Neural Network Models: Their Theoretical Capabilities and Relevance to Biology

Martin Evans

Submitted for the degree of
Doctor of Philosophy

Department of Physics
University of Edinburgh
1990

To the memory of David Joseph Evans.

Declaration

The work presented in this thesis is my own and is a result of the following collaborations and projects.

The work of chapter 2 was carried out in collaboration with Prof D.J.Wallace and C.Zhan. It has been submitted for publication as

"Competition between Symmetry Transform and Hopfield Interactions in a Neural Net"

M.R.Evans, D.J.Wallace and C.Zhan to *J. Phys.A.*

The work of chapter 3 was carried out in collaboration with Prof D.J.Amit, Prof H.Horner and Dr K.Y.M.Wong and the results have appeared in

D.J. Amit, M.R. Evans, H.Horner and K.Y.M. Wong, *J.Phys.A* **23** 3361 (1990).

Some of the results of sections 4.1 – 4.7 have appeared in

M.R. Evans *J.Phys.A* **22** 2103 (1989).

The work presented in chapter 5 was carried out in collaboration with Prof D.J. Amit and Prof M.Abeles and has been submitted for publication as

"Attractor Neural Networks with Biological Probe Records"

D.J.Amit, M.R.Evans and M.Abeles to *Network: Computation in Neural Systems.*

Acknowledgements

I should like to thank Moshe Abeles, Daniel Amit, Heinz Horner, David Wallace, Michael Wong and Cui Zhan, in collaboration with whom parts of this work were carried out, for the benefit of their expertise and in particular Daniel Amit for patience and humour during my time in Jerusalem. Grateful thanks are due to the staff of the Racah Institute of Physics for the hospitality during that stay and to the staff of Imperial College and Mrs Volino for a pleasant summer in London.

I should also like to thank my supervisors: Alastair Bruce, for much advice and thorough reading of typescripts; David Wallace for vital encouragement, and would like to acknowledge the invaluable guidance of the late Elizabeth Gardner at the beginning of my research.

Finally, I have enjoyed many conversations with Meir Griniasty and Nava Rubin, and been saved from much computer despair by Jeremy Craven and Nigel Wilding.

Abstract

After a brief review of the history and philosophy of neural modelling, several attractor neural networks are studied in some detail. Firstly a variation of the Hopfield Model employed to perform symmetry invariant pattern recognition is considered. It is shown that parallel dynamics tend to perform a symmetry-transformation of the network configuration at each update. In contrast serial dynamics tend to drive the network configuration into a symmetry invariant. The component of the interactions that drive the aforementioned dynamic tendencies act as a noise upon the Hopfield interactions. However replica symmetric theory shows that an extensive number of patterns may be stored whilst allowing symmetry invariant pattern recognition.

The performance of Gardner optimal interactions, that optimise the performance of a perceptron, is examined in the context of attractor neural networks. The discussion is restricted to randomly dilute networks for which dynamical equations for the overlaps are available. A general analysis of these equations is performed and the transitions to no memory categorised. In particular the conditions for a point, tricritical in nature, to exist in the α–T plane are derived. Retrieval phase diagrams for the optimal interactions with and without errors in storage are constructed.

The case of sparse spatial coding is then investigated by considering two connection rules, Covariance and Willshaw, that have the storage capacities of the form of the Gardner optimal connections as the bias of the patterns becomes very large. In both cases the choice of threshold is crucial in order to achieve maximum storage, and also controls the basins of attraction of the memories. Both connection schemes exhibit an undesirable high activity attractor, but in the Willshaw case this may be suppressed by introducing an activity dependent inhibition.

In order to bring neural network models into close contact with biological experiment, the problem of firing rates is discussed. A model is then proposed that uses a biologically realistic dynamics and incorporates a variety of other biological features. Graphic displays from computer simulation of the model are presented and associative retrieval can be seen to occur whilst the network functions in a manner that is reminiscent of the results of biological experiments.

Contents

Chapter 1

Neural Modelling: From Hebb to Attractors

1.1 Neural Modelling and two Philosophies

The understanding of cognitive function and memory is a challenge with appeal to a variety of disciplines. The many levels at which studies may be undertaken range from considering the biochemistry through which the neural components function, to theorising about the origin of consciousness. In order to embark on a programme of research within this mélange, one ought first to know the context and limitations within which the investigations are to be carried out. Even if one is working within a purely mathematical framework the motivations and tradition that lead to the models being investigated must not be forgotten or else the resulting work maybe ephemeral and of interest only to the author and colleagues. Any importance will be lost to the rest of the scientific community. With this polemic in mind I will try to sketch the history and philosophy of the tradition of research that I feel the present thesis is continuing.

It was Psychologists who first considered the idea that behaviour and memory in particular could be construed on all levels as sequences of associations. One idea leads to another and action results from stimulation. Hebb was one of the first to theorise on how these associations could be effected by the cortical components of neurons and synapses. His book [1] contains a well developed theory of memory and learning without referring to detailed biology. This illustrates what I hold as a main tenet of neural modelling – too much biological detail may obscure the

underlying principles at work. Hebbs treatise centred around the concept of neural assemblies. These are groups of neurons that work together to perform some general computation. Computations were interpreted as reverberations where different subsets of the neural assembly would become active on response to different stimuli. The stimuli led to an association with an item in the memory store, the reverberation was the act of recall. The items in the memory store were considered to be coded in the synaptic connections between neurons in the assembly. Hebb also proposed a learning mechanism by which these connections could be strengthened in an unsupervised manner to allow the memory to be encoded.

These remarkably clear ideas on cognitive function still hold good today over fifty years after their inception. The work done in the interim has been to construct mathematical models to furnish the theory. These models in turn must evolve to become more and more biologically realistic once the principle in each modelling step has been investigated and understood. This is philosophy with which the present thesis was researched.

An alternative and contrasting approach to modelling is to examine carefully the biological evidence and construct a model that is a faithful representation of all the features. The rationale for this approach is that nature has endowed the system under study with many complex features and it would be precocious of the modeller to decide a priori what is essential and what is peripheral. A paradigm of this approach is the Hodgkin-Huxley equations[2] for the dynamics of the squid giant axon. These equations accurately reproduce some observed phenomenology of a single neuron at the expense of having to solve a system of four first order nonlinear differential equations. The next step in this approach has to be to remove some of the biological detail and see if it changes the properties of the model. I believe that removing detail and hoping that nothing changes gives the researcher less insight than starting with a simple model, adding detail and analysing the changes.

The ultimate aim of this thesis is to present a model neural network that realises Hebb's ideas and at the same time has enough biological detail to convince neurobiologists that the resulting model bears some relevance to their studies. In the spirit of the previous discussion one must first study some simple, unrealistic models in order to gain a solid understanding of principles of the modelling. First

a little biology is required to describe the system that we attempt to model.

1.2 A Little Biology

The fundamental components of the cortex are neurons and synapses. A neuron has three areas: a dendritic tree; the soma or cell body; and the axons. The synapses are the connections between neurons. A synapse is formed between the terminal of the axon of a pre-synaptic neuron and a dendrite of another, post-synaptic, neuron. The cell is separated from the extra-cellular fluid by the cell membrane. This membrane has different conductances to different ionic species. Concentration gradients of the different ions are maintained by a metabolic pump. Due to these concentration gradients, and the different conductances of the membrane to different ions, charge separation occurs at the cell membrane. This causes the cell membrane to be *polarised* so that the inside of the cell membrane is negative with respect to the outside. The polarisation is quantified by the *membrane potential*, V_m, which is the potential of the inside relative to the potential of the outside. The resting membrane potential is about -60 mV. Within the dendrites the membrane potential may become less negative, or *depolarised*, due changes in the conductivities of the membrane. This would be the result of an excitatory synaptic action. Alternatively the membrane potential within the dendritic tree, may become more negative or *hyper-polarised*, which would be the result of inhibitory synaptic action. These changes in membrane potential at the dendrites propagate electrotonically (with attenuation) towards the *trigger zone* of the cell membrane at the cell body. If the membrane potential at the trigger zone is depolarised past a threshold level, an action potential or spike occurs.

The action potential is a sudden increase in the membrane potential which for a short time becomes positive with respect to the outside of the cell. The membrane potential then returns to slightly below the resting potential for a short time, which is known as post-spike hyper-polarisation. Disregarding the effects of any incoming currents the membrane potential then decays back to the resting level. Action potentials are the means by which different neurons communicate with each other. The spike shape of the action potential is transmitted without attenuation along the axons of the cell. When the action potential reaches a synapse, transmitters

are released across the synaptic gap which cause changes in conductance in the post-synaptic cell's dendrites according to whether the synapse is excitatory or inhibitory, as discussed in the previous paragraph. The depolarisation required to cause a neuron to spike is greater than the excitatory post synaptic potential generated from a single incoming action potential from a presynaptic neuron. In order to depolarise to threshold each neuron is able to integrate its afferents over space and time.

This brief sketch of neural function is sufficient background knowledge for the first 4 chapters of this thesis. In chapter 5, the mechanisms by which a cell membrane can integrate afferents and the functioning of inhibitory synapses will be expanded upon. For a good introduction to neural mechanisms the reader is referred to [3].

Hebb's ideas concerning reverberations can be made more precise now that some basic neurophysiology has been discussed. The output of a single neuron is in the form of spiking. Experiments involving isolated single neurons show that when the neuron is stimulated by an external current then the rate of spiking increases with the strength of stimulus. Within an assembly of neurons in the cortex, an increased spike rate of an excitatory neuron will mean that the other neurons to which it has synaptic connections will receive larger excitatory post synaptic potentials which should increase their spiking rates. In this way the elevated spike rates can then sustain themselves and the interpretation of a reverberation becomes a set of neurons whose spike rates are elevated above the background or spontaneous rates. Of course, a particular neuron only makes connections with some of the other neurons in the vicinity. Thus only certain neurons can communicate with each other directly by means of action potentials. The synaptic connections are in general unidirectional so that if cell A synapses onto cell B it does not imply that cell B synapses onto cell A.

In pursuing Hebb's ideas about reverberations one is really dealing with feed-back systems. A feed back system is one where the neurons are densely interconnected so that the output of a neuron is fed back to become the input to other neurons in the assembly which in turn feed-back into the assembly. This is to be contrasted to feed-forward systems where a neuron in an assembly feeds its output forward to neurons in a different assembly which in turn feed-forward to other neurons. Feed-forward networks shall be discussed in section 1.4. Although the connectivity of

most parts of the brain falls somewhere between these two extremes, the cerebral cortex appears suitable for modelling as a feed-back system. This is because each neuron is connected to > 10000 other neurons. We shall interpret our neural assemblies as small regions of the cortex, of approximate volume $1mm^3$, which contain around 10^4 neurons.

1.3 Model Neural Components

For the present we have enough information to discuss the goal of the modelling. Although processes leading to the emission of an action potential are rather complicated, the Hodgkin-Huxley equations[2] do describe accurately the development of an action potential in a single neuron. However the processing involved in any neural computation is distributed over many neurons. It would be unfeasible to use the Hodgkin-Huxley equations to describe a system of say 1000 neurons. On top of that one would have to solve the countless equations for the propagation of action potentials along axons and for currents in the dendritic trees. Modelling would have to be performed for the functioning of synapses and so on. Even if such computation was possible, then to specify a particular system would require the values of a multitude of parameters that are not and cannot be known due to the unreliable nature of the neural components. Clearly simplification is required. Before any simplifications can be made one must consider the aims of a neural model. Any model must be tailored to what is required from it as well as to the physical system for which understanding is sought. When the mechanisms through which memory operates are under investigation then the focus of any theoretical study is how large numbers of neurons can act together in a co-operative manner to produce these mechanisms. The study will be facilitated considerably if the detailed internal mechanics of each neuron are omitted from the model so that the co-operative phenomena are highlighted.

One striking feature of the brain and mental function is the speed at which information processing can take place. For example our visual system collects and processes information so quickly that we can play racket sports, a task beyond present day robotics. What makes this speed even more remarkable is that the basic neural components, neurons and synapses, have operating speeds of the order

of milliseconds. This suggests that to achieve the observed speed, the information must be processed in parallel and distributed over large numbers of components. It is this parallel and distributed information processing that an initial model of memory must capture and the accuracy with which individual neurons are represented is of secondary importance. What must be done however, is to retain the salient features of how the neurons function. This requires an abstraction or idealised model neuron.

The most commonly used model neuron preserves three features from biology. Firstly the neuron summates inputs from other neurons. The result of this summation is compared to a threshold. An output is produced which depends on the result of the comparison. In this thesis the outputs of the model neuron are binary. The model neuron then becomes a simple logic device in the spirit of McCulloch and Pitts[4]. If the summed inputs exceed a threshold the neuron takes one value; if not it takes another. Analogue neurons have also been considered by some authors [5,6] but these shall not be studied in the present thesis. The inputs to a neuron represent the effect of dendritic currents in a biological neuron. The effect of these currents at the trigger zone depends on how much transmitter has been released, how far up the dendritic tree the synapse is located and many other factors. However the overall strength with which an action potential in the pre-synaptic neuron j affects the post-synaptic neuron i can be simply modelled by a synaptic weight J_{ij}.

In this thesis two binary representations for neuron outputs will be used. In chapters 2 and 3 Ising spin Neural Networks where the output of a neuron i is $S_i = \pm 1$ are studied; in chapters 4 and 5 neural components with outputs of 1,0 are investigated. For the present discussion the difference representation is not significant but I will use The Ising spin representation for concreteness. The input–output relation for a single neuron then becomes

$$S_i = \mathrm{Sgn}\,(h_i - \theta) \tag{1.1}$$

where

$$\begin{aligned} \mathrm{Sgn}(x) = &\quad 1 \quad \text{if} \quad x \geq 0 \\ = &\quad -1 \quad \text{if} \quad x < 0. \end{aligned} \tag{1.2}$$

h_i is the local field at site i due to a set of inputs I_j from other neurons j:

$$h_i = \sum_{j=1}^{N} J_{ij} I_j, \tag{1.3}$$

and θ is the threshold level. The local field h_i models the membrane depolarisation of a neuron. The interpretation of the binary neuron states is that the if the neuron takes value 1 then it is active in some sense, whereas if it takes value -1 then it is passive. This rather vague interpretation is unsatisfactory, as shall be discussed in chapter 5, when one begins to compare the results of models to biological experiment. In my view, the only feasible biological interpretation is that the 1 state indicates that the neuron has emitted an action potential as a result of the integration of inputs, whereas the -1 state indicates that the neuron has not. Some authors consider the 1 state to mean that the neuron has an elevated firing rate whereas the -1 state means that the neuron fires at the background rate. However I believe that directly equating the model neurons' outputs to firing rates is only reasonable when analogue neurons are used This is because experiments that artificially stimulate neurons show that the firing rate of a single neuron can vary continuously [3]. The problems of a direct interpretation of the binary neural states will be discussed further in chapter 5. For the present I shall put aside the problems with a biological interpretation of such neurons until that chapter, and refer to them as spins at lattice sites. A network of N sites is considered and the state of the assembly of neurons may be described by the configuration of spins $\{S_i\}$. A Mathematical representation of a reverberation, which we shall refer to as a *pattern*, can now be presented. Each pattern is a vector $\{\xi_i^\mu\}$ where the subscript i runs from 1 to N and is the index that labels the spins corresponding to each component. The superscript μ labels the patterns. Each component of the pattern vector is binary and can take values 1 or -1. The value 1 indicates that when the nominated pattern μ is being recalled the spin i should take value 1; a value of -1 for ξ_i^μ indicates that the spin S_i should take value -1.

