Convergence time and finite size effects in neural networks

1990 J. Phys. A: Math. Gen. 23 2237

(http://iopscience.iop.org/0305-4470/23/11/047)

J. Phys. A: Math. Gen. 23 (1990) 2237-2241. Printed in the UK

COMMENT

Convergence time and finite size effects in neural networks

G A Kohring
HLRZ, c/o KFA Jülich, D-5170 Jülich, West Germany

Received 1 March 1990

Abstract. The time needed for convergence to a stable state starting from an initial state having macroscopic overlap with a stored state is studied numerically for networks up to $N = 10^5$ neurons. Although such systems are much larger than those traditionally used in studying neural networks, they are still too small to extract reliable information about the asymptotic behaviour of the convergence time.

The body of neural network knowledge has grown very quickly during the past few years through a variety of analytical and numerical techniques. For strongly connected networks, analytical methods have been very successful at illuminating many of the static properties [1-4, 5], however, the dynamical properties have resisted analytic solutions and have mainly been studied via numerical means [6, 7]. The use of numerical methods requires careful attention to the effect the finite system size has on the quantities of interest. One of the widely studied dynamical problems is the determination of the critical overlap, or radius of attraction, and in that case Forrest [6] and Kanter and Sompolinsky [7] have shown how to account for the finite size of the network and extract meaningful information about the asymptotic, $N \to \infty$, limit from networks of only a few hundred to a few thousand neurons.

A closely related, though less well studied dynamical problem, is that of the time needed for convergence to one of the stored states starting from an overlap greater than the critical overlap [4, 8]. In this case, there is no simple prescription for dealing with the finite size effects and one must proceed with caution when dealing with 'small' systems. The first question to be answered is: how large is a small system?. Experience in calculating certain static properties of the Hopfield model has shown the necessity of considering system sizes up to $N \approx 3 \times 10^4$ before the large N behaviour could be reliably extracted [9]. Hence, there is a precedent for believing that systems $N \sim 10^3$ are too 'small' to be useful in predicting asymptotic behaviour of neural networks.

Recently Kanter [8] has studied this problem of convergence time in the Hopfield model via numerical simulations on networks of up to a few thousand neurons and drew some unexpected conclusions regarding the asymptotic behaviour of the average convergence time and the distribution of the convergence times. In this paper these extrapolations will be checked against system sizes up to $N = 10^5$ using a multispin coding algorithm [10] running on a Cray-YMP/832.

The numerical algorithm has been discussed in detail previously [9, 10] and here only a sketch of the essentials will be given. In the Hopfield model at zero temperature, the time evolution of the spins is given by the deterministic formula:

$$S_i(t+1) = \mathrm{sgn}(h_i(t)) \tag{1}$$

0305-4470/90/112237+05$03.50

where

$$h_i(t) = \sum_{j \neq i} J_{ij} S_j(t) = \frac{1}{N} \sum_{j \neq i} \sum_{\mu}^{P} \xi_i^\mu \xi_j^\mu S_j(t)$$

$$= \sum_{\mu} m^\mu(t) \xi_i^\mu - \frac{P}{N} S_i(t) = \frac{1}{N} H_i(t) - \frac{P}{N} S_i(t) \tag{2}$$

and $m^\mu(t) \equiv (1/N) \Sigma_j \xi_j^\mu S_j(t)$, $H(t) \equiv \Sigma_{j,\mu} \xi_j^\mu \xi_i^\mu S_j(t)$, and P is the number of stored states. The algorithm stores only the ξ_i^μ and S_i variables and not the coupling matrix J_{ij}. These variables are stored as one bit in an integer word or B spins per word on machines with B bits per word. The calculation of the overlaps is then carried out with logical operations acting on these integer variables:

$$Nm^\mu(t) = N - 2 \sum_{j=1}^{N/B} POPCNT(\zeta_j^\mu \otimes \sigma_j(t))$$

where ζ_j^μ is an integer word containing B spins from the state ξ^μ, $\sigma_j(t)$ is an integer word containing B spins from the state $S(t)$, $\otimes$ stands for the 'exclusive or' logical operator and *POPCNT* is a function which counts the number of bits set to one in the argument. To calculate the local fields, $H_i(t)$, one sums over μ in equation (3.2) by multiplying Nm^μ by $+1$ (-1) if the bit representing ξ_i^μ is 1 (0). Then, the spin S_i is updated by: $S_i(t+1) = (H_i(t) + NP)/(PN + P)$ if $S_i(t) = +1$, or $S_i(t+1) = (H_i(t) + NP)/(PN - P)$ if $S_i(t) = 0$.

When using parallel dynamics, the overlaps need only be calculated once at the beginning of the network update and then stored until used for calculating the local fields. In that way the algorithm is about as fast as a normal procedure which does not use multispin coding, but it can handle systems up to B times as large [9].

