An exact learning algorithm for autoassociative neural networks with binary couplings

1997 J. Phys. A: Math. Gen. 30 2349

(http://iopscience.iop.org/0305-4470/30/7/016)

An exact learning algorithm for autoassociative neural networks with binary couplings

G Milde and S Kobe

Institut für Theoretische Physik, Technische Universität Dresden, D-01062 Dresden, Germany

Received 7 October 1996

Abstract. Exact solutions for the learning problem of autoassociative networks with binary couplings are determined by a new method. The use of a branch-and-bound algorithm leads to a substantial saving of computational time compared with complete enumeration. As a result, fully connected networks with up to 40 neurons could be investigated. The network capacity is found to be close to 0.83.

The training of neural networks with binary couplings is believed to belong to the class of NP-complete problems, i.e. the average computational time required to find a solution scales exponentially with the number of couplings to determine. This exponential scaling is due to the discrete structure of the space of couplings and is obvious in the case of complete enumeration. However, theoretical [1] and numerical [2] studies have shown that it also holds for heuristic approaches (e.g. simulated annealing). Training by complete enumeration has been carried out for small networks with up to 25 neurons [3–5]. Heuristic algorithms [2, 6, 7] were used for networks with up to 1000 neurons. Still, the main disadvantage of heuristic algorithms is the uncertainty about the existence of solutions not found by the algorithm.

Our aim has been to develop an exact algorithm guaranteed to find all possible solutions in considerably less computational time than complete enumeration. In [8], Gardner showed that the space of interactions in neural network models can be treated in a similar way to the phase space of spin glass models. Accordingly, it should be possible to use the branch-and-bound method, already successfully applied to the search for ground states of an Ising spin glass model [9, 10], for the training of neural networks with binary couplings.

Consider an autoassociative network built of N two-state neurons $s_i = \pm 1 (i = 1 \ldots N)$ and fully connected by binary synaptic couplings that can take on the values $J_{ij} = \pm 1$. The self-couplings, J_{ii}, should be set to zero. The task of the network would be to store a set of patterns $\boldsymbol{\xi}^\mu$ $(\mu = 1 \ldots p)$ with elements $\xi_i^\mu = \pm 1$. A training procedure determines couplings that make this pattern's attractors of the discrete network dynamics

$$s_i^{(t+1)} = \operatorname{sgn}\left(\sum_{j=1}^{N} J_{ij} s_j^{(t)}\right) \qquad i = 1 \ldots N. \tag{1}$$

The capacity of the network specifies the number of different patterns that can be stored simultaneously. It is normally expressed as a critical load $\alpha_\mathrm{c} = p_\mathrm{c}/N$.

0305-4470/97/072349+04$19.50 © 1997 IOP Publishing Ltd

For good retrival, one is interested in large basins of attraction. As discussed in [11, 12], these correspond to large values of the pattern stability

$$\kappa^\mu = \min_i \left(\frac{1}{\sqrt{N}} \sum_{j(\neq i)} \xi_i^\mu J_{ij} \xi_j^\mu \right). \tag{2}$$

The maximally stable rule therefore formulates the learning problem as an optimization task. For a given set of patterns, one has to determine an optimal set of couplings that maximizes the network stability

$$\kappa = \min_\mu (\kappa^\mu). \tag{3}$$

As long as there is no symmetry constraint on the matrix of couplings, the optimization task separates into the training of N simple perceptions with $N-1$ input neurons, corresponding to the individual rows of the matrix with the self-coupling excluded. The network stability, κ, emerges as the minimum of the 'perception stabilities' κ_i.