1.3.1 Noise

An additional striking feature of the brain is that as well as the individual components being slow, they are not uniform and are unreliable. The unreliability stems from the fact that properties of neurons, such as the threshold, may fluctuate in time. In addition, the spontaneous activity of neurons outside of the

network which may have connections leading into the network, will mean that the membrane potential generated is not deterministic. Hebb foresaw this noisiness of neural components when he concluded that complete determinism within the association process was not consistent with attention. A way to model this noisiness [7,8] is to introduce stochasticity into the input–output relation:

$$\begin{aligned} S_i &= 1 \quad \text{with probability} \quad \left[1+\exp\left(-\frac{2h_i}{T}\right)\right]^{-1}, \\ &= -1 \quad \text{with probability} \quad \left[1+\exp\left(\frac{2h_i}{T}\right)\right]^{-1}. \end{aligned} \tag{1.4}$$

T is the temperature and parameterises the amount of noise in the network. As $T \rightarrow 0$, (1.4) reduces to (1.1) and the output is deterministic. At the other extreme, as $T \rightarrow \infty$, the output is completely random, with probability 0.5 of being 1 or -1 for any input.

1.4 Perceptrons and Pattern Associators

The idealised neurons described in the previous section first became well known in the context of perceptrons [9] and later for pattern associators [10,11] A perceptron, in its simplest form, is one model neuron that computes its output through (1.1) from a set of N inputs $\{I_j\}$. The perceptron can then classify all the possible input configurations into two classes depending on whether the output is ± 1. Learning algorithms were developed [9] to construct the synaptic connections to perform desired classifications. Minsky and Papert [12] were then able to define the types of classification that could be performed

Pattern associators [10,11] use a set of N model neurons. Pairs of patterns (nominated configurations of the neurons) are then chosen. Rules for constructing the connection strengths were developed so that on presentation of the first pattern of a pair as input to the associator, the second pattern is produced as the output. An association has then been made.

1.5 Attractor Neural Networks

Although perceptrons and pattern associators made use of the basic model neuron and synapse components, they did not fully realise the ideas of Hebb because they did not explore the idea of a self-sustaining reverberations. Perceptrons and pattern associators basically use a single updating of each model neuron so that an input produces an output. In an attractor neural network [5,7,13] the outputs are fed back into the network as inputs for the next updating. This produces a time evolution of the network, with each updating of the neurons corresponding to a time step. The network configuration then evolves in time. If after a long period of time the network configuration explores only a small area of the configuration space, then an attractor has been reached. Attractors realise to some extent the reverberations of Hebb.

1.5.1 Hopfield Model

The paper of Hopfield [14] opened the way for attractor neural networks to become amenable to analysis [15–17] by using symmetric synaptic connection strengths. In the Hopfield model patterns are stored by giving the synaptic connections between spins values according to the rule:

$$H_{ij} = \frac{1}{N}\sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu. \tag{1.5}$$

The index μ runs over the patterns so that P is the number of patterns stored. One may consider this set of synapses as the result of some learning process which may be interpreted as having some of the features proposed by Hebb. However for the purposes of this thesis the theory of how synaptic connections are formed and evolve is not central. Equation (1.5) also assumes that the network is fully connected. The symmetry of the synapses generated by 1.5 allows a configurational energy to be defined. The form of the energy depends on what type of dynamics is used. In the dynamics the process defined in Eq (1.4) is used as the updating procedure for each spin. However the synchronisation and order with which the updating is carried out within the network must also be specified. There are two obvious choices for the synchronisation: parallel dynamics whereby at each time step all neurons are updated simultaneously; asynchronous dynamics where

a time step is divided into N subdivisions and a different neuron is updated in each subdivision. One would hope that both of these dynamics should produce qualitatively similar performances. Hopfield[14] used asynchronous dynamics, in which case the configurational energy, when the threshold θ is taken to be zero, is given by

$$E = -\frac{1}{2}\sum_{i \neq j} H_{ij} S_i S_j. \tag{1.6}$$

Once a configurational energy has been defined, one can consider the shape of the energy surface over the space of network configurations, which is known as the energy landscape. Under the prescribed dynamics and at zero temperature the network will perform downhill motion in the energy landscape.

1.5.2 Content Addressable Memory and the Act of Recall

A content addressable memory is one where the item in the memory store is recalled by presenting the item or part of it or a distorted version of it rather than the address at which the item is to be found. In this way the content of the memory is itself the address. The energy landscape metaphor allows an intuitive understanding of how content addressable memory occurs in the Hopfield model. If a stored pattern is near to the bottom of valleys in this landscape, then the valley will become a basin of attraction for that pattern. From a corrupted version of the pattern, which corresponds to a point on the slope of the valley, the network configuration will evolve towards the valley floor thus recalling the stored pattern. The initial spin configuration forms a content address for the stored pattern. A slight variation of this theme is associative recall (c.f section 1.4). In this interpretation the initial configuration has been associated with a stored pattern. The act of recall has then been performed by an association.

In order to formulate the recall process mathematically one defines a parameter m that measures the overlap of the stored network configuration with the pattern to be recalled (taken to be pattern 1).

$$m = \frac{1}{N}\sum_{i=1}^{N} \xi_i^1 < S_i > \tag{1.7}$$

The local field h_i at a site i, which is given by

$$h_i = \sum_{j=1}^{N} H_{ij} S_j \tag{1.8}$$

can then be decomposed into a signal and a noise:

$$h_i = \xi_i^1 m + \frac{1}{N} \sum_{\mu>1, j\neq i} \xi_i^\mu \xi_j^\mu S_j \tag{1.9}$$

The first, signal, term is linear in the overlap order parameter m and consequently the larger the value of m the more likely the spin i is to line up to ξ_i^1. If the initial value of m is large enough, then after some updating, the spin configuration should match up with the stored pattern and the overlap m approach the value 1. The co-operative nature of the recall process, that has each spin interacting with all the others through synaptic connections, endows the model with a high degree of robustness. The robustness is manifested by a tolerance to various structural noises such as synaptic clipping and synaptic destruction [18], as well as dynamic noise (temperature).

The end result of the aforementioned recall process is that the network remains in the close vicinity of a stored pattern. The configuration of the network has evolved through the dynamics to an attractor. An attractor maybe a fixed point configuration as would be the case for zero temperature asynchronous dynamics; a limit cycle as may occur for parallel dynamics; or some set of configurations each with high overlap with stored pattern. The distinguishing feature of an attractor is that after the network has relaxed into the attractor only a restricted volume of the network's configuration space is explored. This is a loose statement of broken ergodicity.

1.6 Outline of Thesis

The starting point of the present thesis is the Hopfield Model[14], for which an analysis has been developed by Amit, Gutfreund and Sompolinsky[15–17]. This analysis will be presented and extended in chapter 2 where the capability of the Hopfield model for a particular form of associative memory will be explored. The analysis will centre on a variation of the Hopfield model where the interactions

(1.5) are augmented by an interaction component that generates a dynamic tendency to perform a symmetry transform on the network configuration. In the chapter first time step equations shall be derived. These equations describe exactly the dynamics tendencies on the first step of parallel dynamics. They shall demonstrate how a stored pattern and the symmetry transform of a stored pattern may be associated with the same attractor. However these equations also highlight how the symmetry transform component of the interactions acts as a structural noise on the Hopfield interactions. The effect of this noise is quantified by performing a replica symmetric analysis of the model and comparing the results with the replica symmetric analysis of a model with a simpler form of noise. Both calculations demonstrate the robustness of the Hopfield interactions to structural noise. Chapter 2 will also serve as a pedagogical introduction to the techniques used in the analyses and their limitations.

Chapter 3 will then proceed to consider the theoretical capabilities of attractor neural networks by examining how interactions, that optimise certain properties of perceptrons, fare in the more demanding environment of attractor neural networks with dynamical noise. There are two motivations for this. Firstly to understand how perceptron interactions generate attractors, as opposed to performing simple pattern association (see section 1.4); secondly to demonstrate that interactions that are optimal in one environment (the perceptron) may not be optimal in another environment (attractor neural network with dynamical noise). In order to carry out this study, randomly diluted networks must be considered, so that the dynamics for the overlap m (1.7) may be solved exactly. A general theory for the dynamic evolution of the overlap parameter m will be developed. This theory will then be applied to the interactions that optimise the storage of a perceptron.

In chapter 4 a move towards biological realism will be made by the consideration of networks that store patterns with low spatial activity. A pattern with low spatial activity only has small fraction of the spins that take the value 1. In the neural interpretation this corresponds to only a small fraction of neurons being involved in a computation, which seems to be borne out in biological experiment. Two synaptic connections schemes will be studied: the Covariance rule and Willshaw's rule. However the networks are found to be wanting in that as will as having memory attractors they still exhibit attractors with very high spatial activity. In the case of the Willshaw rule, a way to suppress this high activity attractor by

the use of inhibitory interactions, will be investigated. The method is effective, but there is a forfeit in another detail of biological reality.

In chapter 5 the limitations concerning the temporal firing rates of attractor neural networks will be fully discussed. The problem is basically that using dynamics based on (1.4), when the network is in an attractor the spins either take value 1 most of the time they are updated, or take value -1 most of the time. In the neural interpretation this corresponds to a neuron either emitting a spike at every opportunity, or not at all. This neural behaviour is not seen in biological experiment. The lessons from a discussion of why these extreme firing rates occur in the models and from chapter 4 will then be assimilated to present a biologically acceptable network. The model basically involves a complete revision of network dynamics. This network is examined by means of computer simulation, and graphic displays that illustrate some behaviours of interest will be presented. The results of these simulations may be compared directly with biological experiments.

Chapter 2

Storage Capacity for Symmetry Invariant Pattern Recognition Tasks

2.1 Symmetry Invariant Pattern Recognition

2.1.1 The Hopfield Model and Symmetry Invariant Pattern Recognition

In the previous chapter the Hopfield model was introduced and presented as a model for pattern recognition that used the principles of cognitive function. Moreover the attractors of the Hopfield model went some way towards providing realisations of the ideas of Hebb. However from a pragmatist's viewpoint the appeal of the Hopfield model is that one can borrow the techniques developed in the study of spin glasses to gain much analytic insight into the co-operative phenomena on which the pattern recognition mechanism relies.

An analysis of the Hopfield model for a finite number of stored patterns was first performed by Amit, Gutfreund and Sompolinsky[15] . They then developed a more complicated analysis to deal with the case of storing an extensive number of patterns [16,17]. In this chapter I will develop their analysis to examine a model that is an extension of Hopfield's. The model to be studied is primarily concerned with the idea of associative recall within the Hopfield model. As stated in section 1.5.2 the idea of associative recall is that a memory is evoked from a store through

an association with an input. In the Hopfield model this idea is carried out with the proviso that the input has large overlap with the stored pattern. However in pattern associator models [10,11] it is not necessary for the input and output vectors to be correlated in this way. The restricted association capabilities of the Hopfield model are most obvious when the network configuration is pictured as an array of pixels. A spin taking value -1 becomes a dark pixel whereas a spin taking value becomes a light pixel. If a simple symmetry-transformation, such as a rotation or reflection, of a stored pattern is presented to the network, the input configuration will have microscopic overlap m with the relevant stored pattern and the correct association will not be made. However to the eye the rotated array of pixels can still be recognised as the initial pattern. In other words the human brain can perform symmetry-invariant pattern recognition whereas the Hopfield model cannot.

One solution to this problem is to pre-process the initial configurations before presentation to the final Hopfield network [19,20]. This involves finding a mapping from the space of original configurations to the space of processed configurations that are presented to the Hopfield Network, such that original configurations and their symmetry transforms are mapped onto the same processed configuration. What is required is a symmetry invariant of the transformation. However this solution involves more than one network. In this thesis the capabilities of single networks of neurons are explored. It is of interest to see if a single network can in fact make wider associations than that of the original Hopfield model.

In section 1.5.2 the concepts of an Energy landscape and basins of attractions on that landscape were discussed. Within this metaphor one would like to be able to sculpt more complicated basins of attraction. For the example of recognising a stored pattern and its reflection as the same input one in fact requires a disjoint domain of attraction with two basins. One basin is the usual Hopfield model basin of attraction comprising configurations a small Hamming distance away from the stored pattern. The second basin should contain configurations a small Hamming distance away from the reflected pattern. In order to sculpt a channel from the latter basin to the former the Hopfield model must be modified.

Recently several schemes have been proposed to perform such a task. One proposal involves the use of dynamic connection strengths [21,22]. Another idea is to have

competing directions within the configurational flow produced by the dynamics of the network. The ideal that is sought with this approach is a network that symmetry-transforms the presented configuration until a macroscopic overlap with one of the stored patterns is found. At this point in the configurational flow the Hopfield interactions predominate and the pattern is recalled. In effect the network performs its own pre-processing.

To this end several mechanisms have been proposed. Dotsenko [23] has used modifiable thresholdings as the component of the model that causes the symmetry-transformation to occur. Coolen and Kuijk[24] have shown that the connections can be trained by example from pairs of configurations and their symmetry-transforms to perform the desired symmetry transformation. The study presented in the present chapter will be based upon [24] because the formulation of the model is simpler, and more amenable to analysis.

2.1.2 The Symmetry Transform Interaction

A simple interaction [24] that produces a symmetry transformation of the network configuration is given by

$$T_{ij} = a\delta_{j,\pi(i)} \quad i \neq j \tag{2.1}$$

$$= 0 \quad i = j, \tag{2.2}$$

where $\pi(i)$ is the site that i is mapped onto under the symmetry-transformation. Thus (2.1) couple sites to their image sites under the transformation. We will use the term "local" to refer to the fact that the lattice is fully connected so that the sites can be rearranged to leave a site and its image adjacent to each other. In this sense the symmetry transform interaction is local as opposed to H_{ij} where all sites have interactions with each other and the interaction is therefore long-range. Our study will centre on how the local structure of the symmetry transform interactions couples with the long range structure of the Hopfield interactions. In order that the symmetry transform interactions be as local as possible we choose a Z_2 symmetry. This will also give the convenient property of symmetric interactions $J_{ij} = J_{ji}$. In the transformation each spin is acted upon by an element of the group. The spin is either mapped onto itself (by the identity element) or mapped onto another spin. In the latter case the Z_2 constraint $\pi(\pi(i)) = i$ ensures that the pair of spins

at i and $\pi(i)$ are interchanged by the transformation. For simplicity we will set the number of spins mapped onto themselves to zero.

In order to see that a symmetry transformation is indeed performed by the interaction (2.1) consider the local field produced at a site i when this interaction is used by itself:

$$\begin{aligned} h_i(t) &= \sum_{j \neq i} T_{ij} S_j(t) \\ &= a S_{\pi(i)}(t). \end{aligned} \tag{2.3}$$

The field at each site has the sign of the spin at the image of the site and so, at least at zero temperature and parallel dynamics, the transformation will be faithfully performed. For serial dynamics however, a single updating sweep allows the system to converge to a transformation invariant configuration. To see this consider a pair of sites i and $\pi(i)$. If starting from time t, i is visited first in the updating sequence at time t_1, then the spin at i will be updated to $S_{\pi(i)}(t)$ so that $S_i(t_1) = S_{\pi(i)}(t)$. When the site $\pi(i)$ is visited at a later time t_2, the local field will be $h_{\pi(i)}(t_2) = aS_i(t_2) = aS_i(t_1) = aS_{\pi(i)}(t)$ so that the spin at $\pi(i)$ will already be aligned to its local field. Both spins then end up taking the value $S_{\pi(i)}(t)$. Whereas if $\pi(i)$ is visited first both spins will end up taking the value of $S_i(t)$. Clearly the order of updating within the sweep determines the final configuration but whatever order is chosen the final configuration will be symmetry invariant. It is now apparent that serial and parallel updating give contrasting dynamic tendencies: equation (2.3) showed that parallel dynamics faithfully performs the symmetry-transformation; sequential dynamics drives the configuration into a symmetry invariant.

