Physica A 200 (1993) 670–678
North-Holland

*SDI:* 0378-4371(93)E0209-W

PHYSICA A

# The binary perceptron and general aspects of non-self-averaged quantities

I. Kanter and M. Shvartser

*Department of Physics, Bar-Ilan University, Ramat-Gan 52900, Israel*

The possibility of a finite width distribution for the maximal capacity of the binary perceptron in the thermodynamic limit is discussed analytically and supported by a careful analysis of numerical simulations. The results also indicate that the description of quenched random systems could take into account the possibility that in addition to non-self-averaged quantities, other quantities such as the transition temperature might also be sample dependent.

The storage capacity of neural networks has been investigated in the recent past by the theoretical method suggested by Elisabeth Gardner [1,2]. One of the remarkable predictions of this method is the calculation of the maximal storage capacity, which in the case of feedforward networks is defined as the maximal number of input/output relations which can be embedded in the network (see, for example, [3]). Nevertheless, in order to solve a particular task or to verify the predictions of the theory, a learning algorithm has to be defined. The learning algorithm is an effective procedure for finding weights which fulfil the task of the network. Unfortunately, such an algorithm with a proof of convergence is known almost only to the simplest feedforward architecture with continuous weights. This architecture, known as a perceptron, consists of one layer of $N$ binary input units, $\{s_i\}$, which is connected to one binary output, $\{O\}$, via $N$ continuous weights, $\{J_i\}$. The output is a function of the input units and the weights and is fixed under zero-temperature dynamics by

$$O = \operatorname{sgn}\left(\sum_{j=1}^{N} \frac{J_j s_j}{\sqrt{N}}\right). \tag{1}$$

The simplest task of the network consists of mapping of a set of $P$ input patterns $\{\xi_i^\mu\}$, $i = 1, 2, \ldots, N$, $\mu = 1, \ldots, P$, onto a set of $P$ outputs $\{y^\mu\}$, where $y^\mu$ and $\xi_i^\mu$ are equal to $\pm 1$ with equal probability. Since $P$ is proportional to $N$ it is convenient to define the quantity $\alpha \equiv P/N$. It is well

known that the maximal capacity for a perceptron with continuous weights is $\alpha_c = 2$ [1,2], where a particular task can be implemented, for instance, by the perceptron learning algorithm [1]. Nevertheless, discrete weights are more sensible for biological networks and for the construction of real devices.

The binary perceptron, which is at the center of this study, has the same architecture as for the continuous one, but the weights can take only the values $\pm 1$. The first analytical study of this model and within the replica symmetric (RS) assumption gives $\alpha_c = 4/\pi$ [2]. This calculation was made under the assumption that the maximal capacity is obtained as the overlap between two sets of weights which fulfil the task of the network tends to 1. Nevertheless, the actual maximal capacity should be smaller, since information theory gives an upper bound $\alpha_c \leq 1$ [2], and furthermore the RS solution becomes locally unstable above $\alpha \simeq 1.015$ [4]. The exact $\alpha_c$ was found to be $\sim 0.833$, where the entropy of the solutions vanishes, but the overlap between two sets of weights which fulfil the task of the network is less than 1. Note that the exact $\alpha_c$ is obtained within the RS assumption.

Unfortunately, no convergent learning algorithm is known to the problem of the binary perceptron, and there are three main methods in which this problem was examined numerically.

