Correlation of internal representations in feed-forward neural networks

1996 J. Phys. A: Math. Gen. 29 L323

(http://iopscience.iop.org/0305-4470/29/13/003)

J. Phys. A: Math. Gen. **29** (1996) L323–L327. Printed in the UK

LETTER TO THE EDITOR

Correlation of internal representations in feed-forward neural networks

A Engel†

Institut für Theoretische Physik, Otto-von-Guericke Universität, Postfach 4120, D-39016 Magdeburg, Germany

Received 22 April 1996

Abstract. Feed-forward multilayer neural networks implementing random input–output mappings develop characteristic correlations between the activity of their hidden nodes which are important for the understanding of the storage and generalization performance of the network. It is shown how these correlations can be calculated from the joint probability distribution of the aligning fields at the hidden units for arbitrary decoder function between hidden layer and output. Explicit results are given for the parity-, and-, and committee-machines with arbitrary numbers of hidden nodes near saturation.

Multilayer neural networks (MLN) are powerful information processing devices. Because of their computational abilities they are the workhorses in practical applications of neural networks and a lot of effort is devoted to a thorough understanding of their functional principles. At the same time, their theoretical analysis within the framework of statistical mechanics is much harder than that for the single-layer perceptron. It was realized from the beginning that the properties of the internal representations defined as the activity patterns of the hidden units resulting from certain inputs are crucial for the understanding of the storage and generalization abilities of MLN [1–5]. Qualitatively, the flexibility of MLN stems from the fact that the different subperceptrons between input and hidden layer can share the effort to produce the correct output. This *division of labour* gives rise to particular correlations between the activity of the hidden nodes. Near saturation these correlations become a characteristic feature of the decoder function between hidden units and output of the MLN under consideration and determine different aspects of its performance.

Several *ad hoc* approximations have been used to calculate these correlations, e.g., it was assumed that all internal representations giving the correct output (so-called legal internal representations, LIR) are equiprobable [2, 3] or that only internal representations at the decision boundary of the decoder function occur [3]. In this letter we show how these correlations between the hidden units can be calculated for a MLN of tree-architecture and give explicit results for the parity- (PAR), and- (AND) and committee- (COM) machine with arbitrary number K of hidden nodes near saturation.

A MLN of tree-architecture is given by N input nodes ξ_{ik} grouped into K sets of N/K nodes each, K hidden nodes τ_k and one output σ. The inputs $\boldsymbol{\xi}_k = \{\xi_{ik}, i = 1, \ldots, N/K\}$ are coupled to the kth hidden unit by spherical couplings $\boldsymbol{J}_k = \{J_{ik} \in \mathbb{R}, i = 1, \ldots, N/K,\ \boldsymbol{J}_k^2 =$

† E-mail: andreas.engel@physik.uni-magdeburg.de

0305-4470/96/130323+05$19.50

$N/K\}$ according to $\tau_k = \mathrm{sgn}(\boldsymbol{J}_k\boldsymbol{\xi}_k)$. Each hidden node has therefore its own set of inputs (non-overlapping receptive fields). The hidden units τ_k determine the output through a fixed Boolean function $F(\{\tau_k\})$. A set of input–output mappings $\{\boldsymbol{\xi}_k^\mu, \sigma^\mu\}, \mu = 1, \ldots, p$ is generated at random where each bit is ± 1 with equal probability. The couplings $\boldsymbol{J}_k$ are then adjusted in such a way that the MLN gives the desired output σ^μ for each input $\boldsymbol{\xi}_k^\mu$. This is generically possible only if $p/N = \alpha < \alpha_c$.

