ESANN'1994 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 20-21-22 April 1994, D-Facto public., ISBN 2-9600049-1-4, pp. 7-12

Stability and bifurcation in an autoassociative memory model

W.G. Gibson, J. Robinson, C.M. Thomas,

School of Mathematics and Statistics
University of Sydney
N.S.W. 2006
Australia

Abstract The stability and bifurcation structure of a theoretical autoassociative network is studied. The network consists of randomly connected excitatory neurons, together with an inhibitory interneuron that sets their thresholds; both the degree of connectivity between the neurons and the level of firing in the stored memories can be set arbitrarily. The network dynamics are contained in a set of four coupled difference equations. Their equilibrium properties are investigated, analytically in certain limiting cases and numerically in the general case. The regions of parameter space corresponding to stable and unstable behaviour are mapped, and it is shown that for suitable parameter choices the network possesses stable fixed points which correspond to memory retrieval.

1. Introduction

Most recent memory network research has been dominated by models based on analogies with statistical mechanical systems [1]-[3]. This has provided powerful tools for the study of a somewhat restricted class of networks. The basic model [1] is very far from the biology, involving such assumptions as symmetric coupling between 'neurons' which can be simultaneously excitatory and inhibitory, high (50%) firing activity in memory states, artificially set firing thresholds, etc. Although considerable work has been done in relaxing these assumptions (e.g., [4]) the models are still sufficiently unrealistic as to make comparison with biological systems difficult. Thus it is worthwhile returning to earlier models [5], [6] and analysing them more rigorously. These models take their starting point from an actual biological network, and analysis proceeds by conventional statistical and probabilistic methods, not by analogy with another system.

Our basic network consists of randomly connected excitatory neurons, together with an inhibitory interneuron that sets their thresholds. Both the degree of connectivity between the neurons and the level of firing in the stored memories can be set arbitrarily. The memories are stored via a two-valued Hebbian, and evolution from an arbitrary initial state is by discrete, synchronous steps. The theory [7] takes into account both spatial correlations between the learned connection strengths and temporal correlations between the state of the system and these connection strengths. (These correlations were neglected in the earlier theories mentioned above.) Our theory has been applied to the CA3 region of the hippocampus, using parameter values based as far as possible on the known physiology of the CA3 region [8].

The theory leads to a set of four coupled difference equations which describe

ESANN'1994 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 20-21-22 April 1994, D-Facto public., ISBN 2-9600049-1-4, pp. 7-12

the full dynamics of the recall process. For a given set of network parameters, it is easy to simply iterate these equations and hence follow the timecourse of the system. However, it is desirable to understand the network behaviour at a deeper level, and in particular to investigate the conditions under which the memories are stable fixed points of the dynamical system and also to analyze the bifurcations which can occur when parameters are varied. Since our network involves non-symmetric connectivity, it is not possible to use Lyapunov functions; instead, we have used a combination of conventional nonlinear analysis, coupled with numerical methods (in particular, AUTO86) to provide a comprehensive picture.

In this presentation, we limit our analysis to two cases. The first is that of a randomly connected network which does not store memories. In this case, it is possible to give an almost complete picture of the stability and bifurcation structure by purely analytic means. For the case of a network storing memories, the analysis becomes much more complex and we concentrate on the case of a large network storing a limited number of memories; a number of analytic results can be obtained if we further work in an approximation which neglects temporal correlations - numerical work can then be used to demonstrate that the addition of these temporal correlations does not cause a drastic change in the qualitative behaviour of the system, although the precise quantitative details may change.

2. Network

The basic network is shown in Figure 1. The total connection strength from neuron j to neuron i is $W_{ij}J_{ij}$ where W_{ij} is the *intrinsic* contribution given by $P(W_{ij}=1)=c$ and J_{ij} is the *learned* contribution given by the clipped Hebbian prescription $J_{ij} = \max\{Z_i^p Z_j^p : p = 0, 1, \cdots, m\}$ where $\mathbf{Z}^p$ are the memory vectors, with elements 1 or 0 according to $P(Z_i^p=1)=a$.

