Convergence time in infinite-range neural networks with parallel dynamics at zero temperature

Ido Kanter
Joseph Henry Laboratories of Physics, Jadwin Hall, Princeton University, Princeton, New Jersey 08544
(Received 19 January 1989)

In simulations on the Little-Hopfield model it is found that the convergence time to a stable state close to one of the embedded patterns scales like $c(m_0)\log_{10}N$, where N is the size of the network and $c(m_0)$ depends on the initial macroscopic overlap m_0 with the pattern. In a related model known as the pseudoinverse model the convergence time to the pattern is much smaller than $\log_{10}N$. The results are compared with other possible pattern recognition methods.

During the last few years there has been extensive interest in the theory of neural networks that model associative memories. In such models the number of patterns one can store and retrieve is found to be $p=\alpha N$ ($\alpha\leq 2$), where N is the number of neurons (spins) and the number of synapses connecting neurons are of $O(N^2)$. Therefore the embedded information per synapse is of $O(1)$ which might be expected from an information theory point of view. Nevertheless, the quality of a system as an associative memory is a function of many parameters like the quality of the retrieval, basins of attraction, rate of convergence to the patterns, etc., beside the capacity per synapse.

In this paper we present some results of simulation on the convergence time to the embedded memories versus the size of the system, under parallel dynamics where $p\propto N$. The simulations were made on the fully connected Little-Hopfield[1,2] (LH) model and on the pseudoinverse (PI) model[3,4] at zero temperature. In this limit it is simpler to get better results, because the effect of thermal noise is eliminated. This limit may also have practical advantages if one wants to build pattern recognition devices.

In the thermodynamic limit and where $p\propto N$, unfortunately one can only calculate the time evolution of the macroscopic overlap analytically for the few first time steps.[5] The calculation of the macroscopic overlap for a large number of time steps becomes very complicated and is still an unsolved problem.

The models to be discussed here are governed by an energy

$$H=-\tfrac{1}{2}\sum_{i\neq j}^{N}J_{i,j}S_iS_j \tag{1}$$

where $S_i=\pm 1$ and N is the size of the network. The couplings, synaptic efficiencies, are constructed of the p given spin configurations (patterns) and given in the LH case by

$$J_{ij}=\frac{1}{N}\sum_{\mu=1}^{p}\xi_i^{\mu}\xi_j^{\mu} \tag{2}$$

and in the PI case by

$$J_{ij}=\frac{1}{N}\sum_{\mu,\nu=1}^{p}\xi_i^{\mu}(C^{-1})_{\mu\nu}\xi_j^{\nu} \tag{3}$$

where $\{\xi_i^{\mu}\}$ are quenched independent random variables which could take the values ± 1 with equal probability and stand for the ith bit in the μth pattern. Here, C^{-1} is the inverse of the overlap matrix C defined by

$$C_{\mu\nu}=\frac{1}{N}\sum_{i=1}^{N}\xi_i^{\mu}\xi_j^{\nu}\ . \tag{4}$$

Parallel dynamics at zero temperature means that the configuration at time step t is given by the rule

$$S_i(t)=\mathrm{sgn}\left[\sum_{j=1}^{N}J_{ij}S_j(t-1)\right] \tag{5}$$

where all the spins are updated at the same time and $\{S_i(0)\}$ is the initial configuration. The dynamical process evolves until $S_i(\tau)=S_i(\tau-1)$ for $i=1,2,\ldots,N$, and the lowest such τ is defined as the convergence time.

In the simulations the initial configuration was picked at random under the constraint that it has an overlap m_0 with a certain pattern. In order to get good enough statistics for each model and for each system size we used the following procedure: (1) choose n_s different samples, (2) in each sample we start with an initial overlap m_0 with each one of the patterns, and (3) in each such "valley" of one of the patterns we start with n_c different random initial configurations which have an overlap m_0 with the pattern. Therefore the number of measurements for each system size is pn_sn_c. It is obvious that it is better to make the simulations on pn_sn_c different samples. This is due to the fact that many measurements in the same sample and within the same "valley" are correlated. Nevertheless, we choose this way, especially for large systems, in order to limit the slowest computations, which are the calculation of the synaptic efficiencies for each sample.

