Finite Size Scaling in Neural Networks

Walter Nadler*
Institut für Theoretische Chemie, Universität Tübingen, Auf der Morgenstelle 8, D-72076 Tübingen, Germany

Wolfgang Fink†
Institut für Theoretische Physik, Universität Tübingen, Auf der Morgenstelle 14, D-72076 Tübingen, Germany
(Received 25 July 1996)

We demonstrate that the fraction of pattern sets that can be stored in single- and hidden-layer perceptrons exhibits finite size scaling. This feature allows one to estimate the critical storage capacity α_c from simulations of relatively small systems. We illustrate this approach by determining α_c, together with the finite size scaling exponent ν, for storing Gaussian patterns in committee and parity machines with binary couplings and up to $K = 5$ hidden units. [S0031-9007(96)02100-X]

PACS numbers: 87.10.+e, 02.70.Lq, 05.50.+q, 64.60.Cn

Finite size scaling (FSS) has proven to be a powerful method for analyzing phase transitions, which occur rigorously only in the thermodynamic limit, using simulations of systems of finite size [1]. In particular, it has become the prime method for determining numerical values of critical coupling parameters and exponents [2].

Phase transitions are known to occur not only in condensed matter [3] and percolation systems [2] but also in random graphs [4], neural networks [5], and in algorithmic problems such as search [6], and the satisfiability of random boolean expressions [7]. Heuristic derivations of FSS rely on the divergence of a correlation length at a critical point in the infinite system [2,8]. However, Kirkpatrick and Selman [9] have demonstrated recently that FSS can also be used efficiently in problems without any intrinsic length scales, such as the connectivity of random graphs and the satisfiability of random boolean expressions. Abstract neural networks [5] are another class of systems without intrinsic length scale, and we will show in this paper that FSS occurs at the transition from storable to unstorable pattern set sizes, and that it provides a powerful computational method for determining critical storage capacities.

We will concentrate on the particular feed-forward networks of the perceptron class, namely, multilayer perceptrons with N input neurons, K hidden units, and a regular treelike connectivity ($N \bmod K = 0$), see Fig. 1, which are also known as *committee* and *parity* machines (CM, PM) with nonoverlapping receptive fields [10–12]. Input patterns ξ_{ik}, $k = 1, \ldots, K$, $i = 1, \ldots, N/K$, are processed by the following rules: The output of hidden layer cell k is given by

$$O_k = \operatorname{sgn}\left(\sum_{i=1}^{N/K} J_{ik}\xi_{ik}\right), \tag{1}$$

J_{ik} being the coupling between input cell ik and hidden unit k, while the final output is determined by

$$O = \operatorname{sgn}\left(\bigodot_{k=1}^{K} O_k\right), \tag{2}$$

where, in the case of a CM, the majority rule is implemented by $\bigodot \equiv \sum$, while, in the case of a PM, $\bigodot \equiv \prod$. A standard single-layer perceptron corresponds to $K = 1$. Since the majority rule is somewhat problematic in the case of even K, we will restrict ourselves here to CM with K odd.

A perceptron is able to store a particular set of input patterns $\{\xi_{ik}^{\mu}\}$, $\mu = 1, \ldots, p$, if there exists a coupling set $\{J_{ik}\}$ such that—under the action of Eqs. (1) and (2)—a prescribed set of outputs $\{O^{\mu}\}$ is generated. It is well known that for small values of $\alpha = p/N$ such a set of couplings can always be found, while for large enough α the probability for its existence vanishes. For finite systems the fraction of all possible input-output relations $\{(\xi_{ik}^{\mu}, O^{\mu})\}$ of relative size α that can be stored, which we will call $P(\alpha, N)$ [13], undergoes a smooth transition from one to zero. However, in the infinite system, it switches from one to zero at the critical storage capacity α_c.

