Critical storage capacity of the J = ± 1 neural network

1989 J. Phys. A: Math. Gen. 22 L519

(http://iopscience.iop.org/0305-4470/22/11/012)

J. Phys. A: Math. Gen. 22 (1989) L519-L523. Printed in the UK

LETTER TO THE EDITOR

Critical storage capacity of the $J = \pm 1$ neural network

Werner Krauth† and Manfred Opper‡§

† Laboratoire de Physique Statistique, l'Ecole Normale Supérieure, 24 rue Lhomond, 75231 Paris Cedex 05, France
‡ Institut für Theoretische Physik, Justus Liebig Universität, Heinrich Buff Ring 16, D-6300 Giessen, Federal Republic of Germany

Received 2 March 1989

Abstract. For neural networks in which the couplings J_{ij} are allowed to take on the values $J_{ij} = 1$ or $J_{ij} = -1$, we determine numerically the critical storage capacity for random unbiased patterns as a function of the stability. We use an exact enumeration scheme based on the Gray code and a continuous distribution for the patterns to control finite-size effects. Results are presented for $N \leqslant 25$; they indicate an optimal storage capacity of $\alpha_c \simeq 0.82$ ($N \to \infty$).

An important result in the 'modern' theory of neural networks is due to Gardner (1987, 1988). By means of a replica calculation she determined the typical volume of real-valued interactions J_{ij} ($i, j = 1, \ldots, N$; $(N \to \infty)$) with $\Sigma_j J_{ij}^2 = N$ which satisfy

$$\xi_i^\mu \sum_j J_{ij} \xi_j^\mu \geqslant \kappa \sqrt{N} \qquad i = 1, \ldots, N \qquad \mu = 1, \ldots, p. \tag{1}$$

In (1) the ξ_i^μ are fixed independent random patterns, which take on the values $\xi_i^\mu = +1$ and $\xi_i^\mu = -1$ with probability $(1+m)/2$ and $(1-m)/2$, respectively (m is called bias).

This problem is in some respect the inverse of the mean-field theory of spin glasses, where, for example, the properties of a Hamiltonian

$$H = -\sum_{i,j} J_{ij} S_i S_j \tag{2}$$

are considered. There (J_{ij}) is a fixed (symmetric) random matrix and the spins are allowed to vary, i.e. the roles of couplings and spins are interchanged. The spherical condition $\Sigma_j J_{ij}^2 = N$ for the neural network then corresponds to the spherical model of spin glasses ($\Sigma_i S_i^2 = N$) (Kosterlitz *et al* 1976).

The calculation of Gardner marked a significant progress in the theory because it showed that the replica method can be used in the phase space of couplings J_{ij} and, more generally, for systems in which these couplings are not necessarily symmetric ($J_{ij} \neq J_{ji}$). In addition, her results were of much practical importance as they settled the problem of the optimal storage capacity in Hopfield-type neural networks under the zero temperature dynamics

$$s_i^{(t+1)} = \operatorname{sgn} \sum_j J_{ij} s_j^{(t)} \qquad i = 1, \ldots, N. \tag{3}$$

§ Presently at Laboratoire de Physique Théorique, l'Ecole Normale Supérieure, Unité propre du CNRS, associé à l'Ecole Normale Supérieure et à l'Université de Paris-Sud.

0305-4470/89/110519+05$02.50

The stability condition for all patterns $\boldsymbol{\xi}^\mu (=(\xi_1^\mu, \xi_2^\mu, \ldots, \xi_N^\mu))$ under the dynamics (3) is $\xi_i^\mu = \mathrm{sgn}(\Sigma_j J_{ij}\xi_j^\mu)$ ($i = 1, \ldots, N$; $\mu = 1, \ldots, p$); this is equivalent to (1) with stability $\kappa = 0$. The optimal storage capacity $\alpha_c = \alpha_c(\kappa = 0)$ appears in Gardner's elegant formulation as the value of $\alpha = p/N$ at which the typical fractional volume of the sphere $\Sigma_j J_{ij}^2 = N$ which solves (1) goes to zero, at $\kappa = 0$.

