ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 157-162

Exact asymptotic estimates of the storage capacities of the committee machines with overlapping and non-overlapping receptive fields

Chulan Kwon

Department of Physics, Myong Ji University, Yongin, Kyonggi-Do 449-728, Korea

Jong-Hoon Oh

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

Abstract.
We present theoretical investigations via the replica theory of the storage capacities of committee machines with the large M number of hidden units and spherical weights. We discuss the physical implication of the breaking of permutation symmetry and replica symmetry. We obtain the asymptotic estimates α_c of the storage capacities per input unit within the one-step replica symmetry breaking scheme. In the case of the overlapping receptive fields, we find $\alpha_c \simeq (8\sqrt{2}/(\pi-2))M\sqrt{\ln M}$. Through a simple reduction, in the case of the non-overlapping receptive fields, we find $\alpha_c \simeq (8\sqrt{2}/\pi)\sqrt{\ln M}$. Both values satisfy the mathematical bound of Mitchison and Durbin.

1. Introduction

Storage capacity is defined as the maximum number of patterns that can be stored in a network and is given by the value α_c per input unit. Statistical mechanics has proven to be successful since Gardner's pioneering work, especially for single layer perceptrons [1].

There have been further studies on the storage capacities of more complicated networks. Barkai, Hansel, and Kanter obtained $\alpha_c = \ln M/\ln 2$ for a parity machine with non-overlapping receptive fields (NRF) and spherical weights [2]. Their value is exact within the one-step replica symmetry breaking (1RSB) scheme and satisfies the mathematical bound obtained by Mitchison and Durbin [3]. Later Barkai et al. [4] and Engel et al. [5] made extensive progress for committee machines and found many interesting results. The importance of permutation symmetry breaking (PSB) was first pointed out for

ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 157-162

the committee machine with overlapping receptive fields (ORF). In the limit of large M they only obtained the replica symmetric (RS) result for the NRF case that violated the Mitchison-Durbin bound.

In this paper we calculate the storage capacities of the committee machines with ORF and NRF. Most of expressions and discussions from now on are with regards to the ORF machine. We simply shift to the NRF case at the end of calculation. Our calculations are carried out within the 1RSB scheme.

2. Formalism

Let us consider a double layer network with N input units, M hidden units, and one output unit. For the network to be a committee machine, weights in the second layer attached to the output unit are set to 1. Let $\{W_{ji}\}$ be the set of weights between input i and hidden unit j and let $\{\xi_i^\mu\}$ for $\mu = 1, \ldots, P$ be input values at input units i randomly distributed with variance 1. Output value is given by $o^\mu = \text{sgn}\left(M^{-1/2} \sum_j \text{sgn}(h_j)\right)$, where the receptive fields h_j are given by $N'^{-1/2} \sum_i W_{ji} \xi_i^\mu$. In this expression the summation over i depends on the architecture of the hidden layer. For a fully connected machine, all the i are swept and $N' = N$. The h_j are called overlapping in this case. For a tree-structure machine, only M i's are swept for a given hidden unit j and $N' = M$. In this machine, since the i do not overlap for different j's, the h_j s are called non-overlapping. A pattern is classified by an input-output relation, $\{\xi_i^\mu\} \to \sigma^\mu \in \{-1, 1\}$. A pattern can be stored in the machine if $o^\mu = \sigma^\mu$. The storage capacity is then defined as the maximum number P_c of the patterns that can be stored reliably in the machine.

The volume V in weight space satisfying given P input-output relations, originally considered by Gardner, can be found from the partition function $Z = \text{Tr}_{\{W_{ji}\}} \exp\left(-\beta \sum_{\mu=1}^P \Theta\left(-\sigma^\mu o^\mu\right)\right)$. $\Theta(x)$ is the Heaviside step function and the trace is taken over continuous weights with a spherical constraint, $\sum_j^M W_{ji}^2 = M$ and $\sum_i^{N'} W_{ji}^2 = N'$. As the inverse temperature $\beta \equiv 1/T$ goes to ∞, Z leads to the Gardner volume V. Then via the replica theory the average of $\ln Z$ over random input and output values is given by $\langle\langle \ln Z \rangle\rangle = n^{-1} \ln \langle\langle Z^n \rangle\rangle$ in the $n \to 0$ limit where the double bracket denotes the average over the ξ_i^μ. In the random average σ^μ can be set to 1 without loss of generality.