This behavior, together with FSS, is nicely illustrated for the single-layer perceptron with continuous couplings and the ξ_{ik} drawn from a Gaussian distribution, where the exact solution for $P(\alpha, N)$ is known analytically [5,14],

$$P(p/N, N) = 2^{1-p} \sum_{i=0}^{N-1} \binom{p-1}{i}. \tag{3}$$

Figure 2 (top) shows $P(\alpha, N)$ for various values of N. The common intersection of these curves at $\alpha = 2$ is

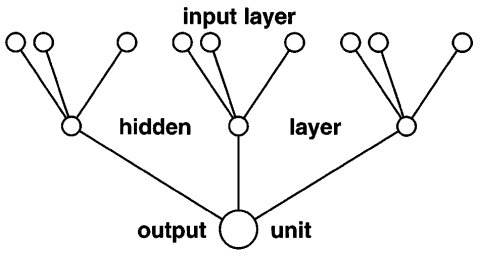

FIG. 1. Treelike multilayer perceptron with $K = 3$ hidden units.

0031-9007/97/78(3)/555(4)$10.00

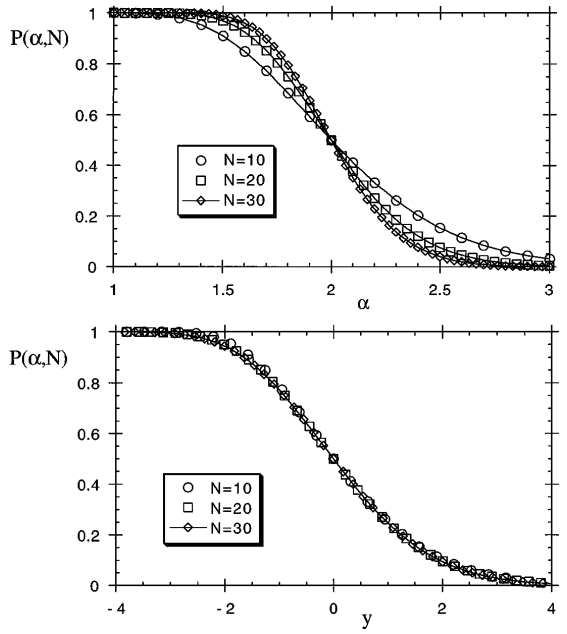

FIG. 2. Finite size scaling in the single-layer perceptron with continuous couplings: (top) Eq. (3), (bottom) finite size scaling as indicated in the text.

noticed immediately. Also, the steepness of the transition increases with system size N.

Under FSS, systems of different size behave in an identical way near the transition under a size-dependent rescaling of the control parameter [9],

$$y = (\alpha - \alpha_c)N^{1/\nu}. \tag{4}$$

Necessarily, the common intersection of the transition curves observed above corresponds to the critical storage capacity α_c. Figure 2 (bottom) shows that a rescaling with $\nu = 2$ and $\alpha_c = 2$ indeed lets all transition curves fall onto a single scaling curve. In this particular case, the numerical value of the FSS exponent ν, together with the analytic form of the scaling function,

$$f(y) = \frac{1}{2} + \frac{1}{2}\,\mathrm{erf}(-y/2), \tag{5}$$

can be derived from the asymptotic behavior of Eq. (3),

$$P(\alpha, N) \to \frac{1}{2} + \frac{1}{2}\,\mathrm{erf}\left(\sqrt{\frac{N}{2\alpha}}\,(2 - \alpha)\right). \tag{6}$$

Figure 2 demonstrates moreover that critical storage capacityα_c and FSS exponent ν can already be estimated from systems of relatively small size.

Simulations of neural networks are plagued by the problem that learning algorithms [5], necessary to determine coupling sets that solve the storage problem, are not guaranteed to reach a solution practically, i.e., under realistic time constraints, even if it exists. Close to α_c the average learning time diverges [15], a behavior reminding one of critical slowing down [3]. The situation is worse for systems with binary couplings, since there the usual learning algorithms are not applicable [16–19].