In the recall of a typical stored pattern, which without loss of generality can be taken to be $\mathbf{Z}^0$, the network starts from state $\mathbf{X}(0)$ (which typically is a random distortion of $\mathbf{Z}^0$) and updates synchronously to generate a sequence $\mathbf{X}(t)$, $t = 0, 1, \ldots$, according to $X_i(t+1) = H(h_i(t) - T(t))$. Here, $H(\cdot)$ is the unit step function, $nh_i(t) = \sum_{j=1}^{n} W_{ij}J_{ij}X_j(t)$ is the input to the ith neuron and $T(t)$ is the threshold which is set by the total activity in the network according to $T(t) = g_0 + g_1 \sum X_i(t)/n$ where g_0, g_1 are constants.

At any point in the recall process, the state of the network is characterised by the quantities x_t and y_t, where x_t is the fraction of neurons in the target memory $\mathbf{Z}^0$ which are active at time t (the "valid" firings) and y_t is the corresponding quantity for all active neurons outside $\mathbf{Z}^0$ (and thus measures the "spurious" firings). For exact recall, we require $x_t \to 1$ and $y_t \to 0$ as $t \to \infty$.

3. Theory

The theory has been developed at three levels [7]. *Level 0* neglects all correlations (and in this respect is equivalent to earlier theories [5], [6]). *Level 1* includes the correlations between the learned connection strengths J_{ij} which arise because the J_{ij}'s are computed from a common set of memories $\{\mathbf{Z}^p\}$. *Level 2* also

ESANN'1994 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 20-21-22 April 1994, D-Facto public., ISBN 2-9600049-1-4, pp. 7-12

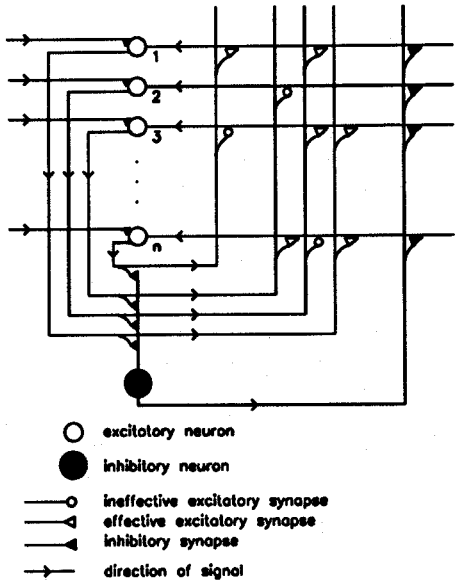

Fig. 1. The basic network, consisting of n excitatory neurons (open circles) and one inhibitory interneuron (filled circle). The excitatory neurons make random connections with each other, the probability of any one connection existing being c. Before learning, these connections are ineffective; after learning, a subset of them becomes effective and in the final state of the network there are excitatory synaptic connections whose strengths are taken to be unity (open triangles) and others whose strengths have remained at zero (open circles). The inhibitory interneuron receives input from all the active excitatory neurons, and in turn sends an inhibitory signal to each of them; no learning occurs here, and all synaptic strengths are fixed. The initial state of the system is set by a firing pattern coming onto the excitatory neurons from some external source, shown by the lines entering from the left. Once the initial state has been set, the external source is removed. The network then updates its internal state cyclically and synchronously.

includes the correlations which build up between the state vector $\mathbf{X}(t)$ and the J_{ij}'s, due to the progressive recall process. The full theory is given in [7]; here we give only the Level 1 theory, which involves two coupled difference equations:

$$x_{t+1} = \Phi\left(\frac{E_1(t)}{\sigma_1(t)}\right), \quad y_{t+1} = \Phi\left(\frac{E_n(t)}{\sigma_n(t)}\right)$$

where $\Phi(\cdot)$ is the normal distribution function and x_t and y_t are the average firing levels, for correct and spurious cells respectively, at time-step t. All the quantities on the right-hand sides can be expressed in terms of x_t and y_t. The expectations are

$$\begin{aligned} E_1(t) &= ac x_t + (1-a)c\rho y_t - ET(t) \\ E_n(t) &= ac\rho x_t + (1-a)c\rho y_t - ET(t), \end{aligned}$$