In the first method, a pattern is chosen randomly and configurations which do not give the desired output are discarded. The process is repeated until $P = P_c + 1$ where no configurations are let [6]. The value of $\alpha_c$ $(N)$ is then averaged over many samples. Since the number of configurations grows exponentially with $N$ it is clear that this method is limited to small systems. The best approximation of this method gives, in the thermodynamic limit, $\alpha_c = 0.75 \pm 0.05$ [5,6], which is below the theoretical prediction [4]. Furthermore, in any extrapolation which assumes that $\alpha_c(N)$ is monotonic with $1/N$ it seems that $\alpha_c$ should be less than 0.833 [6]. The second method suggests the use of Gaussian patterns instead of Ising patterns. The assumption in this method is that the statistical properties of the model depend in the thermodynamic limit only on the first and the second moments of the distribution of the quenched random variables, which was confirmed analytically in the SK model [7]. The advantage of Gaussian patterns over Ising patterns is that the values for the induced field on the output unit $(\Sigma J_j \xi_i^\mu)$ are continuous as in the case of Ising patterns but in the thermodynamic limit. Under this assumption there are two methods for the calculation of the maximal capacity. In the first method $\alpha_c$ is calculated as in the first method and it was found that $\alpha_c \simeq 0.833$ [6]. In the second method, the optimal stability is calculated, $k_{opt} = \max_J \{k_J\}$, where $k_J = \min\{\Sigma J_j \xi_j^\mu / \sqrt{N}\}$. The capacity where $K_{opt}$ vanishes is defined as $\alpha_c$ and is found to be $\simeq 0.82$ [8]. In the third method some heuristic algorithms are done, which are mostly a combination of the perceptron learning algorithm or a gradient descent procedure and stochastic source as in simulated annealing.

The problem with these heuristic algorithms is that the fraction of the volume which fulfils the task of the network drops exponentially to zero with $N$, and therefore the number of steps should grow exponentially. This is the source of the decreasing of $\alpha_c$ with $N$ even below 0.7 [9–11].

In the present situation it is fair to say that the comparison between the theory and the simulations of the binary perceptron is in question. Is the capacity of the binary perceptron with binary patterns indeed 0.833? Is the capacity of the binary perceptron with discrete and Gaussian inputs the same? Is $\alpha_c(N)$ monotonic with $N$?

Another fundamental question (previously mentioned in ref. [4]) is whether the distribution of $\alpha_c$ in the thermodynamic limit is a delta function or has a finite width. The assumption that the distribution of $\alpha_c$ in the thermodynamic limit is a delta function stands as the foundation of Gardner's method. Indeed, in the case where $\alpha_c$ has a finite width in the thermodynamic limit, the average $\alpha_c$ may differ from the maximal theoretical capacity calculated by the Gardner's method. The proof that the width of the distribution of $\alpha_c$ tends to zero in the thermodynamic limit has been obtained, to our knowledge, only for one architecture, the perceptron with continuous weights, [12]. Hence, the calculation of the width of the distribution of $\alpha_c$ is necessary for the physical understanding of the meaning of $\alpha_c$.

Let us now discuss analytically the possibility that even in the thermodynamic limit the distribution of $\alpha_c$ has a finite width. This possible surprising result is supported by simulations which are presented later on.

The number of possible different inputs for the binary perceptron of size $N$ is $2^N$. Each realization of the $N$ weights is implementing one boolean function (BF) among $2^{2^N}$. A BF is a list of the $2^N$ outputs for the correspondent $2^N$ possible inputs. Since each configuration of the weights, among the $2^N$, is implementing a different boolean function, the binary perceptron is implementing $2^N$ different boolean functions (BFs). These $2^N$ BFs define the computational ability of the binary perceptron.

Each BF is implementing $\binom{2^N}{P}$ different cases of $P$ input/output relations. Each such realization of $P$ input/output relations is denoted later as a symbol of size $P$. The maximal number of different symbols of size $P$ is $\binom{2^N}{P}2^P$. The statement that the system has a well defined maximal capacity, $P_c$, indicates that almost all symbols of size $P$ ($P < P_c$) can be implemented. The balance between the number of different symbols of size $P$ and the number of symbols of size $P$ which can be supported by the $2^N$ BFs of the binary perceptron is in the center of the following discussion. In the case that the overlap between the $2^N$ BFs is ignored, the maximal capacity is fixed by the following inequality:

$$\binom{2^N}{P_c} 2^{P_c} \leqslant 2^N \binom{2^N}{P_c}. \tag{2}$$

The left side of eq. (2) is the maximal number of different symbols of size $P_c$ and the right side is an upper bound for the number of symbols of size $P$ which can be obtained from $2^N$ BFs where each symbol appears only once. It is obvious that the solution of eq. (2) is $\alpha_c \leqslant 1$, which is the bound of information theory [6]. The following discussion is centered around the effect of the correlations among the BFs, which decreases the right side of eq. (2) and therefore gives a better estimation for the capacity.