We will concentrate in the following on perceptrons with binary couplings $J_{ik} = \pm 1$, also known as Ising perceptrons. Employing complete enumeration of the couplings for systems up to size $N = 30$, simulation results *independent* of any learning algorithm are obtained. We used Gaussian patterns for the results presented in this contribution. Note that, for binary coupling perceptrons with a finite number of hidden units, information theory gives an upper limit for the critical storage capacity of one, i.e., $\alpha_c \leq 1$ [12,20].

Figure 3 (top) shows simulation results for $P(\alpha, N)$ for the case of a single-layer binary coupling perceptron. Sets of input-output relations were classified as storable or unstorable by complete enumeration of the coupling space [21]. Each data point was sampled with about 10^3 randomly chosen sets of input-output relations, giving a relative error of about 3%. As in the case of continuous couplings, Fig. 2, the curves for various system sizes intersect at the critical storage capacity, here with the numerical value $\alpha_c \approx 0.8$. Figure 3 (bottom) shows the same data under rescaling with Eq. (4) and $\nu \approx 1.7$. Again, all data points fall onto one scaling curve. Note that the value of the scaling function at the transition, $f(0) \approx 0.7$, is different from the continuous case [$f(0) = 0.5$].

Results for the hidden-layer systems of parity and committee type show a behavior qualitatively similar to the one presented in Fig. 3 for the single-layer perceptron. We have collected our results for various values of K in Table I. As is to be expected, α_c increases with the introduction of a hidden layer of neurons. The FSS exponent ν decreases with increasing K to about 1.3 and 1.2 for CM, and to values about one for PM.

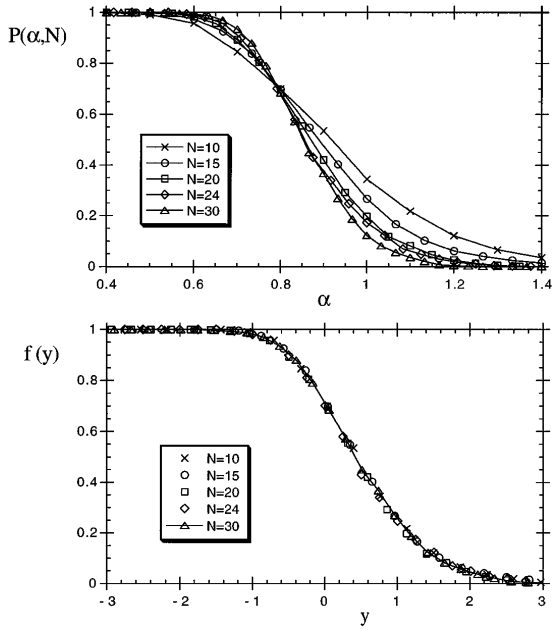

FIG. 3. Finite size scaling in the single-layer perceptron with binary couplings: (top) before, (bottom) after finite size scaling as indicated in the text.

TABLE I. Critical storage capacity α_c, finite size scaling exponent ν, and transition value $f(0)$ of scaling function f, for various binary perceptrons.

K	α_c (SD)	ν (SD)	$f(0)$ (SD)
	Single layer		
1	0.796 (0.010)	1.68 (0.09)	0.70 (0.03)
	Committee machine		
3	0.899 (0.008)	1.28 (0.06)	0.49 (0.03)
5	0.932 (0.012)	1.15 (0.08)	0.36 (0.04)
	Parity machine		
2	0.992 (0.005)	1.02 (0.04)	0.37 (0.02)
3	0.998 (0.005)	0.93 (0.03)	0.22 (0.02)
4	0.999 (0.008)	0.97 (0.04)	0.12 (0.02)
5	0.983 (0.009)	0.91 (0.04)	0.07 (0.01)

[a]In order to perform a reproducible and unambigious error analysis of the data, we used the *bootstrap* method [22]: About 10^3 bootstrap samples were drawn from the original data for all system sizes, and, for each such sample, α_c and ν were determined together by minimizing the mutual mean squared deviation of the interpolating scaling curves; the presented values and the estimated errors are the means and standard deviations, respectively, of α_c, ν, and $f(0)$ in the set of bootstrap samples.