With the advantages of this method, the time needed for convergence to a stable state close to one of the stored states was studied on system sizes up to $N = 10^5$. In figure 1 is shown a histogram of the relative number of starting states needing T steps to reach such a stable state. Both curves are at $\alpha = P/N = 0.10$ and $m(0) = 0.40$. (This value of α should be large enough to avoid the spurious effects that occur as $\alpha \to 0$ and small enough to avoid the problems at $\alpha_c \approx 0.14$, the critical value of α.) The

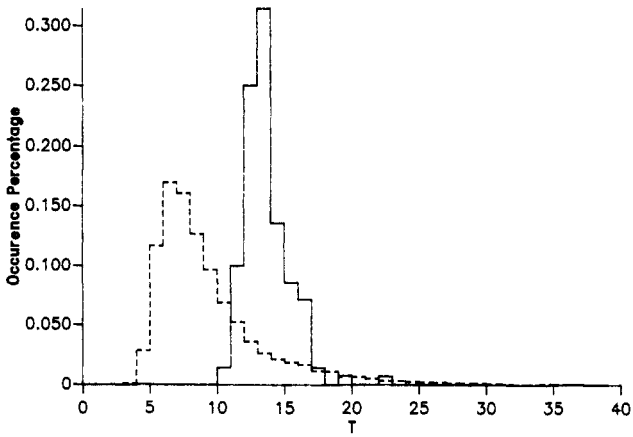

Figure 1. Histograms of convergence time at $\alpha = 0.100$ and $m(0) = 0.400$. The broken curve is for $N = 1088$ and the full curve is for $N = 64\,000$.

broken curve shows the histogram at $N = 1088$ and 12 500 initial states from 50 different sets of patterns. The full curve shows results at $N = 64\,000$ and only 200 initial states from 10 sets of patterns. In all cases the system was allowed to iterate until a stable state or cycle of length two was reached and only those stationary states having a large overlap with a stored state were recorded in the above histograms. (A detailed description of how to select the relevant stationary states can be found in [9].)

There is an obvious difference between the two histograms in that the curve for large N is narrower and higher than that for the smaller N. Figure 2 shows the variance of the distribution in convergence times, $\sigma = \langle T^2 \rangle - \langle T \rangle^2$, as a function of N. For the smallest N ($N = 192$) 10^5 different initial states from 10^3 different sets of patterns were used, while for the largest N ($N = 10^5$) only 100 initial states from 10 sets of patterns could be used with a moderate computational effort. Kanter's [8] calculations were carried out only up to a few thousand nodes, hence it is obvious from the figure how his unexpected conclusion that the width of the distribution increases with N was obtained. After increasing as N increases up to $N \sim 10^3$, the width begins to decrease with further increases in N. This decrease in the width is most probably due to a decrease in the number of metastable states in an annulus centred around a stored state as N increases, an effect which was seen in previous simulations of the Hopfield model [1, 9] and discovered analytically by Komlós and Paturi [4] and Gardner [5]. As these metastable states disappear, the energy landscape becomes 'smoother' and all paths leading to the stored state cross a similar 'terrain'.

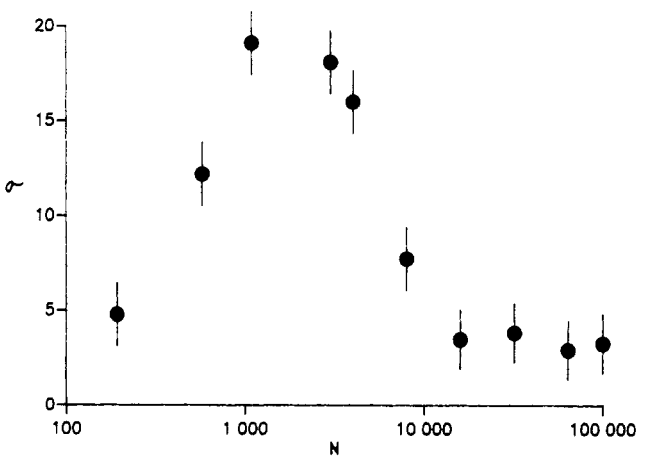

Figure 2. Variance of the distribution of learning times as a function of N for $\alpha = 0.100$, $m(0) = 0.400$.