If the $l$th BF is defined as $\{y_\mu^l\}$, $\mu = 1, \ldots, 2^N$ and $y_\mu^l = \pm 1$, then the magnetization $q^{ll'}$ is defined by

$$q^{ll'} = \frac{1}{2^N} \sum_{\mu=1}^{2^N} y_\mu^l y_\mu^{l'} . \tag{3}$$

Note that since the binary perceptron obeys a global inversion symmetry, for each BF $l$ there is a BF $l'$ such that $q^{ll'} = -1$. It is clear that it is possible that the same symbol may be constructed by many BFs. Hence, the maximal number of symbols of size $P$ which can be constructed by the binary perceptron is obtained in the case where the overlap among the BFs is minimal. This condition is obtained in the case where for any $l \neq l'$ $q^{ll'} = 0$, except the $2^{N-1}$ cases where $q = -1$.

Let us now choose a pair of BFs $l$, $l'$ such that $q^{ll'} = -1$. The number of different symbols of size $P$ which can be constructed from this pair is

$$2\binom{2^N}{P} . \tag{4}$$

Another pair $mm'$ of BFs contributes

$$2\left[\binom{2^N}{P} - 2\binom{2^{N-1}}{P}\right] \sim 2\binom{2^N}{P}\left(1 - \frac{2}{2^P}\right) \tag{5}$$

new symbols of size $P$. Since $q^{ml} = q^{ml'} = q^{m'l} = q^{m'l'} = 0$, $2\binom{2^{N-1}}{P}$ is the number of symbols of size $P$ in the new BF which already exists in the previous pair. Hence, one can verify that the number of different symbols of size $P$ in all the $2^N$ BFs is given by

$$2\binom{2^N}{P} \sum_{l=0}^{2^{N-1}} \left(1 - \frac{2}{2^P}\right)^l \sim \binom{2^N}{P} 2^P . \tag{6}$$

This number is equal to the maximal number of different symbols of size $P$ (see eq. (2)).

Note that eq. (6) gives an upper bound, since in the case of macroscopic

overlaps among the BFs the expression $1-2/2^P$ in eq. (5) should be replaced by $1-[(1+q)/2]^P-[(1-q)/2]^P$, where $q$ is a macroscopic overlap which represents the overlaps among the BFs. Note that the qualitative conclusions of the following discussion are insensitive to a finite width distribution of $q$, but it is clear that smaller $q$ gives a higher number of different symbols.

The distribution of the macroscopic overlap, $q$, was examined numerically for $3 \leq N \leq 27$. For $N \leq 11$ an exhaustive search was carried out over all pairs of the $2^N$ BFs and for $N > 11$ only random subspaces of the BFs were examined. The results indicate, that the averaged $|q^{ll'}|$ scales with $1/\sqrt{N}$, but the minimal non-zero value of $|q^{ll'}|$ scales with $1/N$ (fig. 1). More precisely, for $N = 4m+3$ ($m$ is an integer), there is also a possibility for $q^{ll'} = 0$, where for $N = 4m+1$ this possibility is absent. Nevertheless, in the following we assume that the distribution of $\alpha_c$ has a well defined limit as $N \to \infty$ and the effective minimal $|q^{ll'}|$ is at least of $\mathcal{O}(1/N)$. Using this fact and *assuming* that the only relevant distribution of $x \equiv y_\mu^k y_\mu^l$ $(k \neq l)$ is such that $P(x) = 0.5(1+q)\delta(x-1) + 0.5(1-q)\delta(x+1)$, eq. (6) is given now by