We are interested in the correlations

$$c_n = \frac{1}{\alpha N}\sum_{\mu=1}^{\alpha N} \tau_1^\mu \tau_2^\mu \tau_3^\mu \cdots \tau_n^\mu \tag{1}$$

near saturation, i.e. for $\alpha \to \alpha_\mathrm{c}$. From the statistical properties of the inputs it follows that for permutation invariant Boolean functions $F(\{\tau_k\})$

$$c_n = \langle\langle \tau_{k_1}\tau_{k_2}\cdots\tau_{k_n}\rangle\rangle \tag{2}$$

where $\langle\langle\cdots\rangle\rangle$ denotes the average over the input–output pairs and $k_1, \ldots, k_n$ is any set containing n different natural numbers between 1 and K.

The c_n can be calculated from the joint probability distribution of internal representations

$$P(\tau_1, \ldots, \tau_k) = \left\langle\!\!\left\langle \frac{\int \prod_{k=1}^K \mathrm{d}\mu(\boldsymbol{J}_k)\ \prod_{k=1}^K \theta(\tau_k \boldsymbol{J}_k \boldsymbol{\xi}_k^1) \prod_\mu \theta(\sigma^\mu F(\{\mathrm{sgn}(\boldsymbol{J}_k\boldsymbol{\xi}_k^\mu\}))}{\int \prod_{k=1}^K \mathrm{d}\mu(\boldsymbol{J}_k)\ \prod_{k=1}^K \prod_\mu \theta(\sigma^\mu F(\{\mathrm{sgn}(\boldsymbol{J}_k\boldsymbol{\xi}_k^\mu)\}))} \right\rangle\!\!\right\rangle \tag{3}$$

The calculation of $P(\tau_1, \ldots, \tau_k)$ parallels the determination of the local aligning field distribution for the perceptron [6, 7] (see also [8, 2, 9]). The general result within replica symmetry is

$$P(\tau_1, \ldots, \tau_k) = \left\langle\!\!\left\langle \delta_{\sigma, F(\{\tau_k\})} \int \prod_k \mathrm{D}t_k\ \frac{\prod_k H(Qt_k\tau_k)}{\mathrm{Tr}'_{\eta_k}\prod_k H(Qt_k\eta_k)} \right\rangle\!\!\right\rangle_\sigma \tag{4}$$

where $\delta_{n,m}$ is the Kronecker symbol and the primed trace $\mathrm{Tr}'_{\eta_k} = \mathrm{Tr}_{\eta_k}\,\delta_{\sigma,F(\{\eta_k\})}$ is restricted to the legal internal representations. Moreover $\mathrm{D}t = \mathrm{e}^{-t^2/2}\mathrm{d}t/\sqrt{2\pi}$ and $H(x) = \int_x^\infty \mathrm{D}t$ as usual. We do not display the saddle-point equation necessary to determine $Q = \sqrt{q/(1-q)}$ as a function of α since we are mainly interested in the saturation limit $\alpha \to \alpha_\mathrm{c}$ which is characterized for networks with continuous weights by $q \to 1$ and therefore $Q \to \infty$.

In this limit the integrand in (4) either tends to zero or to one depending on the values of the t_k. When calculating $P(\tau_1, \ldots, \tau_k)$ explicitly for small K and special decoder functions one realizes a simple general rule. Consider the system before learning. All internal representations have an *a priori* probability 2^{-K}. Those already compatible with the desired output are not modified, all the others are shifted by the learning process *to the nearest decision boundary* of the decoder function. This is reminiscent of the aligning field distribution of the simple perceptron [7, 10] and has a natural interpretation within the cavity approach [9, 11]. On the basis of this general rule it is possible to determine $P(\tau_1, \ldots, \tau_k)$ for arbitrary K and arbitrary decoder function.

As examples we derive in the following explicit results for the PAR-, AND- and COM-machines defined by the decoder functions

$$F(\{\tau_k\}) = \prod_k \tau_k,\ F(\{\tau_k\}) = \mathrm{sgn}\left(\sum_k \tau_k - K + \frac{1}{2}\right) \qquad \text{and} \qquad F(\{\tau_k\}) = \mathrm{sgn}\left(\sum_k \tau_k\right)$$

respectively. For the PAR- and COM-machine we can set all outputs equal to $+1$ without loss of generality for symmetry reasons whereas for the AND-machine we have to stick to random outputs $\sigma^\mu = \pm 1$ with equal probability.