The histograms in figure 1 also indicate that the average convergence time is increasing as a function of N. Figure 3 then shows a plot of the average convergence time as a function of $\ln(N)$ for several values of $m(0)$ in the range [0.350, 0.600]. The critical overlap found by Forrest in reference [6] lies inbetween $m(0) = 0.350$ and $m(0) = 0.375$ and is indicated in the plot by the difference in behaviour between the curves for these two values. Up to $N \sim 10^3$ the graphs again show the same behaviour as found by Kanter but, as in the case of the width of the distribution, this quite rapidly changes with increasing N. First, for $m(0) = 0.375$ and $m(0) = 0.400$ there is a clear deviation from the linear behaviour seen for small N. This again is due to the

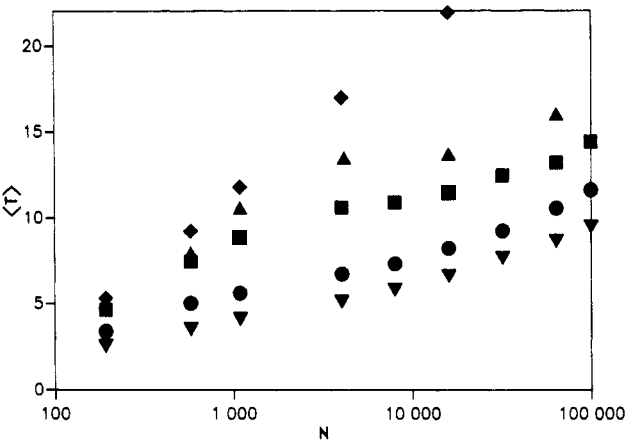

Figure 3. Average time to a fixed point as a function of N at $\alpha = 0.100$. (◆) $m(0) = 0.350$; (▲) $m(0) = 0.375$; (■) $m(0) = 0.400$; (●) $m(0) = 0.500$; (▼) $m(0) = 0.600$. Error bars are about the same size as the data points.

disappearance of intermediary metastable states as N increases leading to a smoother energy landscape. For $m(0) = 0.500$ and $m(0) = 0.600$, however, the initial states are closer to the stored state and thus less affected by the presence or absence of the metasatable states. The plot also shows that the curves for $m(0)$ above the critical overlap, are tending towards being parallel as N increases, although the possibility of convergence cannot be completely dismissed. In any case, the simple linear behaviour of $\langle T \rangle$ with $\ln(N)$ has been ruled out. Furthermore, system sizes up to $N = 10^5$ are clearly not sufficient for determining the asymptotic behaviour of the convergence time in the Hopfield model.

Komlés and Paturi [4] have attempted analytical calculations of the time needed for convergence, but are able to derive an exact result only in the limit $\alpha \to 0$. In this limit, (and as $N \to \infty$) an initial state needs $O(\ln \ln N)$ steps to converge to a stored pattern. The present simulations suggest that such asymptotic behaviour will be very difficult to confirm on present computers.

In summary, it has been demonstrated that the time needed for convergence to a stable state in the Hopfield model at finite values of α is not a simple logarithmic function of the number of neurons, N. Additionally, the width of the distribution of convergence times decreases for large N.

The results presented here bring into question the relevance of the limit $N \to \infty$ for biological systems. Although the human brain has a total of 10^{10} neurons, these neurons are structured into groups of the order of $10^4 \sim 10^5$ neurons [11]. Within these groups there is very high connectivity and it is most likely to such groups that the present model should be applied and not to the brain as a whole. Hence, even though 10^5 neurons is insufficient for determining asymptotic quantities, these asymptotic quantities may themselves be of little physical or biological relevance if they only appear at system sizes much, much larger than those of typical biological systems.

Acknowledgments

I would like to thank D Stauffer for many helpful conversations related to this work.

References

[1] Amit D, Gutfreund H and Sompolinsky H 1985 *Phys. Rev. Lett.* **55** 1530; 1987 *Ann. Phy.* **173** 30
Crisanti A, Amit D J and Gutfreund H 1986 *Euro. Phys. Lett.* **2** 337
[2] Gardner E 1987 *Euro. Phys. Lett.* **4** 481; 1988 *J. Phys. A: Math. Gen.* **21** 257
Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271
[3] Newman C M 1988 *Neural Networks* **1** 223
McEliece R J, Posner E C, Rodernich E R and Venkstesh S S 1987 *IEEE Trans. Inf. Th.* **33** 461
[4] Komlós J and Paturi R 1988 *Neural Networks* **1** 239
[5] Gardner E 1986 *J. Phys. A: Math. Gen.* **19** L1047
[6] Forrest B M 1988 *J. Phys. A: Math. Gen.* **21** 245
Krätzschmar J and Kohring G A 1990 *J. Physique* **51** 223
[7] Kanter I and Sompolinsky H 1987 *Phys. Rev.* A **35** 380
[8] Kanter I 1987 *Phys. Rev.* A **40** 2611; 1990 *Neural Networks and Spin Glasses: Proc. of the STATPHYS 17 Workshop*, ed W K Theumann and R Köberle (Singapore: World Scientific)
[9] Kohring G A 1990 *J. Stat. Phys.* **59** 1077
[10] Penna T J P and Oliveira P M C 1989 *J. Phys. A: Math. Gen.* **22** L719
[11] Eccles J C 1987 *The Understanding of the Brain* 2nd edn (New York: McGraw-Hill)