Order parameters are given by $Q_{jj'}^{\sigma\rho} = N^{-1} \sum_i \langle\langle \langle W_{ji}^\sigma W_{j'i}^\rho \rangle_T \rangle\rangle$ where $\langle \cdots \rangle_T$ denotes the thermal average and σ, ρ are replica indices. There are three independent order parameters: Q^σ for $\sigma = \rho$ and $j \neq j'$, $p^{\sigma\rho}$ for $\sigma \neq \rho$ and $j \neq j'$, $q^{\sigma\rho}$ for $\sigma \neq \rho$ and $j = j'$.

Overlaps between different hidden units, Q^σ and $p^{\sigma\rho}$, are found to be $\mathcal{O}(1/M)$, but make important contributions through rescaling: $\bar{Q}^\sigma = (M-1)Q^\sigma$, $\bar{p}^{\sigma\rho} = (M-1)p^{\sigma\rho}$. We apply the 1RSB scheme for the order parameter matrices. q_1 and $\bar{p}_1$ are matrix elements in diagonal blocks with size m of $q^{\sigma\rho}$ and $\bar{p}^{\sigma\rho}$, respectively. q_0 and $\bar{p}_0$ are matrix elements off the diagonal blocks. The order parameters can be determined by solving the saddle point equations

ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 157-162

obtained from the stationary condition of the free energy F. Then we find α_c using the fact that q_1 goes to 1 as α goes to α_c.

3. Symmetry breaking

The output, therefore the energy, of the ORF machine is invariant under the permutation of hidden units. This property is called permutation symmetry (PS). In the PS phase the specialization in hidden units does not occur so that the overlap between different hidden units and the self-overlap on a hidden unit are not distinguishable. Therefore $q^{\sigma\rho}$, $p^{\sigma\rho} \sim \mathcal{O}(1/M)$. As α increases, the phase transition driven by permutation symmetry breaking (PSB) takes place. In the PSB phase the specialization in hidden units results in the increase of the self-overlap, i.e., $q^{\sigma\rho} > 0$. The Gardner volume shrinks as α increases. In the PSB phase the Gardner volume is decomposed into many islands. Islands correspond to *pure states* or valleys in the weight space with infinite energy barriers. Islands are transformed to each other by permuting hidden units. Replica symmetry breaking is another source of segmentation of the Gardner volume. If a further segmentation within each of islands separated by PSB is possible, RSB is diagnostic to this phenomenon.

Interesting physical properties related to the breaking of these symmetries are found for $\mathcal{O}(\alpha) \geq M$. Let α' be α/M. Investigating the saddle point equations for α up to $\mathcal{O}(M)$, we first find a solution where $q_1 = q_0 = q$, $p_1 = p_0 = p$, and in the leading order $q = p = 0$. This is the solution in the PS phase. Both RS and PS are preserved. There occurs a phase transition at $\alpha' = \alpha'_1$. For α' above this value we find $q_1 > 0$ and $q_0 = 0$. This solution indicates that PS and RS are broken simultaneously. We can imagine the picture in the weight space that there are many islands separated by PSB and each island is also composed of smaller islands separated by RSB.

There is another phase transition at $\alpha' = \alpha'_2$. For α' above this value we have another phase with PSB and RSB, where $q_1 > q_0 > 0$. Both q_1 and q_0 go to 1 for large α'. One might think the solution recovers RS. For large α' we find $1 - q_1 \sim m/\alpha'^8$, $1 - q_0 \sim 1/\alpha'^2$, $-\ln m \sim \alpha'^2$. So $1 - q_0 \gg 1 - q_1$, showing the difference from the RS solution. Recently Urbanczik found $\alpha'_1 \simeq 4.91$ and $\alpha'_2 \simeq 15.4$ [6]. However, the calculation by treating α' to be large but finite gives only information: $\alpha'_c \equiv \alpha_c/M \to \infty$. We need to find the dependence of α'_c on M in the limit of large M.

4. Storage capacities

As α goes to α_c, $1 - q_1$ and $1 - q_0$ become very small. Other small quantities include $1 + \bar{Q}$, $q_1 + \bar{p}_1$, and $q_0 + \bar{p}_0$. We can show these three quantities are zero in the leading order in the limit of large M [7]. It is possible to estimate α_c by investigating the asymptotic behavior of those small quantities. It is fortunate that we can deal with the same number of small quantities as that of the order

ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 157-162

parameters, which makes asymptotic calculation tractable.

For convenience we use the following notations: $w = 1 + \bar{Q} - q_1 - \bar{p}_1$, $v_1 = q_1 + \bar{p}_1$, and $v_0 = q_0 + \bar{p}_0$, where $w' = w/M$, $v_1' = v_1/M$, and $v_0' = v_0/M$. We assume the following scaling: $1 - q_1 - w' = m/c$ and $w = m/d$ with $m \to 0$, $1/c \to 0$, $1/d \to 0$. A similar scaling for q_1, $1 - q_1 = m/c$, was used in the previous study on the storage capacity of the NRF parity machine [2]. We can reduce to the NRF case by simply setting $v_1 = v_0 = 0$, ignoring the integrations over z_1 and z_0, and replacing α' by α. We also assume $1 - q_1 = m/c$ in the NRF case.