$$2\binom{2^N}{P}\sum_{l=0}^{2^{N-1}}\left[1-2\left(\frac{1+q}{2}\right)^P-2\left(\frac{1-q}{2}\right)^P\right]^l \sim \binom{2^N}{P}\frac{2^P}{\cosh(A\alpha)}, \tag{7}$$

where $q = A/N$ is the minimal magnetization. Hence, the fraction of BFs of size $P$ which can be implemented by the binary perceptron is less than 1 for any finite $\alpha$ and is given by $1/\cosh(A\alpha)$. This fraction obviously converges to 1 (0) for $\alpha \to 0$ $(\infty)$.

In the simulations, a careful analysis of the results of an exhaustive search over the $2^N$, $5 \leq N \leq 27$, configurations has been carried out. For each $\alpha$, the

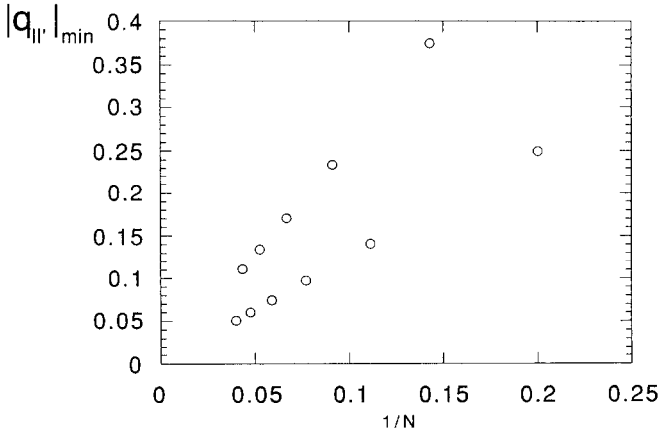


Fig. 1. The minimal $|q^{ll'}|$ versus $1/N$.

fraction $f(N, \alpha)$ of samples which has at least one configuration which classifies correctly the $\alpha N$ patterns is calculated. The results are averaged over 100 000 samples for small systems and over at least 5000 for the largest system. This fraction is then plotted as a function of $\alpha$, where a polynomial fitting (the degree depends on $N$ and $\alpha$) is used to extrapolate $f(N, \alpha)$ among the points. This data is then used to construct $\alpha(N, f)$. It is clear that if the distribution of $\alpha_c$ in the thermodynamic limit is a delta function then the extrapolated $\alpha(\infty, f)$ should be independent of $f$ and $\alpha_c$ is a well defined scalar. In fig. 2, $\alpha(N, f)$ is plotted as a function of $1/N$ for different fractions. From the results for $N < 19$ and $f = 1/2$, for instance, it seems that $\alpha(N, 1/2)$ scales with $1/N$ [6]. However, a remarkable deviation from this scaling is observed for $N \geq 19$. In contrast, an excellent parabolic fitting is found for $\alpha$ as a function of $1/N$. The values of the minima of these parabolas are plotted in the insert of fig. 2. Note that previous works assumed that $\alpha_c$ is well defined and therefore only $\alpha(\infty, 1/2)$ was calculated. A fit to higher order polynomials given non-physical results, since lines of different fractions intersect and $\alpha$ (even for $f < 1/2$) is not monotonic with $N$.