In Figs. 1 and 2 the results for the LH case are presented, where $\alpha=0.1$ and $m_0=0.4$ and $m_0=0.6,0.75$, respectively. For $N=50-100$, $n_s=100-150$ and $n_c=100-150$; for $N=300-500$, $n_s=20-30$ and $n_c=2-4$; and for $N\geq 1000$, $n_s=3-5$ and $n_c=1-2$. We would like to stress that in both of these figures the ferromagnetic point, $p=1$ (or in our cases $N=1/\alpha=10$), is known analytically without any errors. The convergence time in

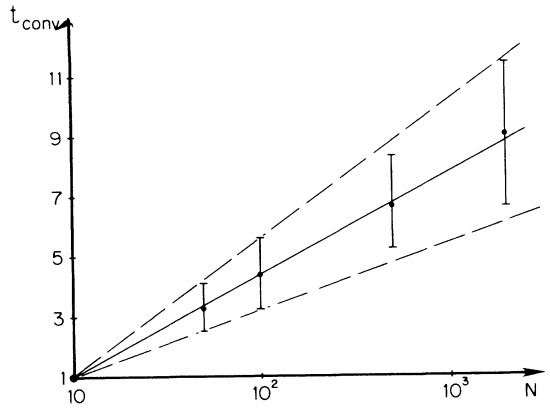

FIG. 1. The time convergence vs $\log_{10}N$ for the LH model, where $\alpha=0.1$ and $m_0=0.4$. The dots indicate the average convergence time, and the dashed lines indicate that the width of the distribution of the convergence time grows also as $\log_{10}N$.

this case is exactly 1, for the reason that after one step one knows undoubtedly that the system is in a stable state without any need to check it.

In these simulations α was fixed to be 0.1 in order to be as far as one can in the region of a finite α, but not too close to the maximum capacity in the thermodynamic limit $\alpha_c \simeq 0.14$,[6] above which the fraction of errors in the retrieval increases dramatically. Nevertheless, in finite systems, the transition as a function of α and the overlap of the retrieval state with the pattern are not sharp functions as they are in the thermodynamic limit. Therefore in the simulations we decided to count only cases where the final overlap was greater than 0.9. We also checked that the fraction of cases, among pn_sn_c measurements, which converge to a stable state with respect to a single spin flip and which have a magnetization greater than 0.9 is an increasing function of the size of the system (cyclic flows with an average magnetization greater than 0.9 are very rare). This is an indication that m_0 is within the basin of attraction, which is true only for $m_0 \geq 0.4$ in our system size.

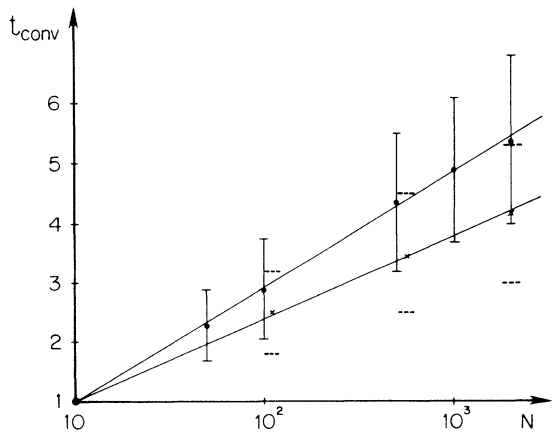

FIG. 2. The time convergence vs $\log_{10}N$ for the LH model, where $\alpha=0.1$. The dots stand for the average convergence time for $m_0=0.6$. The $\times$ signs and the dashed signs stand for the average convergence time and the width of the distribution of the convergence time for $m_0=0.75$, respectively.