This point is rather interesting as it implies that the two different dynamics endow the model with rather different properties. Recalling that in the neural interpretation these two dynamics represent different degrees of synchronicity then this may be an indication that synchronicity could be a parameter that is utilised in neural information processing.

The fact that the interactions are symmetric allows us to write down a configurational energy

$$E = -\frac{1}{2} \sum_{i \neq j} T_{ij} S_i S_j \tag{2.4}$$

$$= \frac{-a}{2}\sum_i S_i S_{\pi(i)}.$$

This simple model can be solved by basic statistical mechanics:

$$\begin{aligned} Z &= \mathrm{Tr}_S \exp(-\beta E) \\ &= \left[4\cosh\left(\frac{\beta a}{2}\right)\right]^{\frac{N}{2}}, \end{aligned} \tag{2.5}$$

where β is the inverse temperature and Z is the partition function. This energy only corresponds to the model that uses random sequential dynamics. As this dynamics drives the network into a symmetry invariant, one expects that an order parameter should be introduced that measures the symmetry invariance of the network. The order parameter, g, associated with this symmetry is given by

$$g = \frac{1}{N}\sum_i \langle S_i S_{\pi(i)}\rangle. \tag{2.6}$$

and corresponds to the magnetisation of a configuration $\{S_i^N\}$ generated by a local gauge transformation

$$S_i^N = S_i S_{\pi(i)}. \tag{2.7}$$

This configuration is invariant under the symmetry transform of the old configuration and is thus a symmetry invariant.

The value of this order parameter can be obtained from

$$\begin{aligned} g &= \frac{2}{\beta N}\frac{\partial \ln Z}{\partial a} \\ &= \tanh\left(\frac{\beta a}{2}\right). \end{aligned} \tag{2.8}$$

At zero temperature ($\beta \to \infty$) we find $g = 1$ if a is positive. In this case the thermodynamic equilibrium phase is that of ferromagnetic order in the positive magnetisation direction of the symmetry invariant configuration $\{S_i^N\}$. However in the spin space S_i this ground state is highly degenerate with the ground state entropy equal to $N\ln(2)/2$. The energy barriers between these ground states are not extensive so that at finite temperature the system will wander between ground states and make the phase paramagnetic.

2.2 The First Time-Step Equations

We now proceed to analyse the full model with synaptic interactions

$$J_{ij} = H_{ij} + T_{ij}. \tag{2.9}$$

Ideally we would like to investigate the dynamics of the model as we are interested in the associative properties of the network, which are determined by the dynamic behaviour. However to solve fully the dynamics of the Hopfield model is an insurmountable task due to the lack of self-averaging(see section2.3.1) in the dynamic evolution equations [25]. Roughly speaking this means that during the dynamic evolution of network, the spin configuration becomes correlated with the set of synaptic strengths in a way that depends on the particular realisation of the network. However the equation for first-time step of parallel dynamics($t = 1$) can be derived, if one assumes that the values of the spins in the initial configuration have been generated independently. The equation derived for the resulting overlap $m(1)$

$$m(1) = f(m(0)) \tag{2.10}$$

is known as a first time-step equation. The information this equation gives is the overlap $m(1)$ averaged over an ensemble of initial conditions corresponding to $m(0)$. A realisation of such an initial condition could be generated by taking the stored pattern and flipping each spin with probability $(1 - m(0))/2$.

In order to derive an equation of the form (2.10) for the model (2.9) we need to classify the sites according to how the symmetry transform interaction affects them.

1. The sites unchanged by the transformation.

$$i = \pi(i) \tag{2.11}$$

2. Pairs of sites that are interchanged by the transformation and which take the same value in the pattern

$$i \neq \pi(i) \text{ and } \xi_i^1 = \xi_{\pi(i)}^1 \tag{2.12}$$

3. Pairs of sites that are interchanged by the transformation and which take values of opposite sign in the pattern

$$i \neq \pi(i) \text{ and } \xi_i^1 = -\xi_{\pi(i)}^1 \tag{2.13}$$

We assume that the nominated patterns are random:

$$P(\xi) = \frac{1}{2}(\delta(\xi - 1) + \delta(\xi + 1)) \tag{2.14}$$

so that to within $\sqrt{N}$ fluctuations there are equal number of sites in classes 2) and 3). In the following we also disregard the possibility of sites being mapped onto themselves, so that there are no sites of class 1. However the equations can be easily generalised to include such a possibility. The overlap order parameters for our remaining two classes of sites are given respectively by

$$m_2 = \frac{1}{N}\sum_i < S_i > (\xi_i^1 + \xi_{\pi(i)}^1) \tag{2.15}$$

$$m_3 = \frac{1}{N}\sum_i < S_i > (\xi_i^1 - \xi_{\pi(i)}^1) \tag{2.16}$$

$$\tag{2.17}$$

so that

$$m = \frac{1}{2}(m_2 + m_3) \tag{2.18}$$

The distinction between the two overlap paramteres is that m_2 measures the overlap of the configuration with the pattern at sites where the pattern is symmetry invariant; m_2 measures the overlap at sites where the pattern is antisymmetric. In the thermodynamic limit the magnitudes of both m_2 and m_3 both lie between 0 and 1.

We must also quantify the ordering due to the symmetry transform interaction. For this purpose we consider

$$g = \frac{1}{2}(g_2 + g_3), \tag{2.19}$$

where

$$g_2 = \frac{1}{N}\sum_i < S_i S_{\pi(i)} > | \, \xi_i^1 + \xi_{\pi(i)}^1 \, | \tag{2.20}$$

$$g_3 = \frac{1}{N}\sum_i < S_i S_{\pi(i)} > | \, \xi_i^1 - \xi_{\pi(i)}^1 \, | \, . \tag{2.21}$$

The order parameter g measures the symmetry invariance of the system. The two orders, overlap with a stored pattern and symmetry invariance, compete directly in sites of class 3. For these sites the spins in pattern 1 at the site and image site are of opposite sign. When these spins align with the stored pattern one has antisymmetry under a symmetry transformation, so that $m_3 = 1$ implies $g_3 = -1$.

I shall now derive the first-time step equation for m_2. First it is convenient to write (2.15) as

$$m_2(1) = \ll \xi_i \langle S_i(1) \rangle \gg_2, \tag{2.22}$$

where the single angular brackets denote a thermal average and double angular brackets denote a composite average: an average over sites of class 2; and an average over the ensemble of initial configurations that have the specified overlaps $m_2(0)$ and $m_3(0)$. The thermal average can be performed straight away to give

$$m_2(1) = \ll \tanh\left(\xi_i \beta h_i(0)\right) \gg_2, \tag{2.23}$$

where the fact that tanh is an odd function has been used to absorb ξ into its argument. One must now assume that the averages over sites of class 2 and the ensemble of initial conditions are equivalent to averaging at a single site over all the possible values the local field with the appropriate probability distribution. This is an assumption that $m_2(1)$ is self-averaging which is why the double angular bracket notation has been used. The details of the term self-averaging will be discussed in section 2.3.1. This assumption results in

$$m_2 = \int dh_i^{(2)} \rho(h_i^{(2)}) \tanh\left(\xi_i \beta h_i^{(2)}\right), \tag{2.24}$$

where $h_i^{(2)}(0)$ denotes the local field at a site i which is of class 2, so that

$$\begin{aligned} h_i^{(2)} &= \sum_j J_{ij} S_j \\ &= \xi_i(m + a S_{\pi(i)}) + \frac{1}{N} \sum_{\mu>1,j} \xi_i^\mu \xi_j^\mu S_j. \end{aligned} \tag{2.25}$$

The first term in (2.25) is the signal and has finite mean. Recalling that i is a site of class 2 , the value of this mean is

$$\xi_i(m + a) \quad \text{with probability} \quad \frac{1}{2}(1 + m_2(0)) \tag{2.26}$$

$$\xi_i(m - a) \quad \text{with probability} \quad \frac{1}{2}(1 - m_2(0)). \tag{2.27}$$

The second term in (2.25) is noise. The noise is composed of a sum of $(P-1)(N-1)$ terms each of value $\pm 1/N$. The initial configuration is not correlated with the uncondensed patterns at all, therefore each term in the sum is independent. The probability distribution for this noise is a Gaussian with mean zero and variance α. The probability distribution for $h_i^{(2)}$ can now be constructed as a weighted sum of two Gaussians each with mean and weight given by (2.27) and both with variance α. $m_2(1)$ then becomes

$$m_2(1) = \int \frac{dh_i^{(2)}}{\sqrt{2\pi\alpha}} \left[\frac{1}{2}(1+m_2(0)) \exp - \left(\frac{(h_i^{(2)} - m(0) - a)^2}{2\alpha} \right) \right. \tag{2.28}$$
$$\left. + \frac{1}{2}(1-m_2(0)) \exp - \left(\frac{(h_i^{(2)} - m(0) + a)^2}{2\alpha} \right) \right] \tanh\left(\beta h_i^{(2)}\right).$$

The corresponding equation for $m_3(1)$ can be derived in a similar manner. As $\beta \to \infty$ ($T \to 0$) one finds

$$m_2(1) = \frac{1}{2}(1+m_2)\ \mathrm{erf}\left[\frac{m+a}{\sqrt{2\alpha}}\right] + \frac{1}{2}(1-m_2)\ \mathrm{erf}\left[\frac{m-a}{\sqrt{2\alpha}}\right] \tag{2.29}$$

$$m_3(1) = \frac{1}{2}(1+m_3)\ \mathrm{erf}\left[\frac{m-a}{\sqrt{2\alpha}}\right] + \frac{1}{2}(1-m_3)\ \mathrm{erf}\left[\frac{m+a}{\sqrt{2\alpha}}\right], \tag{2.30}$$

where the error function is given by

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x dz \exp(-z^2). \tag{2.31}$$

In these equations the order parameters on the r.h.s. are evaluated at $t = 0$.

Starting from the symmetry-transformed pattern ($m_2 = 1, m_3 = -1$) one finds

$$m(1) = m_2(1) = m_3(1) = \mathrm{erf}\left[\frac{a}{\sqrt{2\alpha}}\right]. \tag{2.32}$$

With a high enough value of a the pattern will be accurately reproduced after one parallel iteration. However if a is too large then at the second time step the transformed pattern will be reproduced. To get an idea of this effect we can consider a first time-step starting from the exact pattern. To do this we insert ($m_2 = 1, m_3 = -1$) in equations (2.29, 2.30). The result is

$$m_2(1) = \mathrm{erf}\left[\frac{1+a}{\sqrt{2\alpha}}\right] \tag{2.33}$$

$$m_3(1) = \mathrm{erf}\left[\frac{1-a}{\sqrt{2\alpha}}\right] \tag{2.34}$$

$$m(1) = \frac{1}{2}\ \mathrm{erf}\left[\frac{1+a}{\sqrt{2\alpha}}\right] + \frac{1}{2}\ \mathrm{erf}\left[\frac{1-a}{\sqrt{2\alpha}}\right]. \tag{2.35}$$

The overlap $m(1)$ given by (2.35) is maximised when $a = 0$. This corresponds to the Hopfield model. When a increases $m(1)$ decreases monotonically. For $a \to \infty$, $m(1) \to 0$. From these results one can surmise that increasing the strength of the symmetry-transform interaction a has an adverse effect on the recall capabilities of the model. However (2.32) demonstrates that increasing the value of a improves the symmetry-transformation capabilities of the model. There is a conflict between the two dynamic tendencies of the model which stems from the competition between the Hopfield and symmetry-transformation interactions. This competition will lead to a decrease in the storage capacity of the Hopfield model, which is similar to that caused by a noise, such as destruction of synapses[18]. However if one wishes to interpret the symmetry transform interactions as a noise on the Hopfield interactions, then it is a noise with some coherence. The coherence is demonstrated by the fact that these symmetry-transform interactions can produce a locally ordered zero temperature phase as was analysed in section 2.1.2.

Unfortunately first time-step equations, although straightforward to derive, are rather unsatisfactory. This is because one does not know what happens after the first time-step. For example it is possible that one may move towards the stored pattern at the first time-step and then move away from it later so that there is in fact no attractor corresponding to the stored pattern. In order to investigate the equilibrium properties of the model we must rely on the methods of statistical mechanics. This is in fact the complement of the first-time step equation as we shall be examining the long time behaviour of the model.

2.3 Self-Averaging and the Replica Method

2.3.1 Self-Averaging

Each realisation of the Hopfield model is defined by the choice of stored patterns. From these patterns the synaptic interaction strengths are constructed. The pattern vectors are said to be quenched random variables to denote that they are initially chosen in a random manner but then kept fixed. To derive physical properties of a particular realisation of the model would be an enormous task as one would have to specify each component of each pattern vector. Instead one seeks to

calculate average quantities by performing an average over all possible choices of the quenched random variables $\{\xi_i^\mu\}$. This quenched average is denoted by $\ll\gg$. The averaging amounts to calculating the average physical properties over the ensemble of all realisations of the Hopfield model. An important consideration is whether the quenched average reflects the physical properties of a typical realisation of the randomness. In other words whether there are large fluctuations about the quenched average in the thermodynamic limit. If the fractional deviation of a quantity, over the ensemble of all realisations of the randomness, vanishes in the thermodynamic limit, then the quantity is said to *self-average.*

The question remains as to a rule of thumb for which type of quantities self-average and which do not. Many quantities of physical interest are extensive thermodynamic quantities. These are observables which are proportional to the size of the system. They are often simply configurational averages of local quantities. An example is the overlap parameter m^μ:

$$m^\mu = \frac{1}{N}\sum_i \xi_i^\mu \langle S_i \rangle \tag{2.36}$$

Equation (2.36) is simply a definition. During a calculation one aims to develop an expression for $\xi_i^\mu\langle S_i\rangle$. This expression will depend on, amongst other parameters, the set of quenched random variables at that site:

$$\xi_i^\mu \langle S_i \rangle = f(\{\xi_i^1 \ldots \xi_i^P\}). \tag{2.37}$$

Each site i has its set of quenched random variables chosen independently, so in turn the function f becomes quenched at each site. However it is far more convenient to consider the quenched average of the function f rather than a random realisation of the function at each site. The configurational average, given by the sum over i, in the definition of m^μ is an average of f over sites. If this average is equivalent to a quenched average, so that

$$m^\mu = \ll f(\{\xi_i^1 \ldots \xi_i^P\}) \gg, \tag{2.38}$$

then m^μ is self-averaging. One says that an extensive quantity is self-averaging if the average of the corresponding local quantity over sites is equivalent to an average of the local quantity at a site over all possible realisations of the randomness at that site.

Continuing with the example 2.37, if f depends on the P bits ξ_i^μ assigned to a site i then there are 2^P possible realisations of f, each with equal probability of

occurring. The mean number of times a particular f appears over the N sites is $\frac{N}{2^P}$. The variance in this number is $\frac{N}{2^P}(1-\frac{1}{2^P})$. The fractional deviation in this number is then $\sqrt{\frac{2^P}{N}}$. If the fractional deviation vanishes as $N \to \infty$ then the number of times a particular f appears amongst the N sites varies negligibly from realisation to realisation and is given to within $\sqrt{N}$ by $N/2^P$. The distribution of f over sites for any realisation of the randomness then becomes an exact representation of the probability distribution of f and one has self-averaging. The condition for the fractional deviation to vanish is $2^P \ll N$ which holds for any finite P. To summarise one may say that extensive variables may often be self-averaging.