The new learning algorithm was developed by using the branch-and-bound method, a standard tool of the combinatorial optimization theory [13]. To find a row of the matrix of couplings with maximal stability κ_i, complete enumeration would check the $2^{(N-1)}$ possible configurations for optimal ones. The branch-and-bound method starts with a division into a hierarchy of subproblems: each single coupling is tested with both possible values yet taking into account the state of the previously (on a trial basis) determined couplings, thus forming a binary tree of 'incomplete' configurations. Only the final level of the tree would contain the 'complete' solutions. This division is the 'branching' part of the algorithm. Standing alone it would double the neccesary computational time. Here the 'bounding' (and subsequently cutting) part comes into action: for each node of the binary tree an upper bound for the best possible solution of the remaining subproblem is evaluated. The starting point is an ideal stability, $\kappa_{\mathrm{id}} = N - 1$, which is obtained if one takes all terms in the sum (2) to be positive. (Generally, the maximal stability lies below this ideal stability which can only be achieved if there is just one pattern to store.) When testing a coupling J_{ij}, this bound will be corrected, taking into account the already-fixed part of the configuration. If it falls under a preset value, for example the stability attained by using the clipped Hebb rule, the binary tree is 'cut' at this node, i.e. the subtree of this node does not need to be considered. As a result, only a small percentage of the nodes has to be checked. For $N = 25$ we found that only 10^{-4}–8% of the nodes were evaluated, depending on, e.g. the number of patterns to store.

Assuming that the evaluation of a node of the binary tree is approximately as time consuming as checking one possible configuration during complete enumeration, a comparision of these two methods has been done. As predicted by theory, we still have an exponential scaling of the algorithm. However, if we set the load $\alpha = p/N$–0.5 and look for one solution with positive stability, the algorithm no longer scales with 2^N but with $2^{0.46N}$. In the (worst) case of determining all optimal solutions at $\alpha = 1$, the scaling is $3 \times 2^{0.8N}$. (That would mean that the algorithm still optimizes a 30-neuron network in approximately the computational time needed for the complete enumeration of a 25-neuron network.)

We used the branch-and-bound algorithm to determine the capacity of the network storing random uncorrelated patterns. Only one row of the coupling matrix was considered assuming the stability value to be self-averaging in the thermodynamic limit (cf [8, 3]).

The procedure resembles the one used in [4]. For a given value of N, the stability $\kappa^N(\alpha)$ is determined for an increasing number of patterns until its value becomes negative, signifying that it is no longer possible to store all patterns. Then the capacity $\alpha_{\mathrm{c}}(N)$