The most surprising results are those for PM. Already a system with $K = 2$ hidden units exhibits a storage capacity extremely close to the theoretical limit, and Table I shows that there is practically no improvement in increasing K. In application situations, storing patterns has to be done using finite size perceptrons. Since the FSS scaling function $f(y)$ describes the asymptotic behavior of the fraction of storable patterns, $P(\alpha, N)$, around α_c, the critical capacity has to be considered together with $f(y)$ when assessing the quality of a particular system. Note that $f(y)$ decreases considerably with K in the critical region for PM as well as CM, see $f(0)$ in Table I. These features suggest that a PM with $K = 2$ is already the best *practical* binary perceptron for storing continuous patterns.

Simulation studies of the single-layer binary perceptron have been performed before for the problems of storing binary [16–19,23,24], and Gaussian patterns [23,25], using various approaches, and not always leading to conclusive results. Our result for α_c differs significantly from the analytical result of Ref. [26] ($\alpha_c = 0.833$) obtained using a first order replica symmetry breaking ansatz (RSB), but could be considered compatible—within error bars—with the simulation result of Ref. [25] ("$\alpha_c \approx 0.82$" [27]). This discrepancy between the analytical approximation and our simulation result suggests—provided finite size scaling holds—that the first order RSB is still insufficient for a correct analytical treatment of the $K = 1$ case, despite the claims in [26]. For binary CM and PM storing Gaussian patterns, no analytical or simulation results are available at present, to the best of our knowledge.

It has been hypothesized on the basis of replica studies [25] that the storage capacity for binary and Gaussian patterns is identical. Previous simulation results for $K = 1$ seemed to be compatible with this hypothesis and with the RSB result reported above ($\alpha_c = 0.83$ [16], $\alpha_c = 0.833$ [17–19], however [27]). Since our results differ significantly from the RSB result, this casts some doubt on either this hypothesis, the RSB result, or on the interpretation of the simulation results [27]. For the case of storing binary patterns in CM, simulation results using complete enumeration have been obtained for $K = 3$ in [12], together with analytical results for $K = 3$ ("$\alpha_c \approx 0.92$"), and for $K \to \infty$ ("$\alpha_c \approx 0.95$"), using a replica symmetric (RS) ansatz. Although our simulation results for CM differ somewhat, they can still be considered statistically compatible with those values, in contrast to the $K = 1$ case discussed above. This result supports the hypothesis of [12] that a RS ansatz might be sufficient for CM, and suggests that the hypothesis of an identical α_c for storing binary and Gaussian patterns might hold at least for CM.

In closing, we would like to again draw attention to the fact that the values for $f(0)$ differ strongly between various perceptrons. In particular, with the single exception of the CM with $K = 3$, they differ considerably from $1/2$. On the other hand, the relation $P(\alpha_{0.5}(N), N) = 0.5$ has often been the basis of an extrapolation to the infinite system critical parameter from simulations of finite systems [24,28]. If we define $y_{0.5}$ by $f(y_{0.5}) = 0.5$, then

$$\alpha_{0.5}(N) = \alpha_c + y_{0.5} N^{-1/\nu}. \tag{7}$$

Together with the fact that the FSS exponent ν deviates from one particularly for $K = 1$ and for CM, this feature emphasizes the need for an extrapolation nonlinear, instead of linear, in $1/N$ to correctly obtain the thermodynamic limit value of $\alpha_{0.5}(N)$ [29], and it may be the source of some problems encountered in earlier simulation studies [23,24].

The above results demonstrate that the FSS ansatz not only offers a new and powerful computational approach for evaluating the critical storage capacities of binary perceptrons, but also allows a detailed view on the storage properties in the critical region. We believe that it will prove valuable in analyzing the properties of a wide variety of binary perceptron topologies.

*Electronic address: walter.nadler@uni-tuebingen.de
†Electronic address: wolfgang.fink@uni-tuebingen.de

[1] *Finite Size Scaling and Numerical Simulations of Statistical Systems*, edited by V. Privman (World Scientific, Singapore, 1990).