However there are other quantities to be calculated that may not be extensive. For example the partition function is exponentially related to the size of the system:

$$Z = \exp(-Nf). \tag{2.39}$$

If an average is performed on the partition function, the particular realisations of the randomness that minimise f will dominate the average. In effect the interactions are treated in the same way as the spin variables, and the average is analogous to the configurational trace. This implies that the interactions are in thermal equilibrium and the average is known as an annealed average. Therefore a typical realisation of the randomness will produce a value of Z far from the annealed averaged value and Z does not self-average.

2.3.2 The Replica Method

When using the techniques of statistical mechanics one usually tries to evaluate the partition function of the system as all the thermodynamic quantities can be calculated from it. However, as was noted in section (2.3.1), the partition function is not an extensive quantity and therefore does not self-average. On the other hand the free energy should be self-averaging as it is extensive. One should then start to calculate the quantity

$$f = -\frac{1}{\beta} \ll \ln Z\{J_{ij}\} \gg . \tag{2.40}$$

However the averaging of the ln function cannot be performed directly and one must use what is known as the replica method [26,27] to express the free energy in a more tractable form.

The replica method is based on the expansion

$$x^n = \exp(n \ln x) = 1 + n \ln x + \ldots \tag{2.41}$$

Taking the limit $n \to 0$ and rearranging terms leads to the identity

$$\ln x = \lim_{n \to 0} \frac{x^n - 1}{n}. \tag{2.42}$$

After setting $x = Z$ and inserting a quenched average, one obtains

$$\ll \ln Z \gg = \lim_{n \to 0} \frac{\ll Z^n \gg -1}{n}. \tag{2.43}$$

The quenched averaging on the r.h.s. of Eq. 2.43 is rather easier to perform than that on the l.h.s., due to the form the partition function takes. Z^n is the partition function of n replicas of the original system. The composite system can be considered as one system connected to the same heat bath but with no direct interactions between the replicas. To see this one can write

$$\begin{aligned} Z^n &= \mathrm{Tr}_{\{S_i^\alpha\}} \exp\left[-\beta \sum_{\alpha=1,n} E\left(\{J_{ij}\}, \{S_i^\alpha\}\right)\right] \\ &= \mathrm{Tr}_{\{S_i^\alpha\}} \exp\left[-\beta E'\left(\{J_{ij}\}, \{S_i^1\}, \ldots, \{S_i^n\}\right)\right]. \end{aligned}$$

The primed energy is then a function of the spins in all n replicas, $\{S_i^\varrho, \varrho = 1, \ldots, n\}$. However when we wish to take the limit $n \to 0$, n must become continuous and the physical interpretation of n replicas becomes rather difficult to understand. In fact there is no a priori correct choice for the mathematical form of this limit and in certain regions of the phase diagram the most obvious choice, the "replica-symmetric" one, is incorrect. However the effects due to replica symmetry breaking [28] are usually small corrections to the replica symmetric theory and in this thesis studies of replica symmetry breaking shall not be pursued.

2.4 Replica Symmetric Theory

In this section the replica-symmetric mean-field equations for the Hopfield model with symmetry-transform interactions shall be derived. First I shall give a summary of the approach, as the details of the calculation become rather lengthy and opaque. The starting point of the calculation is the expression for the free energy

$$f = \lim_{n \to 0} \lim_{N \to \infty} -\frac{\ll Z^n \gg -1}{\beta n}. \tag{2.44}$$

The initial calculation task is to perform the quenched averaging $\ll Z^n \gg$. One then proceeds to express this quantity as an integral over a space of order parameters, with the integrand an exponential of an extensive quantity. This extensive quantity is the free energy functional. In the thermodynamic limit the saddle points of the free energy dominate the integrals and give the equilibrium free energy. The equations for the saddle points form a set of mean field equations that can be solved to give the physical values of the order parameters. However to obtain the saddle point equations an ansatz must be made as to the form of the solutions. The ansatz used is replica symmetry.

The configurational energy for the model is

$$E = -\frac{1}{2}\sum_{i\neq j} H_{ij}S_iS_j - \frac{a}{2}\sum_i S_iS_{\pi(i)}. \tag{2.45}$$

We will consider the case where P, the number of stored patterns is proportional to N so that a storage ratio of the network may be defined as

$$\alpha = \frac{P}{N}. \tag{2.46}$$

The partition function of n replicas becomes

$$\ll Z^n \gg = \ll \mathrm{Tr}_{S^\rho} \exp\left[\frac{\beta}{2N}\sum_{ij\rho\mu}\xi_i^\mu S_i^\rho \xi_j^\mu S_j^\rho - \frac{1}{2}\beta nP + \frac{\beta a}{2}\sum_{i\rho} S_i^\rho S_{\pi(i)}^\rho\right] \gg, \tag{2.47}$$

where the i,j indices run from 1 to N; ρ is the replica index that runs from 1 to n; and μ runs over the stored patterns. The term $\frac{1}{2}\beta Pn$ is a correction for the component $i=j$ of the previous sum term. In order to linearise the contributions of the ξ's to the argument of the exponential a Gaussian transformation is used.

$$\exp\left(bx^2\right) = \int_{-\infty}^{\infty}\frac{dz}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}z^2 + \sqrt{2b}zx\right). \tag{2.48}$$

In the first term inside the exponential a Gaussian transformation can be carried out for each pair of indices ρ and μ. The integration variable for each pair shall be denoted m_ρ^μ.

$$\begin{aligned}\ll Z^n \gg \;&=\; \exp\left(-\frac{\beta pn}{2}\right) \ll \mathrm{Tr}_{S_i^\rho} \int\prod_{\mu\,\rho}\frac{dm_\rho^\mu}{\sqrt{2\pi}} \\ &\times\; \exp\left(-\frac{(m_\rho^\mu)^2}{2} + \sqrt{\beta N}m_\rho^\mu\frac{1}{N}\sum_i \xi_i^\mu S_i^\rho + \frac{\beta a}{2}\sum_i S_i^\rho S_{\pi(i)}^\rho\right) \gg\end{aligned} \tag{2.49}$$

However the argument of the exponent, which is the free energy, should be proportional to N. This requirement motivates a rescaling of the m's:

$$m \to \sqrt{N\beta}m, \tag{2.50}$$

and one obtains

$$\begin{aligned} \ll Z^n \gg &= \exp\left(-\frac{\beta p n}{2}\right) \ll \mathrm{Tr}_{S_i^\rho} \left[\frac{\beta N}{2\pi}\right]^{\frac{pn}{2}} \int \prod_{\mu\rho} dm_\rho^\mu \\ &\times \exp\left(-\beta N \frac{(m_\rho^\mu)^2}{2} + \beta N m_\rho^\mu \frac{1}{N} \sum_i \xi_i^\mu S_i^\rho + \frac{\beta a}{2} \sum_i S_i^\rho S_{\pi(i)}^\rho\right) \gg \end{aligned} \tag{2.51}$$

At this point one can see that if a saddle point is taken with respect to m_ρ^μ then one finds

$$m_\rho^\mu = \frac{1}{N} \ll \sum_i \xi_i^\mu < S_i^\rho > \gg . \tag{2.52}$$

However this saddle point is only valid if m is finite, so that it is only for a condensed pattern μ that m_ρ^μ is defined as the overlap order parameter for replica ρ. The main purpose of this chapter is to investigate how the symmetry transform interactions disrupt the retrieval state of the Hopfield model. We therefore assume that only the overlap with a particular pattern (taken to be pattern 1) is finite the others are microscopic ($O(1/\sqrt{N})$). For the patterns with microscopic overlaps the configurations $\{S_i^\rho\}$ that contribute to the free energy are not correlated with the pattern vectors, therefore the quenched averaging over these uncondensed patterns may now be performed. To do this one uses the cumulant expansion

$$\ll \exp(z) \gg = \exp\left[\ll z \gg + \frac{1}{2}\left(\ll z^2 \gg - \ll z \gg^2\right) + \ldots\right]. \tag{2.53}$$

For each i, μ we can set $z = \beta \xi_i^\mu \sum_\rho m_\rho^\mu S_i^\rho$, and find $\ll z \gg = 0$, $\ll z^2 \gg = (\beta \sum_\rho m_\rho^\mu S_i^\rho)^2$. As we assume that these overlaps are microscopic we need only keep the lowest order terms which result from $\ll z^2 \gg$. The total contribution from the microscopic overlaps becomes

$$\left[\frac{\beta N}{2\pi}\right]^{\frac{pn}{2}} \int \prod_{\mu\rho} dm_\rho^\mu \exp\beta\left(-\frac{\beta N}{2} \sum_{\mu,\rho} (m_\rho^\mu)^2 + \beta \sum_{i,\mu,\rho,\sigma} m_\rho^\mu m_\sigma^\mu S_i^\rho S_i^\sigma\right) . \tag{2.54}$$

The m_ρ^μ's can now be integrated out by using the general formula for Gaussian integrals

$$\int \frac{d\mathbf{x}}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}\mathbf{x}^t \mathbf{M} \mathbf{x} + \mathbf{b}^t \mathbf{x}\right) = \exp\left(-\frac{1}{2}\mathrm{Tr}\ln \mathbf{M} + \frac{1}{2}\mathbf{b}^t \mathbf{M}^{-1} \mathbf{b}\right) \tag{2.55}$$

where d is the dimensionality of the matrix M. In the present case the dimensionality of the matrix is the number of replicas n. One obtains

$$\int \prod_{\rho\neq\sigma} dq^{\rho\sigma} \exp\left(-\frac{\alpha N-1}{2}\mathrm{Tr}\ln \mathrm{M}\right) \prod_{\rho\neq\sigma} \delta\left(Nq^{\rho\sigma} - \sum_i S_i^\rho S_i^\sigma\right) \tag{2.56}$$

where

$$\mathrm{M} = (1-\beta)\delta_{\rho\sigma} - \beta q^{\rho\sigma}. \tag{2.57}$$

In Eq 2.56 dirac delta functions were used to introduce the Edwards-Anderson[34] order parameters $q^{\rho\sigma}$. In order that the partition function takes the correct form an integral representation of the dirac delta function may be used

$$\delta(x) = \int_{-i\infty}^{+i\infty} \frac{d\,y}{2\pi i} \exp(xy). \tag{2.58}$$

In Eq. 2.56 the delta function becomes

$$\delta\left(Nq^{\rho,\sigma} - \sum_i S_i^\rho S_i^\sigma\right) = \frac{\alpha\beta^2}{2\pi i} \int_{-i\infty}^{+i\infty} dr^{\rho\sigma} \exp\left(-\frac{N}{2}\alpha\beta^2 r^{\rho\sigma} q^{\rho\sigma} + \frac{1}{2}\alpha\beta^2 r^{\rho\sigma} \sum_i S_i^\rho S_i^\sigma\right) \tag{2.59}$$

The rescaling of the integration variables by $\alpha\beta^2$ is purely cosmetic as it simplifies the eventual saddle point equations. This exponentiation of the delta function has introduced another set of order parameters $r^{\rho\sigma}$. The saddle point of the free energy with respect to $q^{\rho\sigma}$ can be shown to define $r^{\rho\sigma}$ as

$$r^{\rho\sigma} = \frac{1}{\alpha} \ll \sum_{\mu=2}^{p} m_\rho^\mu m_\sigma^\mu \gg \tag{2.60}$$

In order to develop the determinant of the matrix in (2.55) further one must make use of the $n \to 0$ limit so that one can expand to first order in n. To be able to do this one must make some assumptions about the symmetry properties of the matrix $q^{\rho\sigma}$. The simplest scheme is that of replica symmetry

$$q^{\rho\sigma} = q \quad \forall \rho \neq \sigma. \tag{2.61}$$

With this ansatz one finds that

$$\lim_{n\to 0} \mathrm{Tr}\ln \mathrm{M} = n\ln\left[1-\beta(1-q)\right] - n\frac{\beta q}{1-\beta(1-q)}. \tag{2.62}$$

For consistency, the replica symmetric ansatz must now be used on all the saddle point order parameters. Now that the microscopic overlaps have been dealt with,

the superscript 1 shall be dropped from the condensed pattern.

$$m_\rho = m \quad \forall\rho \tag{2.63}$$

$$q^{\rho\sigma} = q \quad \forall\rho \neq \sigma \tag{2.64}$$

$$r^{\rho\sigma} = r \quad \forall\rho \neq \sigma. \tag{2.65}$$

The question of replica symmetry and the $n \to 0$ limit is actually rather subtle. The problems stem from the fact that the space of 0 by 0 square matrices is infinite so that there is no obvious analytic continuation from n by n matrices where n is an integer to $n = 0$. Parisi[29–32] has proposed a rather more sophisticated scheme for the symmetry properties of replica matrices and his replica symmetry breaking scheme is now generally accepted as correct. For a more detailed discussion of these topics a good reference is [33].

The development of the partition function now reads

$$\begin{aligned} \ll Z^n \gg &= \left(\frac{N\beta}{2\pi}\right)^{\frac{1}{2}} \left(\frac{\alpha\beta^2}{2\pi i}\right)^{n(n-1)} \int \prod_\rho dm_\rho \prod_{\rho\neq\sigma} dr^{\rho\sigma} dq^{\rho\sigma} \\ &\times \exp\left[-\frac{N\beta}{2}\sum_\rho (m_\rho)^2 - \frac{nN\beta\alpha}{2} - \frac{nN\alpha}{2}\ln(1-\beta(1-q))\right. \\ &\qquad \left. + \frac{nN\alpha}{2}\frac{\beta q}{(1-\beta(1-q))} - \frac{1}{2}\alpha\beta^2 N \sum_{\rho\neq\sigma} r^{\rho\sigma} q^{\rho\sigma}\right] \\ &\times \ll \exp\left[\sum_i \ln \mathrm{Tr}_{S_i^\rho} \exp\left(\beta \sum_{i,\rho} m_\rho \xi_i S^\rho \right.\right. \\ &\qquad \left.\left. + \frac{1}{2}\alpha\beta^2 \sum_{i,\rho\neq\sigma} r_{\rho\sigma} S_i^\rho S_i^\sigma + \frac{\beta a}{2}\sum_{i,\rho} S_i^\rho S_{\pi(i)}^\sigma\right)\right] \gg . \end{aligned} \tag{2.66}$$

The last exponential of this equation can be dealt with by invoking self-averaging. It then becomes

$$\exp\left[N \ll \ln \mathrm{Tr}_{S_\rho} \exp\left(\beta \sum_{i,\rho} m_\rho \xi_i S_i^\rho + \frac{1}{2}\alpha\beta^2 \sum_{i,\rho\neq\sigma} r_{\rho\sigma} S_i^\rho S_i^\sigma + \frac{\beta a}{2}\sum_{i,\rho} S_i^\rho S_{\pi(i)}^\sigma\right) \gg\right]. \tag{2.67}$$

One can then impose the replica symmetric ansatz and extract the free energy as

$$\begin{aligned} f &= \frac{\alpha}{2} + \frac{1}{2}m^2 + \frac{\alpha}{2\beta}\left[\ln(1-\beta+\beta q) - \frac{\beta q}{1-\beta+\beta q}\right] - \frac{1}{2}\alpha q r\beta \\ &\quad -\frac{1}{Nn\beta} \ll \ln\left[\mathrm{Tr}_{S_i^\rho} \exp\left(\frac{\alpha\beta^2 r}{2}\sum_{i,\rho\neq\sigma} S_i^\rho S_i^\sigma\right.\right. \end{aligned} \tag{2.68}$$

$$+\beta m \sum_i \xi_i^\mu \sum_\rho S_i^\rho + \frac{\beta a}{2} \sum_{i,\rho} S_i^\rho S_{\pi(i)}^\rho \Bigg) \Bigg] \gg$$