In the case of the PAR-machine all internal representations are at the decision boundary of the decoder function. Hence all LIR gain in addition to their *a priori* weight 2^{-K} an equal share from the 2^{1-K} internal representations that are eliminated by the learning process. Therefore for $\alpha \to \alpha_c$ all LIR have equal probability 2^{1-K} which results in $c_n = 0$ for all $n = 1, \ldots, K-1$ and $c_K = 1$. On the other hand it is known that the PAR-machine saturates the upper limit $\log K / \log 2$ for the storage capacity of any two-layer neural network of tree-architecture [12]. Any machine with lower capacity should therefore be characterized by additional non-zero correlations between its hidden units.

In the case of the AND-machine there is only one LIR for the output $\sigma = +1$, namely $\tau_1 = \cdots = \tau_K = 1$. It contributes $\frac{1}{2}$ to all c_n. If $\sigma = -1$ all but one internal representations are LIR. Only those with exactly one $\tau_k = -1$ are at the decision boundary of the decoder function and consequently only their probability is changed by the learning process. For symmetry reason it is clear that all of them get an equal share $2^{-K}/K$ from the elimination of the internal representation $\tau_1 = \cdots = \tau_K = +1$ in addition to their *a priori* weight 2^{-K}. The calculation of the resulting contribution from $\sigma = -1$ to the correlations c_n can be most easily accomplished by observing that $c_n = 0$ for all n before learning. To calculate c_n after learning one has hence only to take into account those LIR with exactly one $\tau_k = -1$. The result is $c_n = \frac{1}{2} - n2^{-K}/K$. As expected all correlations are dominated by the restrictive case $\sigma = +1$ of the output.

For the COM-machine the calculation is more involved. As usual, we only consider odd values of K. The decision boundary is given by all LIR with $\sum_k \tau_k = 1$. All these gain an equal share from the 2^{1-K} internal representation that have to be eliminated by the learning. Hence

$$
\begin{aligned}
P(\{\tau_k\}) &= 2^{-K} && \text{if } \sum_k \tau_k > 1 \\
P(\{\tau_k\}) &= 2^{-K} + \frac{1}{2}\left[\binom{K}{\frac{1}{2}(K-1)}\right]^{-1} && \text{if } \sum_k \tau_k = 1 \\
P(\{\tau_k\}) &= 0 && \text{otherwise}
\end{aligned}
$$

To determine the values of c_n from this $P(\{\tau_k\})$ it is convenient to consider the contribution from the regular part

$$P^{(r)}(\{\tau_k\}) = 2^{-K} \qquad \text{if } \sum_k \tau_k \geqslant 1$$

and that from the extra part

$$P^{(e)}(\{\tau_k\}) = \frac{1}{2}\left[\binom{K}{\frac{1}{2}(K-1)}\right]^{-1} \qquad \text{for } \sum_k \tau_k = 1$$

separately. The regular part contribution to even moments is zero due to symmetry. Its contribution to odd moments is

$$c_n^{(r)} = 2^{-K} \sum_{m=0}^{\frac{1}{2}(K-1)} \sum_{i=0}^{n} (-1)^i \binom{n}{i}\binom{K-n}{m-i} \tag{5}$$

since $i = 0, \ldots, n$ of the $m = 0, \ldots, \frac{1}{2}(K-1)$ minus ones of a LIR can be found in $\tau_1, \ldots, \tau_n$ whereas the remaining $(m-i)$ minus ones are to be distributed between the remaining $(K-n)$ $\tau_{n+1}, \ldots, \tau_k$. After some algebra using properties of binomial coefficents