An interesting conclusion from these results is that asymptotically for large $N$ there is a critical fraction $f_0$, such that for $f < f_0$ the capacity is a decreasing function of $N$ where for $f > f_0$ the capacity is an increasing function of $N$. Such behavior is expected in the case of the perceptron with continuous weights, where $\alpha(N, f \gg 1/2)$ ($\alpha(N, f \ll 1/2)$) decreases (increases) to 2 as a function of $N$. Note that for small $N$ and large $f$ the extrapolation is expected to be poor, since $\Delta\alpha = 1/N$ and $f \to 1$. Nevertheless, results of fig. 2 indicate that the

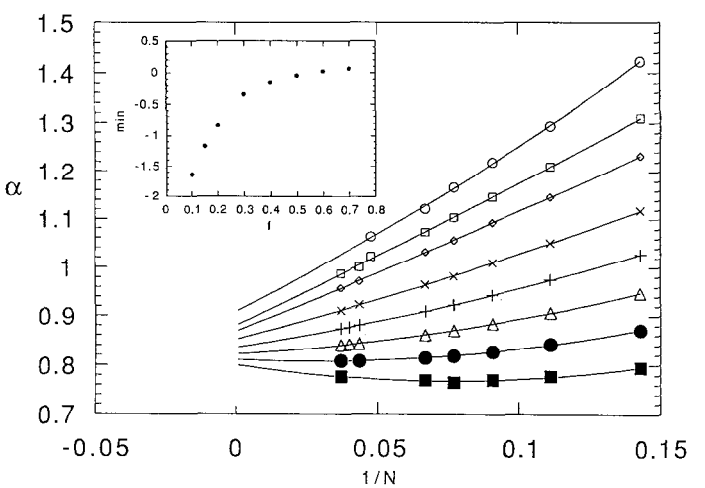


Fig. 2. $\alpha$ versus $1/N$ for different fractions $f$. $f = 0.1$ (○), 0.15 (□), 0.2 (◇), 0.3 (×), 0.4 (+), 0.5 (△), 0.6 (●), 0.7 (■). Insert: the minima of the parabolas versus $f$.

distribution of $\alpha_c$ has a finite width, since $\alpha(\infty, f)$ is a function of $f$, a fact which supports the abovementioned calculation. For $f = 0.7$, 0.6, 0.5, 0.4, 0.3, 0.2 and 0.1 the estimated $\alpha_c$ is $\simeq$0.8, 0.811, 0.822, 0.836, 0.851, 0.868 and 0.90, respectively. It is clear that $\alpha_c(\infty, f)$ is weakly dependent on $f$ as it is expected from the approximation eq. (7), where $d\alpha(\infty, f)/df$ diverges only for $\alpha \to 0$ or 1. Nevertheless, the exact form of the distribution is sensitive to the exact form of the correlations among the BFs.

The maximal capacity was also calculated numerically by the method of ref. [6], where a pattern was added sequentially and configurations which do not give the desired output are discarded. The process is repeated until $P = P_c + 1$ where no configurations are left. The results for the averaged $\alpha_c$ and the variance of the distribution are plotted in fig. 3, where each point is averaged over at least 10 000 samples. The maximal capacity seems to be slightly greater than 0.83 and the variance seems to converge to zero or to a small number in the thermodynamic limit. The solution to this contradiction is that the two discussed numerical methods to estimate the maximal capacity are the same only under the assumption of a sharp transition, $f = \Theta(\alpha_c - \alpha)$. In the case of a finite width distribution, then all possible sets of $\alpha N$ patterns contribute to the statistics of $f(\alpha, N)$, independent of the number of errors. In contrast, in the discussed simulations only sets of $\alpha N$ patterns with *one* error contribute to the statistics.