It is only the final term of Eq. 2.68 that differs from the equivalent expression for the original Hopfield model, which is recovered when $a = 0$. In the Hopfield model calculation this term could be developed by performing a Gaussian transform to linearise the first term inside the exponential for each site i and then factorising over sites i. In the present calculation we can only factorise over pairs of sites $i, \pi(i)$. However the configurational trace may then be taken to give the free energy as

$$\begin{aligned} f &= \frac{\alpha}{2} + \frac{1}{2}m^2 + \frac{\alpha}{2\beta}\left[\ln(1-\beta+\beta q) - \frac{\beta q}{1-\beta+\beta q}\right] + \frac{1}{2}\alpha\beta r(1-q) \\ &\quad -\frac{1}{2\beta} \ll \int Dz \int Dz_\pi \ln\left[2e^{\beta a}\cosh\left[\beta\sqrt{\alpha r}(z+z_\pi) + \beta m(\xi^1 + \xi_\pi^1)\right]\right. \\ &\quad \left. +2e^{-\beta a}\cosh\left[\beta\left(\sqrt{\alpha r}(z-z_\pi) + m(\xi^1 - \xi_\pi^1)\right)\right]\right] \gg \end{aligned} \tag{2.69}$$

where the Gaussian measure is denoted by

$$Dz = \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}\,. \tag{2.70}$$

The values taken by the order parameters appearing in Eq. 2.69 are given by the solution of the saddle point equations. In addition we have the order parameter associated with the symmetry transform interactions

$$g = -2\frac{\partial f}{\partial a}. \tag{2.71}$$

One finds the mean field equations to be

$$m_2 = \int Dz \int Dz_\pi \frac{e^{2\beta a}\sinh\left[\beta\left(\sqrt{\alpha r}(z+z_\pi) + 2m\right)\right]}{e^{2\beta a}\cosh\left[\beta\left(\sqrt{\alpha r}(z+z_\pi) + 2m\right)\right] + \cosh\left[\beta\left(\sqrt{\alpha r}(z-z_\pi)\right)\right]} \tag{2.72}$$

$$m_3 = \int Dz \int Dz_\pi \frac{\sinh\left[\beta\left(\sqrt{\alpha r}(z-z_\pi) + 2m\right)\right]}{e^{2\beta a}\cosh\left[\beta\sqrt{\alpha r}(z+z_\pi)\right] + \cosh\left[\beta\left(\sqrt{\alpha r}(z-z_\pi) + 2m\right)\right]} \tag{2.73}$$

$$\begin{aligned} q &= \frac{1}{2}\int Dz \int Dz_\pi \frac{e^{4\beta a}\sinh\ 2\left[\beta\left(\sqrt{\alpha r}(z+z_\pi) + 2m\right)\right] + \sinh^2\left[\beta\left(\sqrt{\alpha r}(z-z_\pi)\right)\right]}{\left(e^{2\beta a}\cosh\left[\beta\left(\sqrt{\alpha r}(z+z_\pi) + 2m\right)\right] + \cosh\left[\beta\left(\sqrt{\alpha r}(z-z_\pi)\right)\right]\right)^2} \\ &+ \frac{1}{2}\int Dz \int Dz_\pi \frac{e^{4\beta a}\sinh^2\left[\beta\left(\sqrt{\alpha r}(z+z_\pi)\right)\right] + \sinh^2\left[\beta\left(\sqrt{\alpha r}(z-z_\pi) + 2m\right)\right]}{\left(e^{2\beta a}\cosh\left[\beta\left(\sqrt{\alpha r}(z+z_\pi)\right)\right] + \cosh\left[\beta\left(\sqrt{\alpha r}(z-z_\pi) + 2m\right)\right]\right)^2} \end{aligned}$$

$$(2.74)$$

$$r = \frac{q}{(1-\beta+\beta q)^2} \tag{2.75}$$

$$g_2 = \int Dz \int Dz_\pi \frac{e^{2\beta a}\cosh\beta\left[(\sqrt{\alpha r}(z+z_\pi)+2m)\right]-\cosh\left[\beta\sqrt{\alpha r}(z-z_\pi)\right]}{e^{2\beta a}\cosh\beta\left[(\sqrt{\alpha r}(z+z_\pi)+2m)\right]+\cosh\left[\beta\sqrt{\alpha r}(z-z_\pi)\right]} \tag{2.76}$$

$$g_3 = \int Dz \int Dz_\pi \frac{e^{2\beta a}\cosh\beta\left[(\sqrt{\alpha r}(z+z_\pi))\right]-\cosh\left[\beta\left(\sqrt{\alpha r}(z-z_\pi)+2m\right)\right]}{e^{2\beta a}\cosh\beta\left[(\sqrt{\alpha r}(z+z_\pi))\right]+\cosh\left[\beta\left(\sqrt{\alpha r}(z-z_\pi)+2m\right)\right]}. \tag{2.77}$$

In Eqs (2.72-2.77) the average over the independent quenched spin variables ξ^1 and ξ^1_π has been taken. The parameters with subscript 2 result from the contributions with $\xi^1 = \xi^1_\pi$; the parameters with subscript 3 result from the contributions with $\xi^1 = -\xi^1_\pi$.

2.4.1 The Zero Temperature Limit

Obtaining the zero temperature limit of the mean-field equations (2.72-2.77) is a task requiring some patience. The strategy used is to examine the integrands of the equations and determine the regions of the $z - z_\pi$ plane where they do not vanish. In this process one finds that the symmetry $m_3(-a) = m_2(a); g_3(-a) = g_2(a)$ apparent in eqs (2.72-2.77), is broken in the $\beta \to \infty$ limit, because one must restrict the equations to the case $a > 0$. One eventually obtains the following zero temperature equations.

$$m_2 = \frac{1}{2}\,\mathrm{erf}\left[\frac{m+a}{\sqrt{2\alpha r}}\right] + \frac{1}{2}\,\mathrm{erf}\left[\frac{m-a}{\sqrt{2\alpha r}}\right] + \frac{1}{2}\int_{\frac{m-a}{\sqrt{\alpha r}}}^{\frac{m+a}{\sqrt{\alpha r}}} Dz\;\mathrm{erf}\left[\frac{2m}{\sqrt{2\alpha r}} - \frac{z}{\sqrt{2}}\right] \tag{2.78}$$

$$m_3 = \frac{1}{2}\,\mathrm{erf}\left[\frac{m+a}{\sqrt{2\alpha r}}\right] + \frac{1}{2}\,\mathrm{erf}\left[\frac{m-a}{\sqrt{2\alpha r}}\right] + \frac{1}{4}\,\mathrm{erf}^2\left[\frac{m-a}{\sqrt{2\alpha r}}\right] - \frac{1}{4}\,\mathrm{erf}^2\left[\frac{m+a}{\sqrt{2\alpha r}}\right] \tag{2.79}$$

$$\beta(1-q) = \frac{1}{\sqrt{2\pi\alpha r}}\left(e^{-\frac{(m+a)^2}{2\alpha r}} + e^{-\frac{(m-a)^2}{2\alpha r}}\right)\left(1 - \frac{1}{2}\,\mathrm{erf}\left[\frac{m+a}{\sqrt{2\alpha r}}\right] + \frac{1}{2}\,\mathrm{erf}\left[\frac{m-a}{\sqrt{2\alpha r}}\right]\right) + \frac{1}{2\sqrt{\pi\alpha r}}\,\mathrm{erf}\left[\frac{a}{\sqrt{\alpha r}}\right] e^{-\frac{m^2}{\alpha r}} + \frac{1}{4\sqrt{\pi\alpha r}}\left(\mathrm{erf}\left[\frac{m+a}{\sqrt{\alpha r}}\right] + \mathrm{erf}\left[\frac{m-a}{\sqrt{\alpha r}}\right]\right) \tag{2.80}$$

$$r = (1-\beta(1-q))^{-2} \tag{2.81}$$

$$
\begin{aligned}
g_2 &= \operatorname{erf}\left[\frac{m+a}{\sqrt{2\alpha r}}\right] - \operatorname{erf}\left[\frac{m-a}{\sqrt{2\alpha r}}\right] \\
&- \frac{1}{2}\operatorname{erf}^2\left[\frac{m-a}{\sqrt{2\alpha r}}\right] - \frac{1}{2}\operatorname{erf}^2\left[\frac{m+a}{\sqrt{2\alpha r}}\right] \qquad (2.82) \\
&+ 2\int_{\frac{2a}{\sqrt{2\alpha r}}}^{\frac{2(m+a)}{\sqrt{2\alpha r}}} Dz \operatorname{erf}\left[\frac{m+a}{\sqrt{\alpha r}} - \frac{z}{\sqrt{2}}\right] + 2\int_{-\frac{2a}{\sqrt{2\alpha r}}}^{\frac{2(m-a)}{\sqrt{2\alpha r}}} Dz \operatorname{erf}\left[\frac{m-a}{\sqrt{\alpha r}} - \frac{z}{\sqrt{2}}\right] \\
g_3 &= \operatorname{erf}\left[\frac{m+a}{\sqrt{2\alpha r}}\right] - \operatorname{erf}\left[\frac{m-a}{\sqrt{2\alpha r}}\right] \\
&-\frac{1}{2}\operatorname{erf}^2\left[\frac{m-a}{\sqrt{2\alpha r}}\right] - \frac{1}{2}\operatorname{erf}^2\left[\frac{m+a}{\sqrt{2\alpha r}}\right]. \qquad (2.83)
\end{aligned}
$$

The $a = 0$ limit of these equations gives the equivalent equations for the Hopfield model[16,17] with the symmetry order parameters taking the values

$$
\begin{aligned}
g_2 &= \operatorname{erf}^2\left[\frac{m}{\sqrt{2\alpha r}}\right] \\
g_3 &= -g_2
\end{aligned}
$$

2.4.2 The Phase Diagram

Eqs.2.78-2.83 show 2 distinct solutions. At all values of a and α there is a solution characterised by

$$
\begin{aligned}
m &= 0 \\
g_2 &> 0 \\
g_2 &= g_3.
\end{aligned}
$$

These order parameter values correspond to a phase which has some symmetry under the transformation $i \rightarrow \pi(i)$; we shall refer to it as the symmetric solution although the symmetry may not be total ($g < 1$). At low α and a we find the retrieval solution which is characterised by

$$
\begin{aligned}
m_2 &> 0 \\
m_3 &> 0 \\
m_2 &> m_3 \\
g_2 &> 0 \\
g_3 &< 0 \\
|\, g_2 \,| &> |\, g_3 \,|.
\end{aligned}
$$

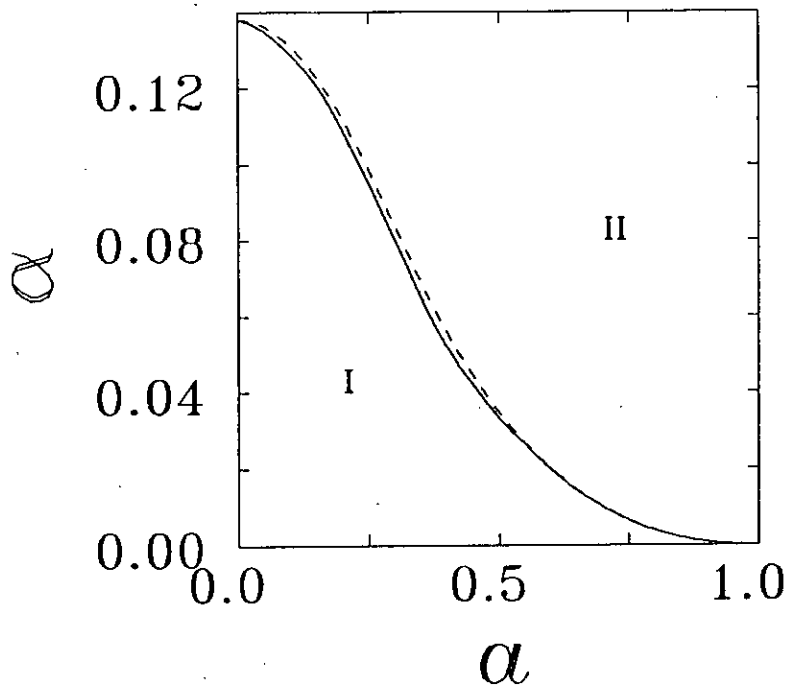

Figure 2.1: Phase Diagram for the Hopfield Model with
a) Z_2 symmetry interactions (full curve)
b) Random External Fields (dashed curve)
The retrieval(I) and non-retrieval(II) phases are described in the text

In Fig.2.1 the area marked I is the region in the space of $a - \alpha$ where the retrieval solution exists, which is the retrieval phase. The rest of the parameter space (II) constitutes the symmetric phase. The transition from retrieval phase to symmetric phase with increasing α is a first order one. At $a = 1.0$ one sees that $\alpha_c = 0$, showing that at this value of a there is no retrieval phase. The symmetric phase replaces the usual spin glass phase. Despite our nomenclature this symmetric phase is rather different from the ground states of the energy (2.4), because we have $q \to 1$. This implies that we have freezing of the spins so that the phase is spin-glass in character.

2.4.3 The Hopfield Model with Random External Fields

As we are interested in how the symmetry transform interactions act as a noise upon the Hopfield interactions it is useful to compare the phase diagram resulting from equations (2.78-2.83) to phase diagrams resulting from other simpler forms of noise. Eq. (2.3) shows that addition of symmetry interactions results in an

additional component of magnitude a to the local field generated by the Hopfield interactions. The sign of this field component is determined by the direction of the spin at the image site. A simpler form of noise results if the signs of these field components are assigned randomly. For this model, which we shall refer to as the Hopfield model with random external fields, the configurational energy is given by

$$E = -\frac{1}{2}\sum_{i\neq j} H_{ij}S_iS_j - a\sum_i \zeta_i S_i \tag{2.84}$$

where each ζ is selected randomly according to the distribution

$$P(\zeta) = \frac{1}{2}\left(\delta(\zeta - 1) + \delta(\zeta + 1)\right). \tag{2.85}$$

The random fields are then quenched in contrast to the fields in (2.3) which are "annealed" in character. The relevant separation of sites into two classes is now according to whether the pattern spin is in the same direction as the random external field. We have

$$\begin{aligned} m &= \frac{1}{2}(m_2 + m_3) \\ m_2 &= \sum_i (\xi_i + \zeta_i) < S_i > \\ m_3 &= \sum_i (\xi_i - \zeta_i) < S_i > \end{aligned}$$

The zero temperature mean-field equations are given by

$$m_2 = \frac{1}{2}\,\mathrm{erf}\left[\frac{m+a}{\sqrt{2\alpha r}}\right] \tag{2.86}$$

$$m_3 = \frac{1}{2}\,\mathrm{erf}\left[\frac{m-a}{\sqrt{2\alpha r}}\right] \tag{2.87}$$

$$m = \frac{1}{2}\,\mathrm{erf}\left[\frac{m+a}{\sqrt{2\alpha r}}\right] + \frac{1}{2}\,\mathrm{erf}\left[\frac{m-a}{\sqrt{2\alpha r}}\right] \tag{2.88}$$

$$\beta(1-q) = \frac{1}{\sqrt{2\pi\alpha r}}\left(e^{-\frac{(m+a)^2}{2\alpha r}} + e^{-\frac{(m-a)^2}{2\alpha r}}\right) \tag{2.89}$$

$$r = (1 - \beta(1-q))^{-2} \tag{2.90}$$

The phase diagram for this model is also shown in Fig.2.1 . The non-retrieval phase is where the only solution to the equations gives $m_3 < 0$ indicating that the spins are aligned to the random external fields rather than to a stored pattern.