[13] this can be simplified to

$$c_n^{(r)} = 2^{-K} \sum_{i=0}^{n-1} (-1)^i \binom{n-1}{i} \binom{K-n}{\frac{1}{2}(K-1)-i} \tag{6}$$

$$= 2^{-K} (-1)^{\frac{1}{2}(n-1)} \frac{(n-2)!!}{(K-2)(K-4)\cdots(K-n+1)} \binom{K-1}{\frac{1}{2}(K-1)} \tag{7}$$

$$= 2^{-K} \frac{\Gamma(\frac{1}{2}n)\ \Gamma(1-\frac{1}{2}K)\ \Gamma(K)}{\sqrt{\pi}\ \Gamma(\frac{1}{2}(n-K+1))\ [\Gamma(\frac{1}{2}(K+1))]^2} \,. \tag{8}$$

Similarly, one gets for the extra part contribution

$$c_n^{(e)} = \frac{1}{2} \left[\binom{K}{\frac{1}{2}(K-1)} \right]^{-1} \sum_{i=0}^{n} (-1)^i \binom{n}{i} \binom{K-n}{\frac{1}{2}(K-1)-i} \tag{9}$$

which results in

$$c_n^{(e)} = \frac{1}{2} (-1)^{\frac{1}{2}(n-1)} \frac{n!!}{K(K-2)(K-4)\cdots(K-n+1)} \tag{10}$$

$$= \frac{\Gamma(\frac{1}{2}n+1)\ \Gamma(1-\frac{1}{2}K)}{2\sqrt{\pi}\ \Gamma(\frac{1}{2}(n-K+1))} \tag{11}$$

for n odd and

$$c_n^{(e)} = -c_{n-1}^{(e)} \tag{12}$$

if n is even. The final result for the correlations of the COM-machine is hence

$$c_n(K) = \frac{\Gamma(\frac{1}{2}(n+1))\ \Gamma(-\frac{1}{2}(K))}{2\ \sqrt{\pi}\ \Gamma(\frac{1}{2}(n-K))} \qquad \text{if } n \text{ even} \tag{13}$$

$$c_n(K) = \frac{\Gamma(\frac{1}{2}(n))\ \Gamma(1-\frac{1}{2}(K))}{\sqrt{\pi}\ \Gamma(\frac{1}{2}(n-K+1))} \left[\frac{\Gamma(K)}{2^K\ [\Gamma(\frac{1}{2}(K+1))]^2} + \frac{n}{2K} \right] \qquad \text{if } n \text{ odd}\,. \tag{14}$$

Note that for n even one has $c_{K-n+1} = (-1)^{(K+1)/2} c_n$. As an example these results are shown in figure 1 for $K = 25$.

It is straightforward to obtain the asymptotic behaviour of the moments for $K \to \infty$. For the COM-machine moments c_n with either n or $K-n$ small remain the largest ones in this limit. Explicitly one gets with the abbreviation $C = 1/\sqrt{2\pi K}$ $c_1 \approx C, c_2 = -1/(2K)$, $c_3 \approx -C/K, c_4 = 3/(2K(K-2)), c_5 \approx 3C/K^2$ and $c_K \approx (-1)^{(K-1)/2}$, $c_{K-1} = (-1)^{(K-1)/2}/(2K), c_{K-2} \approx (-1)^{(K+1)/2}/(2K), c_{K-3} = (-1)^{(K+1)/2} 3/(2K(K-2)), c_{K-4} \approx (-1)^{(K-1)/2} 3/(2K^2)$.

So far we have considered a MLN with fixed decoder function and have determined the correlations c_n resulting near saturation. It is tempting to investigate also the complementary question and to determine the storage capacity of an ensemble of K uncoupled perceptrons with prescribed correlations c_n. For the COM-machine it is known, for example, that already from the prescription of c_1 alone one gets the correct RS-asymptotics $\alpha_c \cong K^2$ for the storage capacity [3] (which is, however, known to be unstable with respect to RSB). It is interesting to see whether the inclusion of other correlations can alter this asymptotics [14].