The simulations suggest the following results. (1) The *average* convergence time to a stable state close to the pattern scales with $\log_{10}N$, which can be seen even in small systems. More precisely, the convergence time is fitted by $t_{\text{conv}}=c(m_0)(\log_{10}N-1)+1$, where the slope $c(m_0)$ is a constant which depends only on m_0. (2) The slope of the average convergence time versus $\log_{10}N$ is a decreasing function of m_0. The slope $c(m_0)$ for the average convergence time is given by 0, 1.4, 2.1, and 3.35 for $m_0=1$, 0.75, 0.6, and 0.4, respectively. It seems that at least for this range of m_0, a good approximation for $c(m_0)$ is given by $c(m_0)\propto 1-m_0$. (3) The width of the distribution of the convergence time also seems to scale with $\log_{10}N$ as shown in Fig. 1.

The result that $t_{\text{conv}} \propto \log_{10}N$ can be compared with information theory. Basically, the task in such pattern recognition systems is to identify (flow to) the pattern which has the maximal overlap with the initial configuration. Let us try to compare these results with other possible methods. The calculations of the overlaps of the p patterns with the initial state can be done in parallel dynamics (or using continuous neurons) even in $O(1)$ time steps. The search of the maximal overlap among the p overlaps using parallel dynamics on a binary tree, for instance, takes $c\ln N$ time steps where $c=1/\log_{10}2$ (see also Ref. 7). Therefore one can do the same task by other methods with time proportional to $\ln N$, but with a bigger coefficient in comparison to the average $c(m_0)$ for $m_0 \geq 0.4$ (at least for the parallel binary search). Nevertheless, from a practical point of view one can terminate the dynamical process after a finite number of steps which ensure an overlap close enough to 1.

The result that the slope $c(m_0)$ is a function of the initial configuration is surprising, because the magnetization for $m_0=0.6$, for instance, after one step and in the thermodynamic limit is greater than 0.75.[5] Therefore it seems natural that the asymptotic behavior of $c(m_0)$ should be independent of the initial overlap. However, from Ref. 5 one can verify that the magnetization in each time step is not only a function of the magnetization in the previous step, but is a function of the magnetizations and the correlations among the configurations in all the previous time steps. Furthermore, in the thermodynamic limit one can explicitly verify from Ref. 5 that $m(1)$ as a function of m_0 is greater than $m(2)$ as a function of $m(1)$ for $m(1)=m_0$, where $m(t)$ is the average magnetization at time t. This result might indicate that also for a large number of time steps the convergence time is slower as m_0 is smaller, which is in agreement with our results.

The result that the broadening of the convergence time distribution scales like the average convergence time is also a surprising result. In order to explain such a behavior, let us consider for simplicity the following convergence process. Assume $\Delta(t)\equiv m(t)-m(t-1)=\epsilon$, then $\Delta(t+1)=\gamma\epsilon$, where $\gamma<1$. The process is terminated when $\Delta(t)<1/N$, yielding a convergence time of $O(\ln N)$. For a fixed γ the distribution of the convergence time is

of $O(1)$, and for a random $\gamma(t)$ with a finite width the distribution is of $O(\sqrt{\ln N})$. Nevertheless, a possible source for a distribution of width of $O(\ln N)$ is that $m(t)$ is a function of $\{m(t_i)\}$ and $\{q(t_i,t_j)\}$, where $t_i,t_j<t$ and $q(t_i,t_j)$ is the overlap between the configurations at t_i and t_j.[5] For instance, one can verify analytically that $dm(1)/dm_0 \neq dm(2)/dm_0$ for $m(1)=m(2)$. Therefore our process is a random process with a long memory and one would expect that also the average $\gamma(t)$ depends on the history of the system. [An example is a process where the average $\gamma(t)$ increases as a function of the average γ in the past.] Another possibility is that this result might in fact be $O(\sqrt{\ln N})$ in larger systems or with a better statistics even in our large systems.

The results for the PI case for $\alpha=0.1$ and $m_0=0.3$ are presented in Fig. 3. Here, the basin of attraction is larger[3] than in the LH case and enables a smaller m_0 to be used, which gives relatively a larger convergence time. In this case, only stable states with an overlap which is equal to 1 are included, for any system size.[3] These results show that in the PI case the convergence time scale is much smaller than $O(\ln N)$, and maybe converges in the thermodynamic limit even to a constant. Furthermore, the width of the distribution of the convergence time does not scale with $\ln N$.