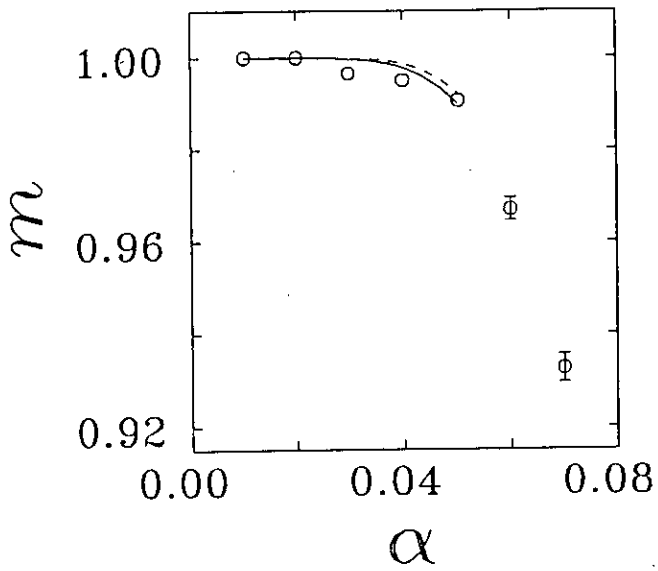

Figure 2.2: Overlap m versus α for a=0.4:
Full curve - Z_2 symmetry interactions
Dashed curve - Random external fields
Symbols - results of simulations (see text for details). Error bars are shown only when they are larger than the symbol size.

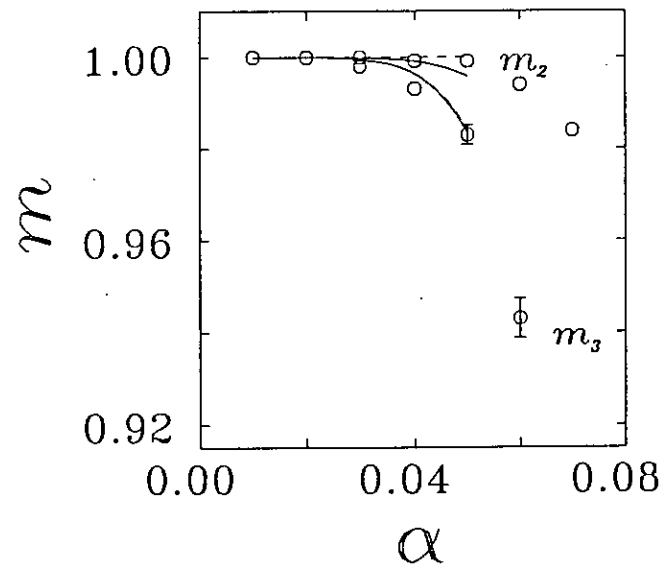

Figure 2.3: Overlaps m_2 and m_3 versus α for a=0.4:
Full curve - Z_2 symmetry interactions
Dashed curve - Random external fields
Symbols - results of simulations (see text for details). Error bars are shown only when they are larger than the symbol size.

Again $a = 1.0$ is the value of a at which there is never a retrieval phase. If one ignores this physical difference in the non-retrieval phase then the two phase diagrams are strikingly similar. To investigate the extent of the similarity figures 2.2 and 2.3 compare the solutions of the zero temperature mean-field equations for the two models. In figure 2.2 the value of the overall overlap m can be seen to be very similar in both models. The random external field model in fact gives a slightly higher value of m. Both models are in reasonable agreement with the result of simulations, up to the phase transition. In figure 2.3 the values of m_2 and m_3 for the two models are compared. The m_2 curves are almost identical, so much so that the random external field curve is obscured. For m_3 however, the random external field model gives a higher value than the symmetry-transform model. This suggests that at sites of class 3 the symmetry-transform interactions act as a stronger noise than random external fields on the Hopfield interactions. To summarise, it appears that the zero temperature mean-field equations for the Hopfield model with random external fields approximate rather well the far more complicated equations for the Hopfield model with symmetry interactions.

2.5 Parallel Dynamics and Invariant Pattern Recognition

We showed in section 2.1.2 that using the symmetry transform under parallel and sequential dynamics gives considerably different configurational flows. If we consider imposing a transformed version of a stored pattern on the network and iterating, then for serial dynamics the direction of the configurational flow induced by the symmetry transform interaction is towards a symmetrised version of the pattern. In terms of the overlap order parameters the transformed pattern is given by $m_2 = 1, m_3 = -1$ and $m = 0$; the symmetrised version is given by $m_2 = 1, m_3 = 0$ and $m = 0.5$. Thus starting from the transformed exact pattern we reach a configuration which is the pattern with 25% noise. If the Hopfield interactions then start to dominate the configurational flow it would only be for relatively low storage levels that the exact pattern could be recalled [35]. In light of this we consider the use of serial dynamics, less suitable for invariant pattern recognition than the use of parallel dynamics.

In section 2.2 the competition between the Hopfield and symmetry-transform interaction was emphasised. The effectiveness of the symmetry transform interactions was gauged by deriving and examining the first time step equations. The mean-field equations have served to gauge the effectiveness of the Hopfield interactions with respect to pattern storage. From the two sets of equations in tandem the optimum value of a may be gauged by using the following guidelines.

$$a > \sqrt{2\alpha} \tag{2.91}$$

$$\alpha < \alpha_c(a), \tag{2.92}$$

where $\alpha_c(a)$ is the phase boundary plotted in Fig.2.1. The first condition comes from the requirement that the transformed pattern be mapped accurately onto the pattern after one parallel update; the second condition comes from the requirement that the symmetry interactions should not disrupt the the retrieval phase of the Hopfield model. The guidelines are rules of thumb rather than quantitative bounds because on the one hand, the basins of attraction of the fully connected Hopfield model cannot be simply parameterised, which makes the first condition (2.91) rather arbitrary. On the other hand parallel dynamics appear more suitable for symmetry invariant pattern recognition whereas the second condition (2.92) applies to random sequential dynamics. However even if the guidelines are held in no greater esteem than rules of thumb, then the optimal value of a that they yield may well give the true optimal value of a if one does not specify the accuracy to greater than 1 decimal place. To determine the optimal value of a from (2.91-2.92) one simply searches for the maximum value of $\alpha_c(a)$ subject to the constraint (2.91). This approximation suggests an optimal value of $a \approx 0.4$ at which $\alpha_c \approx 0.06$. These numbers indicate that if we desire invariant pattern recognition, then although the maximum capacity of the network is reduced from the Hopfield case, we can still store an extensive number of patterns.

In Fig. 2.4 the results of numerical simulations of invariant pattern recognition using parallel dynamics are presented. In these simulations the transformed pattern was presented to the network and 20 parallel iterations were performed. The final overlaps were then calculated. The even number of iterations is convenient because if the symmetry transform interactions dominate and the configuration is symmetry transformed at each time step, then after an even number of iterations the configuration will return near to the transformed pattern, which has small overlap with the nominated pattern. We also found that 20 iterations was enough

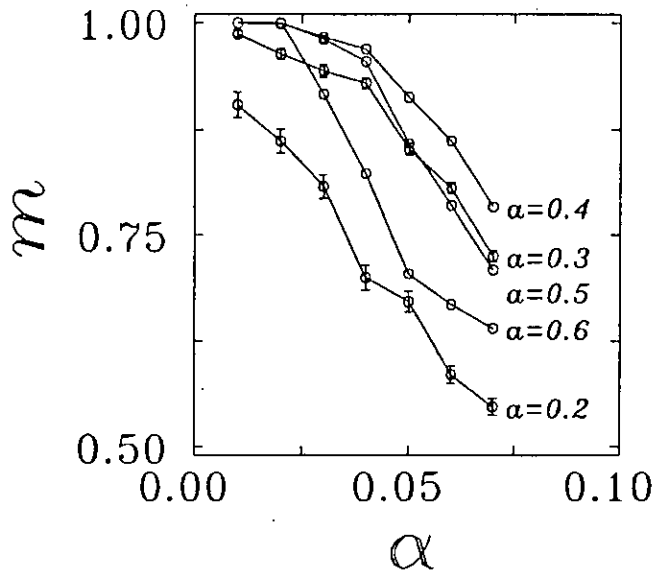

Figure 2.4: Simulation results for recognising transformed patterns under parallel dynamics. Error bars are shown only when they are larger than the symbol size. The lines are drawn solely in order to guide the eye between simulation points for constant a.

to ensure that a fixed point was reached if one existed. The figure indicates the existence of an optimal value of a because the curves rise and fall, to some extent, as a increases. The optimal value appears to be $a \approx 0.4$, as predicted from (2.91–2.92). However the retrieval quality m appears to deteriorate when α passes 0.05, which is lower than the α_c predicted from (2.91–2.92). At $a = 0.2$ one may note that even for very low α the recall of the transformed pattern is not good, indicating that the dynamic tendency to transform the pattern is not strong enough. In contrast, for $a = 0.6$ the recall is excellent at $\alpha < 0.02$, but deteriorates rapidly as α increases. This reflects the disruption of the retrieval attractors at low α, for high a.

2.6 Discussion

In this chapter we have investigated the Hopfield model with Z_2 symmetry transform interactions with two aims in mind. Firstly we have demonstrated how basins of attraction on the energy surface may be sculpted by introducing appropriate interactions. The Z_2 symmetry transform interactions that were used are a specific example that illustrate the general principle. Increasing the basin of

attractions and therefore the content addressability results in a decrease in absolute storage capacity. The second aim was to demonstrate how the available techniques may be employed to probe analytically the properties of a model. The two techniques that were used were the first time step dynamic equations and replica-symmetric mean-field theory. Both calculations yielded quantitative results that exposed the competition between the two types of interaction, Hopfield and symmetry-transform. However a full solution of the model, which would be realised by obtaining self-averaging dynamic equations for the evolution of the order parameters at all times, is not possible.

If we recall the discussion of the first chapter that attempted to justify the simplifications inherent in the Hopfield model when used as a model of neural information processing, then the present chapter has offered the reader more compelling material. Here I have presented the fruits of the simplification which come in the form of clear analytic insight.

As well as being a pedagogical example this chapter has also covered new ground in neural network theory. Most obviously the chapter has presented quantitative results on the performance of the Hopfield model applied to symmetry-invariant pattern recognition. However several secondary lines of investigation have also been pursued. The first was the contrasting properties of serial and parallel dynamics. This is a bit of a dark horse. On the one hand it can be said that serial and parallel dynamics represent opposite degrees of synchronicity within the neurons of the network. Therefore the different properties of serial and parallel dynamics reflect the fact that the level of synchronicity is a degree of freedom that may be utilised constructively within neural processing. On the other hand one may demand that the properties of the Hopfield model and its variants should not be strongly affected by the particular implementation of the dynamics. Within that context the present work becomes a counter-example that highlights the shortcomings of the Hopfield model and its dynamics.

The other subsidiary theme of the chapter was how the local (in the sense defined in sec. 2.1.2) symmetry-transform interactions act as a noise upon the long range Hopfield interactions. Even though there were some sites (those of class 2) at which the two sets of interactions sought to align the spins in the same manner the calculations showed that these particular local and long-range interactions always

competed rather than co-operated. This statement is backed up by the random-external field model. Here the additional elements to the Hopfield interactions, which are the quenched random external fields, are uncorrelated with the patterns and therefore truly a noise. However the phase diagrams for the random external field model and symmetry-transform model are virtually identical, apart from the interpretation of the non-retrieval phases.

Chapter 3

The Theoretical Capabilities of Attractor Neural Networks

3.1 Introduction

In the previous chapter an attractor neural network with explicit synaptic interactions designed for a specific pattern recognition task was examined. Within the framework of attractor neural networks one would like to consider what is the best choice of synaptic interactions. Of course *best* is not a well defined word until one has some definite criterion for comparison. Fortunately the analysis of chapter 2 did produce a quantitative measure of performance - the storage capacity α_c. However a programme of choosing sets of synaptic interactions and then analysing to find the storage capacity that they yield, would be rather time consuming. Gardner[36,37] has circumvented this problem by creating a theoretical framework in which the space of all synaptic interactions is searched to find the fractional volume of this space that can recall exactly the stored patterns. When this fractional volume vanishes the maximum storage capacity has been reached. The subtleties of this approach, its implications and the results it produces will be discussed fully in the next section. For the moment I will assume that it can yield a set of optimal interactions that maximise α subject to a constraint, the simplest constraint being that all patterns are stored exactly.

Having found a set of optimal interactions the question remains as to the scope of their optimality. Interactions that are optimal with respect to one criterion may

not be optimal with respect to another closely related criterion. Additionally we have the question of robustness. A set of interactions may well have the optimal storage capacity but in the presence of some sort of noise the performance may deteriorate drastically. Both of these points can be envisaged as a system highly adapted to a particular environment. When the environment changes, will the system still be successful or will a less highly adapted system be more successful in a variety of environments? In the living brain the environment is constantly changing. Neurons and synapses die and parameters such as thresholds and the quantity of synaptic transmitter released fluctuate. The synaptic strengths used for information processing must be able to cope with this noisy environment, they must be robust.

One aim of this chapter is to examine how the optimal interactions of Gardner perform in the presence of noise. The discussion of the previous paragraphs has basically argues that there is no a priori reason to expect that these interactions are optimal in the presence of noise. As discussed in Chapter 1 noise may be parameterised by a temperature T. At zero temperature the Gardner interactions are optimal because they maximise the number of patterns that can be stored exactly. As the temperature increases these interactions may perform less well. We shall make a comparison with the Hopfield interactions to show that as the noise increases there is some cross over point where the Hopfield interactions out perform those of Gardner. The more general aim of this chapter is to examine how the Gardner interactions perform in the context of attractor neural networks. This is because the Gardner calculation uses the framework of perceptrons (see section 1.4). The question is then whether the optimal interactions for one framework (the perceptron) also perform well in a related framework (attractor neural network with noise).

In the next section I shall review Gardner's calculation[36–38] so that the aims of the present chapter can be more clearly defined in section 3.3. The theoretical technique, random dilution, by means of which the study will be carried out will be reviewed in section 3.4. Random dilution allows the overlap dynamics of a network to be explored through an order parameter map, once the distribution of the local fields is known. This distribution will be calculated in section 3.5. A general analysis of the order parameter map, which will lead to understanding of the attractor structure and transitions in the attractor structure, will be developed

in section 3.6. The results of this analysis will then be applied in section 3.7 to investigate the aims set out in section 3.3.

3.2 The Gardner Volume

Gardner considered a pattern associator to which an input vector was presented and an output produced according to the parallel dynamic zero temperature updating rule

$$S_i(t+1) = \mathrm{Sgn}\left(\sum_j J_{ij} S_j(t)\right). \tag{3.1}$$

For pattern storage one desires a nominated input vector (the pattern) $\{\xi_i^\mu\}$ to be reproduced by the updating procedure (3.1). This shall be referred to as a pattern being stable. If interactions can be found to perform this task then under a sequential updating sweep the input vector will again be reproduced. Furthermore under iterative dynamics, where the output becomes the input for the next time-step, the nominated vector will be a fixed-point of the dynamics. Therefore the pattern will also be stable in a noiseless attractor neural network.