ESANN'1994 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 20-21-22 April 1994, D-Facto public., ISBN 2-9600049-1-4, pp. 7-12

where $\rho = 1 - (1-a^2)^m$ and

$$ET(t) = g_0 + g_1 a x_t + g_1(1-a)y_t,$$

and the standard deviations $\sigma_i(t)$ are given by somewhat more complicated expressions [7]. The Level 2 approximation is similar to the above, but introduces two further (non-observable) quantities x'_t and y'_t and consists of a set of 4 coupled equations [7].

4. No memories

We commence our stability analysis by treating the case where all $J_{ij} = 1$, so the system stores no memories. However, the neurons are still randomly connected through W_{ij}, and we can show that parameter space can be divided into three regions: ***extinction***, where the only stable fixed point corresponds to all firing activity zero; ***stable***, where there is a stable fixed point with non-zero activity; ***unstable***, where there can be a variety of behaviours, ranging from orbits of period 2 to chaos.

5. Memories

In the case of a network storing memories, it is only possible to do a reasonably full analysis of the equations in certain simplified cases. One such case is the Level 1 equations for a network containing a large number of neurons. Taking the limit $n \to \infty$ (and also for simplicity setting $g_0 = 0$), the Level 1 theory reduces to a ***single*** equation for the equilibrium number of valid firings ($x = \lim x_t$ as $t \to \infty$):

$$x = \Phi(\lambda x + \mu),$$

where

$$\mu = \frac{c\rho - g_1}{\sqrt{\gamma}c}, \qquad \lambda = \frac{(c-g_1)}{\sqrt{\gamma}c}\frac{a}{(1-a)}\frac{1}{\Phi(\mu)},$$

and γ is a known function of m and a. This equation is formally similar to one studied by Amari [9], so a parallel analysis can be applied. (Note, however, that although the equations are formally similar, the interpretation of the results is different, since Amari's network does not store memories.) Some results are summarized in Figure 2, where the bifurcation lines are given by

$$\mu_\pm = \pm\sqrt{\log(\lambda^2/2\pi)} - \lambda\Phi\left(\pm\sqrt{\log(\lambda^2/2\pi)}\right).$$

Memory recall can only occur for parameter values in the hatched region.

Some further insight into the recall region can be obtained by plotting μ and x as functions of the inhibition strength parameter g_1. Figure 3(a) shows the appearance and subsequent disappearance of a bistable region (two stable and one unstable branches) as g_1 is increased; the upper branch corresponds to memory retrieval (the spurious activity, y, is small throughout this region). Figure 3(b) relates this to the behaviour of μ, with bistability for $\mu_+ < \mu < \mu_-$.

ESANN'1994 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 20-21-22 April 1994, D-Facto public., ISBN 2-9600049-1-4, pp. 7-12

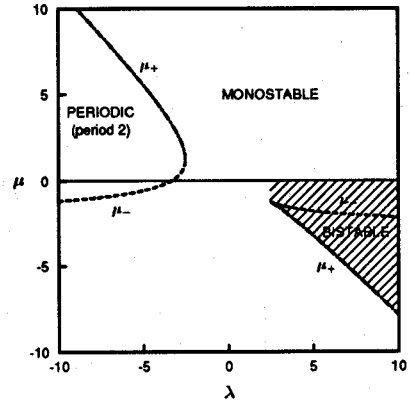

Fig. 2. The stability regions for a network storing memories. The analysis is based on the Level 1 equations in the limit as the number of neurons n becomes large. The intrinsic threshold g_0 is taken to be zero, and the quantities λ and μ are not free to vary, but are related to the remaining parameters in the model as given in the text. The regions are: MONOSTABLE, in which x (the average valid firing activity in a memory) has only one stable fixed point; BISTABLE, where x has two stable fixed points; PERIODIC, where x has a stable cycle of period 2 and there are no other attractors. Memory recall can only occur in the hatched region $\lambda > \sqrt{2\pi}$, $\mu_+ < \mu < 0$.