We also use the following expressions to write equations in simpler forms: $Q_0^* = 1 - q_0 - (v_1' - v_0')$, $Q_0' = (q_0 - v_0')Q_0^{*-1}$, $Q_0 = (q_0 - v_0)Q_0^{*-1}$, and $W = \left((1 - q_0 - (v_1 - v_0))Q_0^{*-1}\right)^{1/2}$. We assume that $Q_0^* \to 0$, $Q_0' \to \infty$, $Q_0 \to \infty$, and $W \to 1$, which can be shown selfconsistently from the final result.

We write $-m\beta F/NM = G_0 + \alpha' G_r$. G_0 can be found exactly in the case of spherical weights, given by

$$2G_0 = \ln\left(1 + c\,Q_0^*\right) + \frac{c(q_0 - v_0')}{1 + c\,Q_0^*} + \frac{1}{M}\ln\left(1 + d(v_1 - v_0)\right) + \frac{d v_0}{M(1 + d(v_1 - v_0))} \,. \tag{1}$$

It is much complicated to compute G_r. It is written by:

$$G_r = \int Dz_0 \int \prod_j Dt_{0j} \,\ln\left(\sum_{\{\eta_j = \pm 1\}} \int Dz_1 \prod_j \int_0^\infty Dt_{1j}\, A^m \right) \,, \tag{2}$$

$$A = \sum_{\{\tau_j = \pm 1\}} \Theta\left(\sum_j \tau_j\right) \int Du \prod_j H\left[\left[\frac{c\,Q_0^*}{m}\right]^{1/2} \tau_j \eta_j \bar{t}_{1j} + \frac{iu\tau_j}{\sqrt{M}}\left(1 - \frac{c}{Md}\right)^{1/2} \right] \,,$$

where $\bar{t}_{1j} = t_{1j} + \frac{iWz_1\eta_j}{\sqrt{M}} + \sqrt{Q_0'}\eta_j t_{0j} + \frac{i\sqrt{Q_0}\eta_j z_0}{\sqrt{M}}$. In the limit of large c we can expand G_r as: $G_r = f^{(0)} + f^{(1)}$, where $f^{(0)}$ and $f^{(1)}$ are of the zeroth and the first order in $1/\sqrt{c}$, respectively. $f^{(0)}$ is obtained from a partial sum in Eq. (2) over the η_j and the τ_j where $\sum_j \eta_j < 0$ and $\eta_j = -\tau_j$ for all j. Then $A^m \to 1$ as $m \to 0$ and $c \to \infty$. $f^{(1)}$ is obtained from another partial sum where each term is given by the condition that $\eta_j = \tau_j = 1$ for one j and $\sum_{j'(\neq j)} \eta_{j'} = 0$, $\eta_{j'} = -\tau_{j'}$ for $j' \neq j$. In this case $A^m \to \exp\left(-\tilde{c}\,\bar{t}_{1j}^2/2\right)$, where $\tilde{c} = c\,Q_0^*(1 + c/Md)^{-1}$. $\tilde{c} \to \infty$ is assumed, which can also be shown to be true selfconsistently.

Technical steps for the computation of $f^{(0)}$ and $f^{(1)}$ are given in detail elsewhere [7]. We find

$$f^{(0)} = -\frac{1}{4}\,\frac{\frac{2}{\pi}\sin^{-1}\left(\frac{Q_0'}{1+Q_0'}\right) - \frac{2}{\pi}\frac{Q_0}{1+Q_0'}}{1 - \frac{2}{\pi}\sin^{-1}\left(\frac{Q_0'}{1+Q_0'}\right) - \frac{2}{\pi}\frac{W^2}{1+Q_0'}} \tag{3}$$

ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 157-162

and

$$f^{(1)} = \sqrt{\frac{M(1-\frac{2}{\pi})}{2\pi}}\sqrt{\frac{1+\frac{c}{Md}}{c}}\frac{1}{1-\frac{2}{\pi}\sin^{-1}\left(\frac{Q_0'}{1+Q_0'}\right)} \,. \quad (4)$$