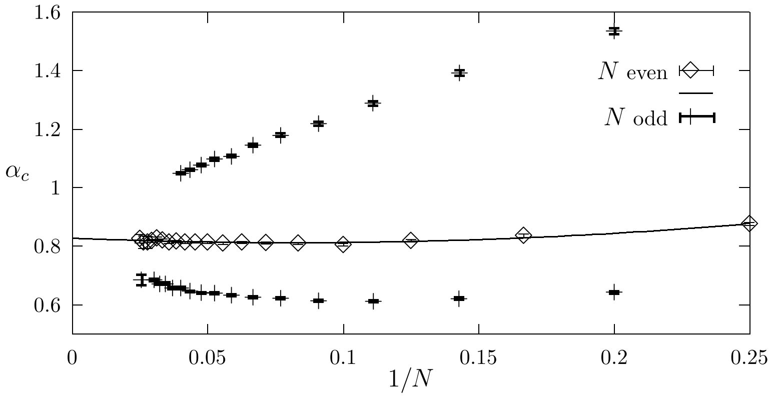

Figure 1. Network capacity in the case of random uncorrelated ±1-patterns

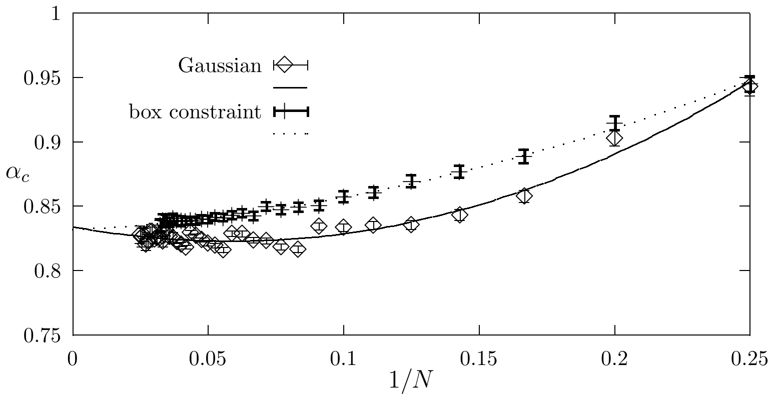

Figure 2. Network capacity in the case of continuous distributed patterns

is determined by linear interpolation between the last positive $\kappa^N(\alpha_+)$ and the negative $\kappa^N(\alpha_-)$. If the patterns are binary-valued, $\xi_i^\mu = \pm 1$, $\kappa^N(\alpha)$ takes on discrete values with a spacing of $2/\sqrt{N}$. For N odd, this discreteness results in two values of $\alpha_c(N)$ corresponding to the first and last occurence of $\kappa^N(\alpha) = 0$. This procedure was carried out for networks with $N = 4 \ldots 40$. To reduce finite size (discretization and parity) effects, we also used continuous distributed patterns (cf [3]). We considered a normalized Gaussian distribution as well as patterns with elements evenly distributed in the interval $-1 \leqslant \xi_i^\mu \leqslant +1$ (box constraint) to examine the influence of the pattern distribution. Figures 1 and 2 show the dependence of this capacity on the network size as well as on the nature of the patterns. The error bars correspond to twice the mean deviation of the average value (statistical error). Sample size varied between 10 000 for small systems and 100 for $N = 40$. The ±1-patterns exhibit a strong parity effect which should, however, vanish in the thermodynamic limit. In figure 2, the values for the Gaussian patterns show a periodicity which is probably a result of the linear interpolation as the period of 6 corresponds to the passing of the zero-line of a stability value $\kappa^N(\alpha)$. (Remember that the critical capacity is approximately $\frac{5}{6}$ and α is restricted to rationals N/p.) Quadratic fits are given as a guideline to the eye (cf [4]).

There is no scaling theory for this problem, however, our numerical data suggest that the extrapolation to $N \to \infty$ could not be a linear one. A tentative quadratic extrapolation yields $\alpha_c = 0.834$ for Gaussian distributed patterns, $\alpha_c = 0.832$ for the box constraint and $\alpha_c = 0.827$ for ±1-patterns and N even. In the case of ±1-patterns and N odd, a quadratic fit is clearly unadmissible.

A second approach followed the procedure by Krauth and Opper [3] to determine $\kappa_\alpha(N)$ for different values of α and a subsequent extrapolation to $N \to \infty$. The capacity for Gaussian distributed patterns is determined as $\alpha_c = 0.833$.

We aimed to examine the possibilites and limits of combinatorial optimization when used for the training of autoassociative neural networks with binary couplings. The developed branch-and-bound algorithm allowed us to extend the exact investigation to systems with up to 40 neurons. We were not able to leave the region of strong finite size effects but could confirm theoretical [14] and numerical [3, 4] studies with additional numerical evidence. The possibility to determine all solutions of the learning problem also opens the way for an analysis of the space of solutions similar to the one already done for the ground states of the Ising spin glass model [10].

Acknowledgments

We are grateful to W Kinzel for drawing our attention to the problem, to T Klotz, A Hartwig and J Weißbarth for many helpful discussions and to M Opper for additional information. GM appreciated a scolarship of the Hans–Böckler–Stiftung.

References

[1] Horner H 1992 *Z. Phys.* **86** 291

[2] Patel H-K 1993 *Z. Phys.* **91** 257

[3] Krauth W and Opper M 1989 *J. Phys. A: Math. Gen.* **22** L519

[4] Derrida B, Griffits R B and Prügel-Bennet A 1991 *J. Phys. A: Math. Gen.* **24** 4907

[5] Uezu T and Nokura K 1996 *Progress of Theoretical Physics* **95** 273

[6] Köhler H 1990 *J. Phys. A: Math. Gen.* **23** L1265

[7] Fontani J F and Meir R 1991 *Network* **2** 353

[8] Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257

[9] Kobe S and Hartwig A 1978 *Comput. Phys. Commun.* **16** 1

[10] Klotz T and Kobe S 1994 *J. Phys. A: Math. Gen.* **27** L95

[11] Penney R W and Sherrington D 1993 *J. Phys. A: Math. Gen.* **26** 3995

[12] Wong K Y M and Sherrington D 1990 *J. Phys. A: Math. Gen.* **23** L175

[13] Bronstein I N and Semendjajew K A 1979 *Ergänzende Kapitel zum Taschenbuch der Mathematik* 6th edn (Stuttgart: Teubner)

[14] Krauth W and Mezard M 1989 *J. Physique* **50** 3057