The results are by now well known: $\alpha_c = 2$, for unbiased patterns (cf Cover 1965, Venkatesh 1986); for strongly biased patterns the result is $\alpha_c \sim 1/[(m-1)\log(1-m)]$ ($m \to 1$).

Gardner's calculation for this 'spherical' model $\Sigma_j J_{ij}^2 = N$ is unambiguous: replica symmetry has to be assumed but can be shown to remain intact. Thus, the solution she proposed is at least locally stable. The predictions for $\alpha_c(\kappa)$ have been confirmed numerically with an optimal stability algorithm (Krauth and Mézard 1987).

In this letter we will be concerned with a system in which the coupling matrix consists not of arbitrary real variables, but of Ising-type variables $J_{ij} = +1$ or $J_{ij} = -1$. This system may be thought to model more faithfully practical neural networks in which the precision of the J_{ij} is fixed. Conceptually, the $J = \pm 1$ model offers some advantages over the spherical model, especially since the information content of the coupling matrix can be determined. It follows from information theory that the capacity of this model should be smaller than or equal to 1 (Gardner and Derrida 1988).

This $J = \pm 1$ model of neural networks is related to the Sherrington–Kirkpatrick model of spin glasses in the same way as the $\Sigma_j J_{ij}^2 = N$ model to the spherical one of Kosterlitz *et al*. This suggests that the two network models might have quite different properties as well.

There has been some study of this model before. A replica symmetric calculation is still possible (Gardner and Derrida 1988) (it gives $\alpha_c = 4/\pi$), but the replica symmetry is broken above a de Almeida–Thouless (AT) line, which passes through $\alpha \simeq 1.03$ for $\kappa = 0$. Thus, the critical capacity is basically not known.

The simulation of the $J_{ij} = \pm 1$ neural network is also more complicated. (See Amaldi and Nicolis (1988) and Gardner and Derrida (1988) who obtained rough estimates of α_c.) Since there is no perceptron-type learning algorithm (guaranteed to converge if there exists a solution), not to speak of a well-behaved optimal stability algorithm, we resort to an exact enumeration method. We are thus avoiding the drawbacks of Monte Carlo simulations. To cope with the equally annoying problem of small system sizes, we use here a sophisticated algorithm, which allows us to reach rather large sizes and a continuous probability distribution for the patterns, to reduce finite-size effects.

As in the work of Gardner we consider just one row of the matrix (J_{ij}) (with fixed index i). For one sample of patterns we define the variable $\kappa_J = \min_\mu \{\Sigma J_{ij}\eta_j^\mu/\sqrt{N}\}$, and the optimal stability as $\kappa_{\mathrm{opt}} = \max_J\{\kappa_J\}$. The critical storage capacity $\alpha_c(\kappa)$ is then found as the inverse of $\kappa_{\mathrm{opt}}(\alpha)$. Calculating the optimal stability involves, for one given sample of random patterns $\eta^\mu = \xi_i^\mu \xi^\mu$, to check all the 2^N possible vectors $\boldsymbol{J} = (J_{i1}, J_{i2}, \ldots, J_{iN})$. This is less of a brute force method than it might seem: there is room for algorithmic subtleties, and the use of them in combination with a powerful vector computer allows us to consider (with good statistics) systems up to the size $N = 25$.

After generation of a set of random patterns (we restrict ourselves to unbiased patterns with $m = 0$), the calculation of the optimal stability is not complicated. One just determines for each of the 2^N possible vectors the variable κ_J. Since we are free to choose the order in which the possible vectors $\boldsymbol{J}$ are scanned, we use the minimal change order provided by the Gray code (Reingold *et al* 1977). In the Gray code one vector $\boldsymbol{J}$ is derived from the previous one by flipping just one of the J_{ij}. This simplifies enormously the calculation of the stabilities $\Delta^\mu = \Sigma_j J_{ij}\eta_j^\mu$ at each step, compared to an order of the $\boldsymbol{J}$ analogous to counting in the binary system.