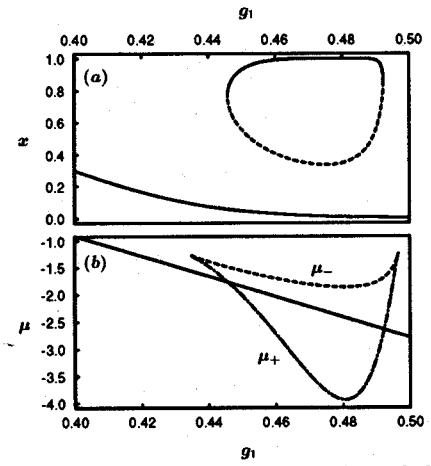

Fig. 3. The behaviour of the fixed points as a function of the inhibition strength parameter g_1 for the case $m = 100$, $a = 0.11$, $c = 0.5$, according to the Level 1 theory in the large-n limit. The upper figure (a), produced using AUTO86, shows the locus of the fixed points of x, with the solid lines corresponding to stable fixed points and the broken line to unstable ones. The upper solid line represents memory retrieval. The lower figure, (b), shows μ, along with μ_- and μ_+, on the same g_1-scale. Comparison of (a) and (b) shows that a change from monostable to bistable behaviour occurs when μ enters the region between μ_- and μ_+.

ESANN'1994 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 20-21-22 April 1994, D-Facto public., ISBN 2-9600049-1-4, pp. 7-12

6. Conclusion

We have investigated the stability properties of a particular autoassociative neural network which incorporates a number of biologically realistic features. The network dynamics is contained in a set of coupled difference equations which govern the time evolution of the firing activity. For the network to store and retrieve memories, it is necessary to establish that in the long-time limit these equations possess equilibrium solutions which are stable fixed points with firing activity close to that of the stored memory patterns. This has been achieved analytically in the limiting case of an infinitely large network storing a finite number of memories, if temporal correlations are neglected. This analysis also indicates that, because of uniform convergence, a large network will have a behaviour which is similar to that of the infinite one. Numerical work has confirmed this, and has also shown that the stability and bifurcation structure is similar when temporal correlations are included.

References

1 J.J. Hopfield: Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Sciences, U.S.A., 79, 2554-2558, 1982.

2 D.J. Amit: Modeling brain function: the world of attractor neural networks, Cambridge U.P., Cambridge, 1989.

3 J. Hertz, A. Krogh, R.G. Palmer: Introduction to the theory of neural computation, Addison-Wesley, New York, 1991.

4 D. Golomb, N. Rubin, H. Sompolinsky: Willshaw model: Associative memory with sparse coding and low firing rates. Physical Review A, 41, 1843-1854, 1990.

5 D. Marr: Simple memory: a theory for archicortex. Philosophical Transactions of the Royal Society of London B, 262, 23-81, 1971.

6 A.R. Gardner-Medwin: The recall of events through the learning of associations between their parts. Proceedings of the Royal Society of London B, 194, 375-402, 1976.

7 W.G. Gibson, J. Robinson: Statistical analysis of the dynamics of a sparse associative memory. Neural Networks, 5, 645-661, 1992.

8 M.R. Bennett, W.G. Gibson, J. Robinson: Dynamics of the CA3 pyramidal neuron autoassociative memory network in the hippocampus. Philosophical Transactions of the Royal Society of London B (in press).

9 S. Amari: Characteristics of randomly connected threshold-element networks and network systems. Proceedings of the IEEE, 59, 35-47, 1971.

Acknowledgments

We wish to thank Professor M.R. Bennett for many useful discussions. Support under ARC Grant AC9031997 is acknowledged.