Finally it should be noted that the results obtained in this letter rely on the assumption of replica symmetry for the determination of the aligning field distribution whereas it is well known that replica symmetry breaking is crucial for the calculation of the storage

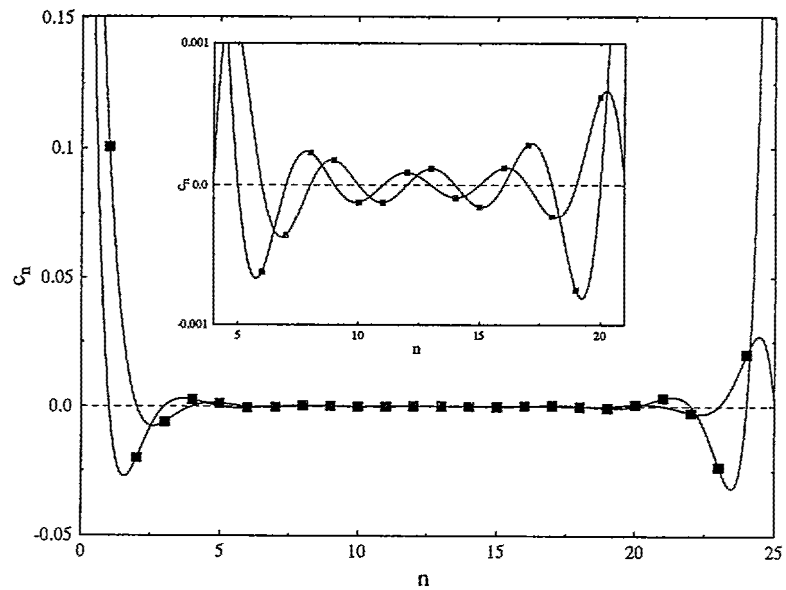

Figure 1. Moments of internal representations for a committee machine with $K = 25$ hidden units. The symbols are the results for integer n following from (5) and (9), the full and dotted line are given by (13) and (14), respectively. The inset shows an enlarged region of the plot.

capacity of MLN. On the other hand it is merely the *qualitative* behaviour of the aligning field distribution that is important for the determination of the c_n. Since this is known to be hardly modified by RSB [15, 16] it seems likely that the results for the correlations will not be significantly altered by the inclusion of RSB.

It is a pleasure to thank the Minerva Center for Neural Networks for hospitality during the initial stages of this work that were performed at BarIlan University in Ramat Gan. Moreover I am indebted to Professor I Kanter for several interesting remarks on the manuscript.

References

[1] Mezard M and Patarnello S 1989 On the capacity of feedforward layered networks *LPTENS preprint*
[2] Griniasty M and Grossman T 1992 *Phys. Rev.* A **45** 8924
[3] Priel A, Blatt M, Grossman T, Domany E and Kanter I 1994 *Phys. Rev.* E **50** 577
[4] Schottky B 1995 *J. Phys. A: Math. Gen.* **28** 4515
[5] Monasson R and Zecchina R 1995 *Phys. Rev. Lett.* **75** 2432
[6] Kepler T B and Abbott L F 1988 *J. Physique* **49** 1657
[7] Gardner E 1989 *J. Phys. A: Math. Gen.* **22** 1969
[8] Barkai E, Hansel D, and Sompolinsky H 1992 *Phys. Rev.* A **45** 4146
[9] Griniasty M 1993 *Phys. Rev.* E **47** 4496
[10] Opper M 1988 *Phys. Rev.* A **38** 3824
[11] Wong K Y M 1995 *Europhys. Lett.* **30** 245
[12] Barkai E, Hansel D and Kanter I 1990 *Phys. Rev. Lett.* **65** 2312
[13] Knuth D E 1967 *The Art of Computer Programming* vol 1 (Reading, MA: Addison-Wesley)
[14] Malzahn D, Engel A and Kanter I 1996 in preparation
[15] Erichsen R and Theumann W K 1993 *J. Phys. A: Math. Gen.* **26** L61
[16] Majer P, Engel A and Zippelius A 1993 *J. Phys. A: Math. Gen.* **26** 7405