Now the free energy is given, from Eqs. (1), (3), and (4), by: $-m\beta\frac{F}{NM} = G_0 + \alpha'(f^{(0)} + f^{(1)})$. The five saddle point equations are obtained from the stationary condition of F with respect to c, d, Q_0^*, $v_1 - v_0$, and v_0. Solving the saddle point equations, we find the following result: $c \simeq ((\pi-2)^3/2048)M\alpha'^4$, $d \simeq \sqrt{2/\pi}\,c$, $v_1 - v_0 \simeq (32/(\pi-2))(M\alpha'^2)^{-1}$, $v_0 \simeq (\pi-2)(M\alpha')^{-1}$, $Q_0^* \simeq (128/(\pi-2)^2)\alpha'^{-2}$. An additional equation comes from the condition that $\partial F/\partial m = 0$ as $m \to 0$, given by:

$$0 \simeq \ln(1 + c\,Q_0^*) + \frac{1}{Q_0^*} - \alpha'\frac{\pi-2}{4\sqrt{2}}\frac{1}{\sqrt{Q_0^*}} \,. \quad (5)$$

This gives the asymptotic value α_c of the storage capacity per input unit

$$\alpha_c \simeq \frac{8\sqrt{2}}{\pi-2}M\sqrt{\ln M} \,. \quad (6)$$

The reduction to the NRF case can be made easily by replacing $\pi - 2$ by π and α' by α in the above result. Therefore we find $c \simeq (\pi^3/2048)M\alpha^4$, $1 - q_0 \simeq (128/\pi^2)\alpha^{-2}$ and

$$\alpha_c \simeq \frac{8\sqrt{2}}{\pi}\sqrt{\ln M} \,. \quad (7)$$

We observe the storage capacities per weight are smaller than the mathematical bound $\sim \ln M$ obtained by Mithison and Durbin.

5. Discussions

In a usual spin-glass phase, q_0 becomes smaller as q_1 gets larger. It implies that the distance between islands becomes larger as the volume of each island gets smaller. In this study both q_1 and q_0 go to 1 although q_1 approaches to 1 much faster. We might explain this rather unusual phenomenon by supposing the following landscape picture in the weight space. Let us imagine a group of islands in the weight space. As each island shrinks, the distance between islands also decreases. This is possible when the overall boundary surrounding islands gets contracting.

Recently Monasson and Zecchina presented an interesting paper [8]. They used a formalism different from the conventional Gardner approach. Interestingly, they argued that the RSB calculation might be avoided, i.e., the RS calculation is acceptable. They reproduced the known result for the NRF parity machine and obtained a new result for the NRF committee machine, $\alpha_c \simeq (16/\pi)\sqrt{\ln M}$. This is larger by the factor $\sqrt{2}$ than the value in Eq. (7),

ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 157-162

obtained in this paper. We have applied their approach to the ORF committee machine and the result shows the same difference of the factor $\sqrt{2}$ [9]. As discussed before, RSB is diagnostic to the existence of many separated islands in the weight space. The phase transition signals such segmentation of the weight space. In their approach based on the RS calculation, however, no phase transition is likely to occur, about which we cannot give a clear explanation in the present stage. One can criticize that the 1RSB calculation in our approach might not be exact. Difference might be reduced if a higher-step RSB scheme, which seems to be a very difficult task, is applied in one or both of the two approaches.

Acknowledgments

This work is supported in part by the non-directed fund, Korea Research Foundation, 1995. It is also supported in part by the KOSEF through the Center for the Theoretical Physics of the Seoul National University and in part by the Ministry of Education through the Basic Science Research Institute of POSTECH.

References

[1] E. Gardner, Europhys. Lett. **4**, 481 (1987); J. Phys. A **21**,257 (1988); E. Gardner and B. Derrida, *ibid.* **21**, 271 (1988); W. Krauth and M. Mézard, J. Phys. France **50**, 3057 (1989).

[2] E. Barkai, D. Hansel, and I. Kanter, Phys. Rev. Lett. **65**, 2312 (1990).

[3] G. J. Mitchison and R. M. Durbin, Boil. Cybern.**60**, 345 (1989).

[4] E. Barkai, D. Hansel, and H. Sompolinsky, Phys. Rev. A **45**, 4146 (1992).

[5] A. Engel, H. M. Köhler, F. Tschepke, H. Vollmayr, and A. Zippelius, Phys. Rev. A **45**, 7590 (1992).

[6] R. Urbanczik, J. Phys. A **28**, 7907 (1995).

[7] C. Kwon and J.-H. Oh, submitted to Phys. Rev. E (1997)

[8] R. Monasson and R. Zecchina, Phys. Rev. Lett., **75**, 2432 (1995).

[9] Y. S. Xiong, J.-H. Oh, and C. Kwon, submitted to Phy. Rev. E (1997).