These results could be understood by the fact that in a configuration which is one of the patterns, the local field on each spin is equal to $1-J_{ii}\simeq 1-\alpha$.[3] This result is in contrast to the behavior in the LH case, where there are many spins with small negative and positive local fields.[6] Furthermore, even far away from the pattern the average overlap, within the replica symmetric approximation,[3] is given by $m=\tanh(Jm/T)$, where T is the temperature, m is the only macroscopic overlap, and J is given explicitly in Ref. 3. This result is very similar to the result in the ferromagnetic case. Therefore one would expect that the behavior of the system would be similar to a ferromagnetic system, and hence the convergence time is much faster.

This result immediately raises the question of whether one can build other pattern recognition systems, for instance a layered network, which can solve the task with

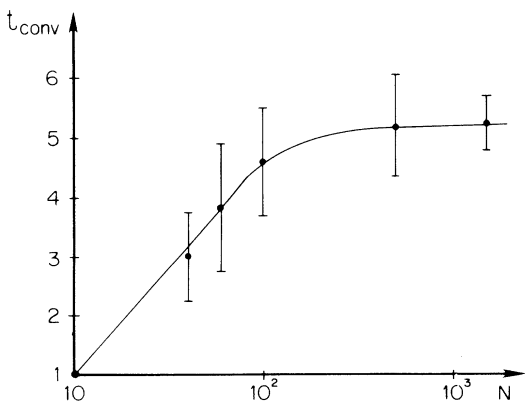

FIG. 3. The time convergence vs $\log_{10}N$ for the PI model, where $\alpha=0.1$ and $m_0=0.3$. The dots indicate the average convergence time.

$O(1)$ time steps. The answer to this question is affirmative, and let us now briefly describe the structure of such a layered network. The first layer contains p neurons $\{S_l^1\}$ taking continuous values, such that $S_l^1=m_l$, where m_l is the overlap of the input with the lth pattern. In the next layer there are $p(p-1)$ two-state neurons $\{S_{lp+k}^2\}$, where $l=1,2,\ldots,p$, $k=1,2,\ldots,p-1$, and $S_{lp+k}^2=\Theta(m_l-m_k)$, where $\Theta(x)$ is the Heaviside step function. In the third layer there are again p continuous neurons such that $S_l^3=\sum_{k\,(\neq l)}S_{lp+k}^2$. The pattern with the maximal overlap is l if and only if $S_l^3=p-1$. The weak point of this method is that the number of neurons and synapses scale like N^2 in the case where $p\propto N$. Nevertheless, in cases where there is only one macroscopic overlap, or in cases where the difference between the maximal overlap and the second maximal overlap is greater than the differences among the other overlaps, one can easily use macroscopic thresholds in the second layer such that it is necessary to have only $O(N)$ neurons. For instance, in the second layer there are pq neurons, where q is of $O(1)$ and $S_{lq+k}^2=\Theta(m_l-m_k-\epsilon)$, where ϵ is smaller than the difference between the largest and the second largest overlap. It is obvious that the maximal overlap is l if $S_l^3=q$. The constraint of one strongly dominant overlap is not artificial, since otherwise in the models studied here the retrieval is not ensured. It is also important to stress that this system is much better than a simple two-layer grandmother-cell circuit,[8,9] since the important parameter is the differences among the overlaps and not the value of the largest overlap.

Since not only the convergence time plays a role in the analysis of parallel algorithms, but also the size (number of neurons), one approach to estimating the quality of a parallel algorithm is the product of time by size.[10] It is important to note that the best product measure cannot be any less than the lower bound time for the same sequential solution. The reason is that it is possible to sequentialize any parallel dynamics. On one hand, the result for the PI system is reasonable, due to the fact that the product of time by size seems to scale like $O(N)$. This order of complexity is the lower bound for a search of the maximal number among N (disordered) numbers. On the other hand, in the studied models each neuron is a processor of 1 bit, where in other methods each processor is of $O(\ln N)$ bits [to represent and to compare overlaps in the range $(-N,N)$]. However, our neural network systems are a shared memory system, which means that the same data (spin configuration) are shared by all the neurons. For instance, *all the neurons* in the next time step *read* the information of the ith neuron from the previous time step, and *all the neurons* together decide what to *write* in the ith neuron in the next time step. Furthermore, each neuron is not a processor which can make a comparison between two numbers, but can sum N numbers in each time step. These two concepts of unrestricted range of shared memory and processors each one of which can sum $O(N)$ numbers in each time step give a strong computational power together with a simple circuit structure to the neural network systems. In order to achieve these goals it is necessary to have some analog devices with nonlinear characterizations, like analog neu-