The resulting algorithm is amazingly fast. For our largest systems with $N = 25$ it enumerates all possible couplings in 400 s. Of this time it spends about 1/7 on counting (Gray code) and 6/7 on other computing (calculating stabilities, taking minima and maxima, bookkeeping). This excellent ratio is due to the fact that the numerical part can be performed in parallel.

As we are interested in the optimal stability κ_{opt} for large systems it may seem natural to calculate $\max_J\{\min_\mu \Sigma J_{ij}\eta_j^\mu\}/\sqrt{N}$, to average over many samples and to try one's luck with an extrapolation to $N\to\infty$. This has in fact been done, but without success. The fact that the possible values of the stability are discrete with a spacing of $2/\sqrt{N}$ for binary-valued patterns seems to preclude this approach. In fact, the resulting curves are rather erratic, and there is a large parity effect (see figure 1).

It is for this reason that we add one more trick: instead of restricting the values of the η_j^μ to binary values ± 1 we use random variables η_j^μ with a continuous distribution. For the present model it becomes clear from the theoretical treatment that the statistical properties depend only on the first two moments of the distribution for $N\to\infty$ (cf the situation in the Sherrington–Kirkpatrick (1975) model, in which a Gaussian disorder was used for convenience only).

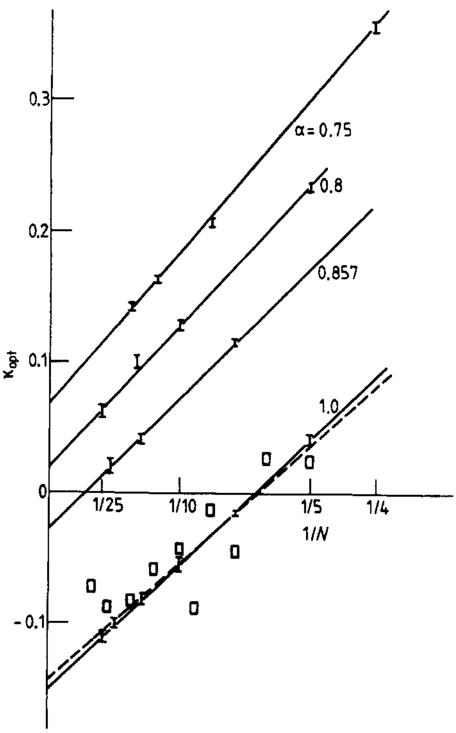

Figure 1. Values of $\kappa_{\text{opt}}(N, \alpha)$ as a function of $1/N$ for $\alpha = \frac{3}{4}, \frac{4}{5}, \frac{6}{7}$ and 1 (plotted as data points with error bars) and their extrapolations to $N\to\infty$. For comparison we display the corresponding values from earlier simulations with $\xi_i^\mu = \pm 1$ for $\alpha = 1$ (plotted as open rectangles).

The model with continuous η has the advantage of yielding continuous values for the stabilities as well. As we will see, the limit $N \to \infty$ is much smoother and an extrapolation becomes possible. For concreteness we choose the normalised Gaussian distribution

$$p(\eta_j^\mu) = (1/\sqrt{2\pi}) \exp[-\tfrac{1}{2}(\eta_j^\mu)^2]. \tag{4}$$

The mean values of the optimal stabilities $\kappa_{\mathrm{opt}}(N, \alpha)$ are found by averaging over a large number of samples (e.g. 10 000 samples at $N = 10$, 500 samples at $N = 25$) for values of α ranging from 0.25 to 2.0. The results for $\alpha = 0.75$, 0.8, 0.857 and 1.0 are given in figure 1. An extrapolation in $1/N$ seems to pose no problem, and leads to rather precise predictions for the value of $\alpha_c(\kappa)$ in the thermodynamic limit $N \to \infty$. To show what we have gained by using continuous patterns, we display for comparison the results of earlier simulations with patterns $\xi_i^\mu = \pm 1$ for $\alpha = 1$. (We stress again that in the limit $N \to \infty$ both distributions of patterns must yield the same $\alpha_c(\kappa)$.)