There is only one way to explain the results of this work with the results of the replica calculation, where a sharp transition occurs at $\alpha = 0.833$. As a result of correlations among the BFs, only a small fraction of symbols have entropy of $\mathcal{O}(N)$. The fraction of such symbols either shrinks to zero at $\alpha_c$ or the product of their probability by their volume becomes less than 1 at $\alpha_c$, which is

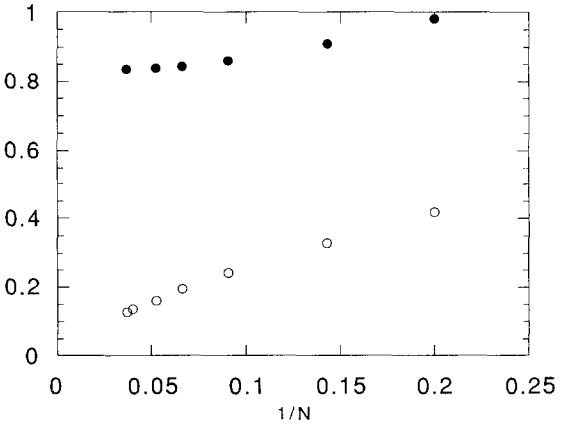


Fig. 3. $\alpha_c$ versus $1/N$ (●) and the dispersion versus $1/N$ (○).

consistent with the fact that even below $\alpha_c$, $f < 1$. Furthermore, the lack of a finite fraction of symbols with entropy of $\mathcal{O}(N)$ above $\alpha_c$ cannot rule out the possibility of a finite fraction of symbols with entropy of $\mathcal{O}(1)$.

The main reason for the lack of such observation within the framework of the replica method is that the calculation is done only in the leading order. The average volume is given by $e^{nNG}$, where the value of $G$ is fixed by the saddle point equations. In the case where higher order terms diverge or $G$ vanishes, then higher order terms become dominant. As a consequence, it is clear that in the case where a finite fraction of the symbols has zero volume, the average volume per symbol can be less than 1. This might be the physical interpretation for negative entropy above $\alpha_c$, even in the case of discrete space. Note that the entropy does not diverge to $-\infty$ at $\alpha_c$ [3], as is expected where no solution exists.

The results of this work raise a fundamental question which were formerly considered trivial. The role of the capacity is similar in some sense to the inverse temperature. The entropy decreases, for instance, as $\alpha$ increases. Therefore, it is a strong assumption that the transition temperature in the thermodynamic limit, for instance, is a self-averaged quantity. Note that in the known theories of disordered spin systems, non-self-averaged quantities are found under the assumption that the transition temperature is fixed for any sample. However, if this assumption is violated, then the physical interpretation of the functional order parameter in the Parisi scheme, for instance, should be reconsidered [13]. Furthermore, if such a behavior is relevant to Hamiltonian systems then a statement such as "this sample is bad" by an experimentalist should also be reconsidered.

Discussions and comments on the manuscript of B. Derrida, E. Domany and T. Grossman are gratefully acknowledged. The research is supported by The Basic Research Foundation administered by The Israel Academy of Science and Humanities.

## References

[1] E. Gardner, J. Phys. A 21 (1988) 257.
[2] E. Gardner and D. Derrida, J. Phys. A 21 (1988) 271.
[3] J. Hertz, A. Krogh and R. Palmer, Introduction to the Theory of Neural Computation (Addison–Wesley, Reading, MA, 1991).
[4] W. Kraut and M. Mézard, J. Phys. (Paris) 50 (1989) 3054.
[5] E. Gardner and B. Derrida, J. Phys. A 22 (1989) 1983.
[6] B. Derrida, B. Griffiths and R.B. Prugel-Bennett, J. Phys. A 24 (1991) 4907.
[7] D. Sherrington and S. Kirkpatrick, Phys. Rev. Lett. 35 (1975) 1792.

[8] W. Kraut and M. Opper, J. Phys. A 22 (1989) L519.
[9] H. Kohler, J. Phys. A 23 (1990) L1265.
[10] H. Horner, Z. Phys. B 86 (1992) 291.
[11] H. Gutfreund and Y. Stein, J. Phys. A 23 (1990) 2613.
[12] T. Cover, IEEE Trans. Electron. Comput. 14 (1965) 326.
[13] M. Mézard, M.A. Virasoro and G. Parisi, Spin Glass Theory and Beyond (World Scientific, Singapore, 1987).