rons in layered networks or analog membranes (to sum the input) in the discussed models. One of the main future goals is to understand the computational power of systems based on these concepts to solve more complicated problems. It seems that one can gain a logarithmic time [=(time)×(size)] in comparison to the usual sequential or parallel dynamics. The possibility of gaining higher orders of computational time in more complicated problems certainly deserves further study, and may help to understand the advantages of such dynamics in biology in correspondence with information theory.

Simulations on larger systems and with better statistics or analytical progress are necessary in order to verify our results.

Finally, we would like to mention that the limit of a finite connectivity, where each neuron receives a finite number of inputs, seems to be a limit where one can simulate much larger systems. However, from some simulations that we made it seems that the fluctuations are much larger in this limit. Nevertheless, a convergence time of $O(\ln N)$ seems to be the natural time scale, due to the fact that almost all loops in the system are of $O(\ln N)$.[11] Therefore it takes $O(\ln N)$ steps for a local event to influence the whole system. We would like also to stress that in asymmetric systems[11] with finite connectivity one can analytically prove that the convergence time is of $O(\ln N)$.[12]

Note added. After this work was completed, I received a copy of unpublished work by Komlos and Paturi which proves that in the thermodynamic limit of the LH model and in the case of a perfect retrieval [$\alpha \neq O(1)$] the convergence time under parallel dynamics is $O(\ln \ln N)$. Furthermore, in the case of a finite α after $\ln(1/\alpha)$ parallel steps the overlap is greater than or equal to $1-\exp(-1/4m_0)$ and remains within this distance. Therefore this work proves that, in the thermodynamic limit and for large enough initial macroscopic overlap, in practice one can terminate the dynamical process after a small number of steps.

I would like to thank especially D. S. Fisher for numerous stimulating discussions and his critical reading of the manuscript. I thank C. M. Newman for sending me a copy of unpublished work by Komlos and Paturi and P. W. Anderson, E. B. Baum, C. Doty, and H. Sompolinsky for helpful conversations. I thank AT&T Bell Laboratories at Murray Hill for the hospitality and the opportunity to use their computer facilities during the time most of this work was being done. This work is supported by the Weizmann Foundation and also in part by the National Science Foundation Grant No. DMR 8719523.

[1]W. A. Little, Math. Biosci. **19**, 101 (1974).
[2]J. J. Hopfield, Proc. Natl. Acad. Sci. USA **79**, 2554 (1982); **81**, 3088 (1984).
[3]I. Kanter and H. Sompolinsky, Phys. Rev. A **35**, 380 (1987).
[4]L. Personnaz, I. Guyon, and G. Dreyfus, J. Phys. (Paris) Lett. **46**, L359 (1985).
[5]E. Gardner, B. Derrida, and P. Mottishow, J. Phys. (Paris) **48**, 741 (1987).
[6]D. J. Amit, H. Gutfreund, and H. Sompolinsky, Phys. Rev. Lett. **55**, 1530 (1985).
[7]E. Domany and H. Orland, Phys. Lett. A **125**, 32 (1987).
[8]D. W. Tank and J. J. Hopfield, Proc. Natl. Acad. Sci. USA **84**, 1896 (1987).
[9]E. B. Baum, J. Moody, and F. Wilczek, Biol. Cybern. **59**, 217 (1988).
[10]D. Harel, *Algorithmics: The Spirit of Computing* (Addison-Wesley, Reading, MA., 1987).
[11]B. Derrida, E. Gardner, and A. Zippelius, Europhys. Lett. **4**, 167 (1987).
[12]I. Kanter (unpublished).