The extrapolated values of $\alpha_c(\kappa)$ are given in figure 2. We use the opportunity to display our numerical results together with the results for $\alpha_c(\kappa)$ in the replica symmetric approximation and for the AT line. The critical capacity for the spherical model is included also.

The evolving picture of our numerical investigation is the following: the critical storage ratio α_c of the $J = \pm 1$ neural network seems to be close to 0.82. For all values of κ we find results for $\alpha_c(\kappa)$ which are below the AT line and, as it should be, below the critical capacity of the spherical model. It is needless to caution that these statements (for $N \to \infty$) critically depend on the validity of our finite-size scaling hypothesis, which

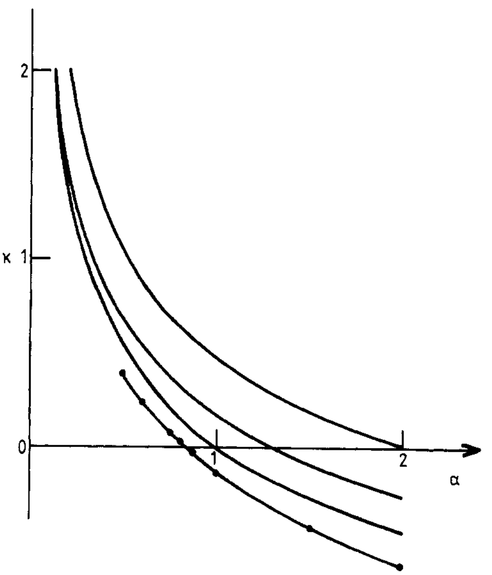

Figure 2. Critical lines for Hopfield-type models. The points on the lowest curve give the values of κ_{opt} as determined numerically for $\alpha = \frac{1}{2}, \frac{4}{5}, \frac{6}{7}, 1, \frac{3}{2}, 2$ (the curve is a guideline to the eye). The other curves give, from above, the critical storage capacity $\alpha_c(\kappa)$ for the 'spherical' model ($\Sigma_j J_{ij}^2 = N$), the storage capacity for the $J = \pm 1$ model in the replica-symmetric approximation, and the de Almeida–Thouless line, above which replica symmetry is broken.

should be checked at larger values of N and which should be given a theoretical foundation. Certainly this model merits further theoretical study.

We would expect that the efficient enumeration provided by the Gray code may be used in different circumstances, as for exhaustive studies of spin glasses.

We would like to thank J Anlauf, M Mézard and J-P Nadal for helpful discussions. The calculations were performed on a Convex C1 vector computer supported by GRECO 70 'Expérimentation Numérique'. We acknowledge financial support from Studienstiftung des deutschen Volkes (WK) and from Deutsche Forschungsgemeinschaft (MO).

References

Amaldi E and Nicolis S 1988 *Preprint* Rome University No 642
Cover T M 1965 *IEEE Trans.* **EC-14** 326
Gardner E 1987 *Europhys. Lett.* **4** 481
—— 1988 *J. Phys. A: Math. Gen.* **21** 257
Gardner E and Derrida B 1988 *Preprint* Saclay SPhT/88-198
Kosterlitz J M, Thouless D J and Jones R C 1976 *Phys. Rev. Lett.* **35** 1792
Krauth W and Mézard M 1987 *J. Phys. A: Math. Gen.* **20** L745
Reingold E M, Nievergelt J and Deo N 1977 *Combinatorial Algorithms: Theory and Practice* (Englewood Cliffs, NJ: Prentice Hall)
Sherrington D and Kirkpatrick S 1975 *Phys. Rev. Lett.* **35** 1792
Venkatesh S 1986 *Proc. Conf. on Neural Networks for Computing* (Snowbird, Utah) (AIP Conf. Proc. **151**) ed J S Denker (New York: American Institute of Physics)