[2] D. Stauffer and A. Aharony, *Introduction to Percolation Theory* (Taylor & Francis, London, 1992).
[3] *Phase Transitions and Critical Phenomena*, edited by C. Domb and M. S. Green (Academic, London, 1972–1976), Vols. 1–6; *ibid.*, edited by C. Domb and J. L. Lebowitz (Academic, London, 1983–1992), Vols. 7–15.
[4] E. M. Palmer, *Graphical Evolution* (Wiley, New York, 1985).
[5] J. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computation* (Addison-Wesley, Reading, MA, 1991).
[6] C. Williams and T. Hogg, Comput. Intell. **9**, 221 (1993).
[7] D. Mitchell, B. Selman, and H. J. Levesque, *Proceedings of the 10th International Conference on Artificial Intelligence* (AAAI Press, Menlo Park, CA, 1992), p. 459.
[8] S. Kirkpatrick and R. H. Swendsen, Commun. ACM **28**, 363 (1985).
[9] S. Kirkpatrick and B. Selman, Science **264**, 1297 (1994).
[10] G. J. Mitchison and R. N. Durbin, Biol. Cybernet. **60**, 345 (1989).
[11] E. Barkai, D. Hansel, and I. Kanter, Phys. Rev. Lett. **65**, 2312 (1990).
[12] E. Barkai, D. Hansel, and H. Sompolinsky, Phys. Rev. A **45**, 4146 (1992).
[13] We note that $P(\alpha, N)$—being the probability to find, in the ensemble of all sets of input-output relations of relative size α, a set of relations that can be stored in a perceptron of size N—differs considerably from the coupling space volume $V(\alpha, N)$ that is usually determined in analytical calculations, e.g., based on the replica approach (see, for example, Ref. [5]). The latter is a fluctuating quantity, and care has to be taken with its ensemble averaging [$\overline{\ln(V)}$ vs $\ln(\overline{V})$]. This does not apply to $P(\alpha, N)$. Moreover, $P(\alpha, N)$ is actually the more fundamental observable for the storage problem [5]. However, a lack of the ability to calculate it analytically for many interesting systems has led to some neglect of this important observable in the current literature.
[14] T. M. Cover, IEEE Trans. Electron. Comput. **14**, 326 (1965).
[15] A. Priel, M. Blatt, T. Grossmann, E. Domany, and I. Kanter, Phys. Rev. E **50**, 577 (1994).
[16] H. M. Köhler, J. Phys. A **23**, L1265 (1990).
[17] H. Horner, Z. Phys. B **86**, 291 (1992).
[18] H. Horner, Physica (Amsterdam) **200A**, 552 (1993).
[19] H.-K. Patel, Z. Phys. B **91**, 257 (1993).
[20] E. Gardner and B. Derrida, J. Phys. A **21**, 257 (1988).
[21] A more detailed account of the computational methods used will be given elsewhere; W. Fink and W. Nadler (to be published).
[22] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap* (Chapman and Hall, London, 1993).
[23] B. Derrida, R. B. Griffiths, and A. Prügel-Bennett, J. Phys. A **24**, 4907 (1991).
[24] I. Kanter and M. Shvartser, Physica (Amsterdam) **200A**, 670 (1993).
[25] W. Krauth and M. Opper, J. Phys. A **22**, L519 (1989).
[26] W. Krauth and M. Mézard, J. Phys. (France) **50**, 3057 (1989).
[27] Note that most previous simulation results are distinguished by a complete lack of an estimate for the error due to Monte Carlo sampling of the extrapolated value for α_c. On the other hand, the minuscule error reported in [19] (even smaller than that for several individual data points used in the extrapolation) is hard to believe. Note also that, when learning algorithms are employed in simulations, systematic errors may arise, and further analysis has to be supplied with some theory on its performance for large system sizes [17,18]. Otherwise, the choice of what system sizes to include in extrapolations becomes somewhat arbitrary (as in [16,19]).
[28] E. Eisenstein and I. Kanter, Europhys. Lett. **21**, 501 (1993).
[29] This problem was already noted in Ref. [9].