In order that each nominated pattern be stable a set of constraints, one for each spin i in each pattern μ, must be satisfied by the interactions

$$\Delta_i^\mu \geq \kappa \tag{3.2}$$

where

$$\Delta_i^\mu = \frac{\xi_i^\mu \sum_j J_{ij} \xi_j^\mu}{\sqrt{\sum_j J_{ij}^2}}. \tag{3.3}$$

The parameter κ is a stability parameter. For the pattern to be stable one requires $\kappa \geq 0$ at each site. In the Gardner approach one considers the volume of the space of interactions that satisfies these constraints. However, if there is a set of interactions that satisfy the constraints (3.2) then simply multiplying each interaction by a positive factor will generate another set of interactions that satisfy (3.2). In this way an infinite family of sets of interactions that satisfy (3.2) could be generated. Therefore for the volume of interactions to be well defined one must put a constraint on the possible sets of interactions. Gardner chose the spherical constraint

$$\sum_j J_{ij}^2 = N \quad \forall i \tag{3.4}$$

The sets of constraints (3.2) at each site i are independent because the interactions J_{ij} and J_{ji} are independent variables. One need only consider the typical space of interactions at a site i. This reduces the problem to that of a perceptron at a site i. In order to do this some self-averaging property must be invoked. First one must determine which quantity will self-average.

One may consider V_C, the volume of interactions that satisfy the constraints (3.2) and (3.4) at a site i.

$$V_C = \int \prod_j dJ_{ij} \prod_\mu \Theta\left(\Delta_i^\mu - \kappa\right) \delta\left(\sum_j J_{ij}^2 - N\right). \tag{3.5}$$

However if V_T, the volume that satisfies the spherical constraints (3.4), is evaluated one finds

$$V_T = \int \prod_j dJ_{ij} \, \delta\left(\sum_j J_{ij}^2 - N\right) = \exp\left(\frac{N}{2} + \ln(2\pi)\right). \tag{3.6}$$

This suggests that the volume V_C will be of the form $\exp(Ng)$. This quantity will not be self-averaging as discussed in section (2.3.1). Instead one must consider $\ll \ln V_C \gg$.

The argument that one should average the logarithm of the required volume rather than the volume is reminiscent of the one that demanded the averaging of the logarithm of the partition function in chapter 2. In fact Gardner and Derrida[38] have taken this analogy further. If one rewrites the Heaviside function as

$$\Theta(x) = \lim_{h\to\infty} \exp\left(-h\Theta(-x)\right), \tag{3.7}$$

and writes J_{ij}' to denote that the values over which the synaptic interactions are integrated are those allowed by the spherical constraint, then the volume appears as

$$V_C = \lim_{h\to\infty} \int \prod_j dJ_{ij}' \exp\left(-h \sum_\mu \Theta(\kappa - \Delta_i^\mu)\right). \tag{3.8}$$

This resembles the zero temperature limit of a partition function with the configurational trace replaced by the integrations over interactions; the inverse temperature replaced by h and the configurational energy replaced by a cost function

$$C_i = \sum_{\mu=1}^{P} \Theta(\kappa - \Delta_i^\mu). \tag{3.9}$$

Within this interpretation the synaptic interaction strengths are the degrees of freedom in thermal equilibrium. They become the variables that are replicated in the calculation and for them one obtains an Edwards Anderson order parameter

$$q^{\rho\sigma} = \frac{1}{N}\sum_j \langle J_{ij}^{\rho} J_{ij}^{\sigma} \rangle. \tag{3.10}$$

The angular brackets are analogous to a thermal average. The physical interpretation is that it is an average, weighted with the Boltzmann like factor in (3.8), over all sets of interactions. The limit $h \to \infty$ corresponds to projecting out the sets of synaptic interactions that yield the ground state of the cost function (3.9). The limit $q \to 1$, in turn indicates that a finite fraction of the space of interactions has been singled out. It is this limit that in fact yields the optimal storage capacity. The value of α at which $q \to 1$ is the storage level at which the volume of interactions satisfying the stability constraints shrinks to a point.

Within this cost function representation of the volume one may also investigate the situation where the constraints (3.2) cannot all be satisfied at a site i. This will occur if there are too many patterns P. The patterns μ for which the constraints are not satisfied, $\Delta_i^{\mu} < \kappa$, have then been incorrectly stored and *errors* have occurred. The cost function (3.9) is equal to the number of errors at a site. In this case the sets of interactions that minimise the cost function 3.9 are those that produce the least errors, $P f_{min}$. The minimum fraction of errors, f_{min}, may be evaluated by

$$f_{min} = \frac{1}{P}\lim_{h\to\infty} -\frac{d}{dh}\ln V_C(h). \tag{3.11}$$

The key results that arose from the papers by Gardner[36,37] and Gardner and Derrida[38] are

$$\alpha_P(\kappa, f_{min}) = \left(\int_{\kappa-x}^{\kappa} Dz(z-\kappa)^2\right)^{-1} \tag{3.12}$$

where the error fraction is given by

$$f_{min} = \int_{-\infty}^{\kappa-x} Dz. \tag{3.13}$$

If one demands that there be no errors so that $f_{min} = 0$, then one finds $x \to \infty$ and one recovers Gardner's original result[36]

$$\alpha_P(\kappa, 0) = \left(\int_{-\infty}^{\kappa} Dz(z-\kappa)^2\right)^{-1} \tag{3.14}$$

which for $\kappa = 0$, corresponding to minimal stability constraints, yields $\alpha_P(0,0) = 2$.

When the stability of the replica symmetric solution to the saddle point equations was checked it was found[38] that replica symmetry is broken when

$$\frac{x}{\sqrt{2\pi}} \exp\left[\frac{(\kappa - x)^2}{2}\right] > \kappa \int_{\kappa - x}^{\kappa} Dz(\kappa - z). \tag{3.15}$$

For $f_{min} = 0 (x \to \infty)$ replica symmetry is never broken if $\kappa \geq 0$. Whereas for $f_{min} > 0$ (x finite) replica symmetry may be broken and in the regions of the κ–x plane where this occurs, (3.12) is invalid.

3.3 Aims of the Chapter

3.3.1 Storage Capacities — α_P and α_c

The maximal storage capacity is denoted α_P in (3.12) to distinguish it from the critical storage capacity α_c that was the subject of calculation in the previous chapter. This is because it is only for a zero temperature ANN and Gardner interactions with error fraction equal to zero, that the two quantities coincide. $\alpha_P(\kappa, f_{min})$ gives the maximum number of patterns that may be stored with stability κ and error fraction f_{min}, and is a parameter concerned with the training of the network. In contrast α_c, as calculated in chapter 2, is a parameter concerned with the performance of a network during retrieval. In that chapter the augmented Hopfield interactions were considered and they formed a family of sets of interactions parameterised by α, the number of patterns stored, and a the strength of the symmetry transform component. $\alpha_c(a)$ then gave the maximum number of patterns that could be stored and had attractors associated with them, for the particular value of a. If Temperature had been considered in that chapter then $\alpha_c(a, T)$ could have been calculated. It should be noted that α_c is a critical value of a quantity α that parameterises a particular family of sets of interactions.

The Gardner interactions on the other hand, are parameterised by κ and f_{min}. If we consider first the case of no errors $f_{min} = 0$, the interactions are parameterised

by κ which can be converted into a parameterisation by α_P through (3.14). In this case the definition of $\alpha_c(T)$ is the maximal value of α_P at which there are attractors associated with each pattern. More fundamentally we are considering $\kappa_c(T)$ the minimal value of the stability parameter κ at which the Gardner interactions give attractors. At $T = 0$ the patterns will be fixed points of the dynamics for all positive κ so that $\kappa_c(T = 0) = 0$ and $\alpha_c(T = 0) = 2$.

The situation is more complicated when f_{min} is not restricted to zero. In this case α_P does not fully parameterise a family of interactions because different pairs of values for κ and f_{min} may produce the same α_P. In order to define an α_c one must fix either f_{min}, as was done in the previous paragraph, or κ. If one fixes κ then the fundamental critical quantity is $f^c_{min}(T, \kappa)$, which is the maximum error fraction at which Gardner interactions retain attractors. This can be converted into $\alpha_c(T, \kappa)$, the maximal value of α_P for optimal interactions with stability parameter κ that generate attractors.

3.3.2 Performance of Optimal Interactions in ANNs

Although α_c has to be carefully defined the physical questions associated with the quantity are more obvious. The Gardner optimal interactions are optimal for a noiseless network because they maximise the number of patterns that are stored perfectly, but will there be retrieval at finite noise levels? If so, how will the retrieval quality vary with the noise level? These questions are concerned with the robustness of the Gardner optimal interactions. To answer these questions one would like the equivalent of the phase diagram in a T–α plane.

A second set of questions is concerned with how storage errors affect retrieval. When the optimal storage allows for violations of the stability condition at some sites, the question of the existence of attractors becomes non-trivial, even in the absence of noise ($T = 0$). This is particularly true for the Gardner Derrida scheme which, as shall be calculated in section 3.5 yields a field distribution Fig (3.1 b). It might have been the case that while all the violations of the stability condition produced sites at which $\Delta^\mu < \kappa$ they still produced $\Delta^\mu > 0$. This would have meant that the pattern would still be a fixed point of the dynamics at $T = 0$. However, what actually happens is that any site that violates the stability

constraint has $\Delta^\mu < 0$. The first question to be answered in this case is: when do the networks with minimised error number have attractors associated with the (incorrectly) stored pattern? Again we are after the equivalent of a phase diagram in the T–α plane. Equation (3.12) shows that α_P increases as the error fraction is allowed to increase. One may wonder whether this increase in storage is actually useful in retrieval. One may also consider the interplay between the parameters κ and f_{min}, the stability at sites stored correctly and the fraction of sites stored incorrectly. At different temperatures the combination of these two parameters that gives the best performance may differ. Answers to any of the above questions must rely on a method of ascertaining exactly when attractors exist.

3.4 Random Dilution and Dynamics

For fully connected networks no method of analysing the attractors generated by the Gardner interactions has yet been found. The reason is that on the one hand the interactions are not symmetric, precluding the application of statistical mechanics. On the other hand, dynamical equations become very complicated beyond the first step as discussed in section 2.2. However there exists a class of models where the dynamic equations for the overlaps maybe written down explicitly. These are models defined on the randomly diluted lattice[40]. Gardner[41] has shown that optimal interactions may be studied on such a lattice. The randomly diluted lattice is defined by cutting the bonds of a fully connected lattice with probability $1 - C/N$. In this dilution process the bond from site j to i is considered independent of the bond from site i to j, thus models defined on the randomly diluted lattice are always asymmetric.

In order to understand why the dynamics are exactly soluble let us consider first deterministic dynamics. The value at time t of the spin at a site i , $S_i(t)$, is determined by the values at the previous time step, of the spins at the sites to which i is connected.

$$S_i(t) = \mathrm{Sgn}(\sum_j J_{ij} S_j(t-1)) \tag{3.16}$$

In section 2.2 it was pointed out that one could derive a first time step equation for $t = 1$ because $\{S_j(0)\}$ were uncorrelated, thus one could average each $S_j(0)$ independently. However to know the values of the set $\{S_{j(i)}(t-1)\}$ when $t > 1$

(the notation $S_{j(i)}$ indicates a site j to which site i is connected) one must know the values at time $t-2$ of all the spins to which $\{S_{j(i)}\}$ are connected. In this way one traces back to $t=0$ a tree of all the ancestors of site i. If the connectivity is C, there are C^t ancestors in this tree. In the case of full connectivity $C=N$ each site appears N^{t-1} times in the tree. However when the network is diluted different sites will appear in the tree a different number of times. If no site appears more than once in a tree then the tree is said to contain no loops. In this case, within each level of the tree, the spins will be uncorrelated with each other. This is because they share no common ancestors in the tree. Each spin may therefore be averaged independently.

The question remains as to under what conditions the tree of ancestors contains no loops. This has been derived by Derrida and Weisbuch[39]. Mathematically one would like to know when the probability P of the tree containing no loops tends to one. This probability is difficult to calculate exactly. One may consider a related situation with a probability of no loops $\tilde{P}$, where $\tilde{P} < P$. The condition at which $\tilde{P} \to 1$ then is a sufficient condition for $P \to 1$. This related situation is to consider a randomly chosen set of C^t sites. The probability that all the sites are different is

$$\tilde{P} = \prod_{n=1}^{C^t-1} (1 - \frac{n}{N}). \tag{3.17}$$

If $C^t \ll N$ this probability may be expanded to first order in n/N to give

$$\begin{aligned} \tilde{P} &= \exp\left[-\sum_{n=1}^{C^t-1} \frac{n}{N}\right] \\ &= \exp\left[-\frac{C^{2t}-C^t}{2N}\right] \end{aligned} \tag{3.18}$$

One then obtains the sufficient condition for $P \to 1$

$$C^t \ll \sqrt{N} \tag{3.19}$$

or for finite times t

$$C \ll \ln N. \tag{3.20}$$

In the following sections that involve randomly diluted networks, in particular sections 3.4.1, 3.5 and 4.3, we will assume that C obeys (3.20). For the Gardner framework this dictates a modification of the spherical constraint (3.4) to

$$\sum_j J_{ij}^2 = C \ \ \forall i\,. \tag{3.21}$$

3.4.1 Derivation of the Order Parameter Map

We consider the quantity $m_i(t) =< S_i(t)\xi_i^1 >$. Where the angular brackets indicate a thermal average as well as an average over an ensemble of initial conditions. The dynamics is stochastic, namely the probability of the spin at site i taking on the state S_i at time $t+1$ is

$$\Pr[S_i(t+1)] = \frac{1}{1+\exp[-2\beta h_i(t)S_i(t+1)]}. \tag{3.22}$$

The thermal average leads to

$$\begin{aligned} m_i(t+1) &= \langle \tanh\left(\beta\frac{1}{\sqrt{C}}\xi_i\sum_j J_{ij}S_j(t)\right)\rangle \\ &= \langle \int dy \tanh[-\beta y]\delta\left(y-\xi_i^1\sum_j J_{ij}S_j(t)\right)\rangle \\ &= \langle \int \frac{dx\,dy}{2\pi}\tanh[\beta y]\exp ix\left(y-\xi_i^1\sum_j\frac{J_{ij}S_j(t)}{\sqrt{C}}\right)\rangle. \end{aligned} \tag{3.23}$$

The averaging over the ensemble of initial conditions then becomes an average over the $\{S_j(t)\}$. When these spins are uncorrelated, the conditions for which were discussed in 3.4, each S_j can be averaged independently according to the probability distribution.

$$p(S_j(t)) = \frac{1}{2}(1+m(t))\delta(S_j-\xi_j^1)+\frac{1}{2}(1-m(t))\delta(S_j+\xi_j^1) \tag{3.24}$$

This is an average over all configurations$\{S_j(t)\}$ that have overlap $m(t)$ with pattern 1. This average can be justified if one assumes that it is equivalent to averaging over all initial configurations$\{S_j(0)\}$ that have overlap $m(0)$, which is the average that was carried out in the first time step equation. The average is performed by taking the cumulant expansion (2.53) to second order. Using the definition of Δ_i^1 (3.2) and imposing the spherical constraint (3.4), one obtains

$$\begin{aligned} m_i(t+1) &= \int \frac{dx\,dy}{2\pi}\tanh[\beta y]\exp\left(ixy-imx\Delta_i^1-\frac{1-m^2}{2}x^2\right) \\ &= \int \frac{dy}{\sqrt{2\pi}}\tanh[\beta y]\frac{1}{\sqrt{1-m^2}}\exp-\frac{(m\Delta_i^1-y)^2}{2(1-m^2)} \\ &= \int Dy\tanh\left\{\beta\left[m\Delta_i^1+\sqrt{1-m^2}y\right]\right\}, \end{aligned} \tag{3.25}$$

where m on the right hand side is the value at t. The convenience of using (3.24) is now clear: $m_i(t+1)$ becomes a function only of $m(t)$ and Δ_i^1. In order to simplify

further and obtain the order parameter map, one writes the site average as

$$\begin{aligned} m(t+1) &= \frac{1}{N}\sum_i m_i(t+1) \\ &= \int_{-\infty}^{\infty} d\Lambda \rho(\Lambda) \int_{-\infty}^{\infty} Dy \tanh\{\beta[m\Lambda + \sqrt{1-m^2}y]\}, \end{aligned} \tag{3.26}$$

where $\rho(\Lambda)$ is known as the field distribution.

