Berlin: Springer)
[4] Barkai E, Hansel D and Kanter I 1990 *Phys. Rev. Lett.* **65** N18, 2312
[5] Mézard M, Parisi G and Virasoro M A 1987 *Spin Glass Theory and Beyond* (Singapore: World Scientific)
[6] Mitchison G J and Durbin R M 1989 *Biol. Cybern.* **60** 345
[7] Barkai E, Hansel D and Sompolinsky H 1992 *Phys. Rev.* E **45** 4146
[8] Engel A, Köhler H M, Tschepke F, Vollmayr H and Zippelius A 1992 *Phys. Rev.* E **45** 7590
[9] Kang K, Oh J-H, Kwon C and Park Y 1993 *Phys. Rev.* E **48** 4805
Kang K, Oh J-H, Kwon C and Park Y 1996 *Phys. Rev.* E **54** 1811
[10] Monasson R and O'Kane D 1994 *Europhys. Lett.* **27** 85
[11] Derrida B, Griffiths R B and Prugel-Bennett A 1991 *J. Phys. A: Math. Gen.* **24** 4907
[12] Biehl M and Opper M 1995 *Neural Networks: The Statistical Mechanics Perspective* ed Jong-Hoon Oh, Chulan Kwon and Sungzoon Cho (Singapore: World Scientific)
[13] Engel A and Weigt M 1996 *Phys. Rev.* E **53** R2064
[14] Monasson R and Zecchina R 1995 *Phys. Rev. Lett.* **75** 2432
Monasson R and Zecchina R 1996 *Phys. Rev. Lett.* **76** 2205
[15] Monasson R and Zecchina R 1996 *Mod. Phys. Lett.* B **9** 1887
[16] Cocco S, Monasson R and Zecchina R 1996 *Phys. Rev.* E **54** 717
[17] Urbanzik R 1997 *J. Phys. A: Math. Gen.* **30** L337
[18] Xiong Y S, Oh J-H and Kwon C 1997 *Phys. Rev.* E **56** 4540
[19] Kwon C and Oh J-H 1997 *J. Phys. A: Math. Gen.* **30** 6273