3.5 Calculation of the Field Distribution

We now have to consider the distribution

$$\rho_\kappa(\Lambda) = \langle \delta(\Delta^\nu - \Lambda) \rangle. \tag{3.27}$$

The physical interpretation of this quantity is that it is the probability that on picking a set of interactions from the ensemble that minimise the cost function (3.9), and choosing a site i at which to examine the stability of pattern ν, the value of Δ_i^ν is Λ. The angular brackets in (3.27) then indicate an average over sites and sets of interactions that minimise the cost function. This average can be written out explicitly and (3.27) becomes

$$\rho_\kappa(\Lambda) = \frac{p_\kappa(\Lambda)}{V_C}, \tag{3.28}$$

where

$$p_\kappa(\Lambda) = \frac{1}{N} \int \prod_{ij} dJ_{ij} \left[\sum_i \delta\left(\Lambda - \Delta_i^1\right) \right] \prod_i \lim_{h\to\infty} \exp -h\left(C_i\right) \prod_i \delta\left(\sum_j J_{ij}^2 - C \right). \tag{3.29}$$

Here we are consider a randomly diluted network so that the index j always runs over the sites from which the bond to i has not been cut. V_C is given by equation (3.8) with the spherical constraint suitably modified to (3.21).

$\rho_\kappa(\Lambda)$ depends on the particular realisation of the stored patterns. However we can see from equation (3.29) that it is extensive. This is because we have a sum over i to give an extensive part and the non-extensive part in the numerator should cancel with V_C in the denominator so that we end up with a well-defined probability distribution. Therefore $\rho_\kappa(\Lambda)$ can be assumed to be self-averaging and

the distribution, $\overline{\rho_\kappa(\Lambda)}$, for a typical realisation of the random patterns is given by

$$\overline{\rho_\kappa(\Lambda)} = \ll \rho_\kappa(\Lambda) \gg, \tag{3.30}$$

where the double angular brackets represent the quenched average over the distribution of the random patterns. The quenched average then allows the index i to be dropped, so that J_{ij} becomes J_j, because the contributions to the sum in the square brackets of equation (3.29) will all be identical.

Although we have invoked self-averaging, we still have the problem that the denominator, V_C, depends on the stored patterns in a non trivial way. This will make the quenched averaging rather difficult to carry out. To alleviate this difficult we can introduce replicas in a slightly different way from chapter 2:

$$\overline{\rho_\kappa(\Lambda)} = \ll \lim_{n\to 0} Z^{n-1} p_\kappa(\Lambda) \gg . \tag{3.31}$$

When equation 3.31 is written out in full,

$$\begin{aligned}\overline{\rho_\kappa(\Lambda)} \;=\; & \ll \lim_{n\to 0} \int \prod_{\alpha,j} dJ_j^\alpha \delta\left(\Lambda - \Delta_1^1\right) \\ & \times \prod_{\mu,\alpha} \left[\Theta\left(\Delta_\alpha^\mu - \kappa\right) + \exp(-h)\Theta\left(\kappa - \Delta_\alpha^\mu\right)\right] \prod_\alpha \delta\left(\sum_j (J_j^\alpha)^2 - C\right) \gg,\end{aligned} \tag{3.32}$$

one can see that the probability distribution now takes a form that will eventually be amenable to quenched averaging. In order to develop this form one must exponentiate the delta functions according to (2.58) and use the corresponding form of the Heaviside function

$$\Theta(z-\kappa) = \int_\kappa^\infty \frac{d\lambda}{\sqrt{2\pi}} \int_{-\infty}^\infty \frac{dx}{\sqrt{2\pi}} \exp\left[ix(\lambda - z)\right]. \tag{3.33}$$

The delta function for Δ_1^1 makes it unnecessary to exponentiate the Heaviside functions in the cost function associated with Δ_1^1 because the delta function identifies Δ_1^1 as Λ.

$$\begin{aligned}\overline{\rho_\kappa(\Lambda)} \;=\; & \ll \lim_{n\to 0} \int \prod_{\alpha,j} dJ_j^\alpha \int \frac{dy}{2\pi} \exp iy\left(\Lambda - \Delta_1^1\right) \left[\Theta\left(\Lambda - \kappa\right) + \exp(-h)\Theta\left(\kappa - \Lambda\right)\right] \\ & \times \int \prod_{\mu=2,\alpha} dx_\alpha^\mu \left[\int_\kappa^\infty + \mathrm{e}^{-h} \int_{-\infty}^\kappa\right] \prod_{\mu=2,\alpha} d\lambda_\alpha^\mu \exp i \sum_{\mu=2,\alpha} x_\alpha^\mu \left(\lambda_\alpha^\mu - \Delta_\alpha^\mu\right) \\ & \times \int \prod_{\alpha=2} dx_\alpha^1 \left[\int_\kappa^\infty + \mathrm{e}^{-h} \int_{-\infty}^\kappa\right] \prod_{\alpha=2} d\lambda_\alpha^1 \exp i \sum_{\mu=2,\alpha} \prod_{\alpha=2} x_\alpha^1 \left(\lambda_\alpha^1 - \Delta_\alpha^1\right) \\ & \times \int \prod_\alpha \exp i \sum_\alpha \phi_\alpha \left(\sum_j (J_j^\alpha)^2 - C\right) \gg\end{aligned} \tag{3.34}$$

The quenched averaging may now be performed. The method used is the same as in chapter 2. In the present calculation there are two cases $\mu = 1$ and $\mu \neq 1$. For $\mu \neq 1$ one finds

$$\ll \exp -i\sum_{\alpha} x_{\alpha}^{\mu}\Delta_{\alpha}^{\mu} \gg = \exp -\frac{1}{2}\sum_{\alpha\beta} x_{\alpha}^{\mu}x_{\beta}^{\mu}\frac{1}{C}\sum_{j} J_{j}^{\alpha}J_{j}^{\beta} = \exp -\frac{1}{2}\sum_{\alpha\beta} x_{\alpha}^{\mu}x_{\beta}^{\mu}q^{\alpha\beta}, \tag{3.35}$$

where $q^{\alpha\beta}$ is the Edwards Anderson order parameter in the space of interactions

$$\begin{aligned} q^{\alpha\beta} &= \frac{1}{C}\sum_{j} J_{j}^{\alpha}J_{j}^{\beta} \quad \text{for} \quad \alpha \neq \beta \\ &= 1 \qquad\qquad\qquad\quad \text{for} \quad \alpha = \beta. \end{aligned} \tag{3.36}$$

For $\mu = 1$ we have

$$\ll \exp -i\left(y\Delta_{1}^{1} + \sum_{\alpha\neq 1} x_{\alpha}^{1}\Delta_{\alpha}^{1}\right) \gg = \exp -\frac{1}{2}\left(y^2 + 2y\sum_{\alpha\neq 1} x_{\alpha}^{1}q^{1\alpha} + \sum_{\alpha\beta\neq 1} x_{\alpha}^{\mu}x_{\beta}^{\mu}q^{\alpha\beta}\right). \tag{3.37}$$

$\overline{\rho_{\kappa}(\Lambda)}$ may now be written as

$$\begin{aligned} \overline{\rho_{\kappa}(\Lambda)} &= \lim_{n\to 0}\int \prod_{\alpha}\frac{d\phi_{\alpha}}{2\pi}\prod_{\alpha\neq\beta}\frac{d\epsilon^{\alpha\beta}dq^{\alpha\beta}}{2\pi} \\ &\times \exp\left\{-iC\sum_{\alpha}\phi_{\alpha} + iC\sum_{\alpha\neq\beta}\epsilon^{\alpha\beta}q^{\alpha\beta}\right. \\ &\qquad +C\ln\left[\int\prod_{\alpha}dJ^{\alpha}\exp\left(i\sum_{\alpha}\phi_{\alpha}(J^{\alpha})^2 - i\sum_{\alpha\neq\beta}\epsilon^{\alpha\beta}J^{\alpha}J^{\beta}\right)\right] \\ &\qquad +(\alpha C-1)\ln\left[\int\prod_{\alpha}\left[\int_{\kappa}^{\infty} + e^{-h}\int_{-\infty}^{\kappa}\right]\prod_{\alpha}d\lambda_{\alpha}\right. \\ &\qquad\qquad \left.\left.\times \exp i\sum_{\alpha}x_{\alpha}\lambda_{\alpha} - \frac{1}{2}\sum_{\alpha\beta}q^{\alpha\beta}x_{\alpha}x_{\beta}\right]\right\} \\ &\times \int\prod_{\alpha}\left[\int_{\kappa}^{\infty} + e^{-h}\int_{-\infty}^{\kappa}\right]\prod_{\alpha}d\lambda_{\alpha} \\ &\times \int\frac{dy}{2\pi}\exp\left\{iy\Lambda - \frac{1}{2}y^2 - y\sum_{\alpha\neq 1}x_{\alpha}q^{1\alpha} - \frac{1}{2}\sum_{\alpha\beta\neq 1}q^{\alpha\beta}x_{\alpha}x_{\beta} + i\sum_{\alpha}x_{\alpha}\lambda_{\alpha}\right\} \\ &\times \left[\Theta(\Lambda-\kappa) + e^{-h}\Theta(\kappa-\Lambda)\right] \end{aligned} \tag{3.38}$$

For large C the integrals over ϕ, ϵ, q may be evaluated by the method of steepest descent. As usual replica symmetry will be assumed at the saddle point. From this saddle point we will develop an expression for q in terms of α and h. However

when the limit $n \to 0$ is taken the contribution from the exponential will vanish as it is of the form $\exp(nCG)$. $\overline{\rho_\kappa(\Lambda)}$ will then be given by the contribution from the second part of (3.38).

First we will deal with the saddle point integrals. Assuming replica symmetry

$$\begin{aligned} q^{\alpha\beta} &= q \quad \text{for} \quad \alpha \neq \beta \\ \epsilon^{\alpha\beta} &= \epsilon \quad \text{for} \quad \alpha \neq \beta \\ \phi^\alpha &= \phi \quad \forall \alpha, \end{aligned} \tag{3.39}$$

making the transformations

$$\begin{aligned} i\phi &= -\frac{1}{2}E \\ i\epsilon &= -\frac{1}{2}F, \end{aligned} \tag{3.40}$$

and taking the $n \to 0$ limit one finds

$$\begin{aligned} G &= \frac{E}{2} + \frac{F}{2} + \ln(\pi i) - \frac{\ln(E+F)}{2} + \frac{F}{2(E+F)} \\ &+ \alpha \int \mathcal{D}z \ln\left[\left[\int_\kappa^\infty + e^{-h}\int_{-\infty}^\kappa\right] \frac{d\lambda}{\sqrt{2\pi(1-q)}} \exp -\frac{1}{2}\frac{(\lambda + z\sqrt{q})^2}{1-q}\right]. \end{aligned} \tag{3.41}$$

The saddle point equations for E and F have solution

$$E = \frac{1-2q}{(1-q)^2} \tag{3.42}$$

$$F = \frac{q}{(1-q)^2}. \tag{3.43}$$

Inserting these in Eq (3.41) one obtains

$$\begin{aligned} G &= \frac{1}{2}\ln(1-q) + \frac{1}{2(1-q)} + \alpha \int Dz \\ &\times \ln\left[\left[\int_\kappa^\infty + e^{-h}\int_{-\infty}^\kappa\right] \frac{d\lambda}{\sqrt{2\pi(1-q)}} \exp -\frac{1}{2}\frac{(\lambda + z\sqrt{q})^2}{1-q}\right]. \end{aligned} \tag{3.44}$$

The minimum of this function will yield an equation relating α, q, h. However in order to find the ground state of the cost function the limit $h \to \infty$ must be taken. This limit is related to the limit $q \to 1$ that signals that the volume of interactions contributing to the partition function (3.8) has shrunk to a point. In order to take the two limits simultaneously, one defines a parameter x such that

$$h = \frac{x^2}{2(1-q)}. \tag{3.45}$$

One then replaces h in (3.44) with (3.45) and expands for $q \to 1$. The leading order terms are singularities of form $1/(1-q)$. Retaining only these terms one obtains

$$G = \frac{1}{2(1-q)}\left(1 - \alpha\int_{\kappa-x}^{\kappa} Dz(\kappa - z)^2 - \alpha x^2 \int_{-\infty}^{\kappa-x} Dz\right). \tag{3.46}$$

In order to extremise (3.46) with respect q it is simplest to write q in terms of x through (3.45) and extremise with respect to x. One then obtains the results listed in sec 3.2 that were derived by Gardner and Derrida[38].

We now turn to the part of Equation 3.38 that will yield the form of the probability distribution. After performing the y integral, the integrals over x_α and taking the $n \to 0$ limit one obtains

$$\begin{aligned}\overline{\rho_\kappa(\Lambda)} &= \frac{1}{\sqrt{2\pi(1-q)}}\left[\Theta(\Lambda-\kappa) + \mathrm{e}^{-h}\Theta(\kappa-\Lambda)\right] \\ &\quad \times \int Dz \frac{\exp -\frac{(z\sqrt{q}+\lambda)^2}{2(1-q)}}{\left[\int_\kappa^\infty + \mathrm{e}^{-h}\int_{-\infty}^{\kappa}\right]\frac{d\lambda}{\sqrt{2\pi(1-q)}}\exp -\frac{(z\sqrt{q}+\Lambda)^2}{2(1-q)}}.\end{aligned} \tag{3.47}$$

In order to take the $q \to 1$ limit one must evaluate the forms which the denominator of the z integral in Eq. 3.47 takes. To do this one makes use of equation (3.45) and the expansion of the gauss error function for large argument:

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x dz \exp(-z^2) = 1 - \frac{\exp(-x^2)}{\sqrt{\pi}x} + \ldots. \tag{3.48}$$

The denominator then takes the forms

$$\begin{cases} 1 & \text{for} \quad z < -\kappa \\ \sqrt{\frac{1-q}{2\pi}}\frac{1}{(z+\kappa)}\exp -\frac{(\kappa+z)^2}{2(1-q)} & \text{for} \quad -\kappa < z < x-\kappa \\ \mathrm{e}^{-h} & \text{for} \quad z > x-\kappa \end{cases} \tag{3.49}$$

which give

$$\begin{aligned}\overline{\rho_\kappa(\Lambda)} &= \frac{1}{\sqrt{2\pi(1-q)}}\left[\Theta(\Lambda-\kappa) + \exp\left(-\frac{x^2}{2(1-q)}\right)\Theta(\kappa-\Lambda)\right] \\ &\quad \times \left\{\int_{-\infty}^{-\kappa} Dz \exp -\frac{(z+\Lambda)^2}{2(1-q)}\right. \\ &\quad + \int_{-\kappa}^{x-\kappa} Dz\sqrt{\frac{2\pi}{(1-q)}}(\kappa+z)\exp\left(\frac{-(z+\Lambda)^2+(z+\kappa)^2}{2(1-q)}\right) \\ &\quad \left. + \int_{x-\kappa}^{\infty} Dz \exp\frac{-(z+\lambda)^2+x^2}{2(1-q)}\right\}.\end{aligned} \tag{3.50}$$

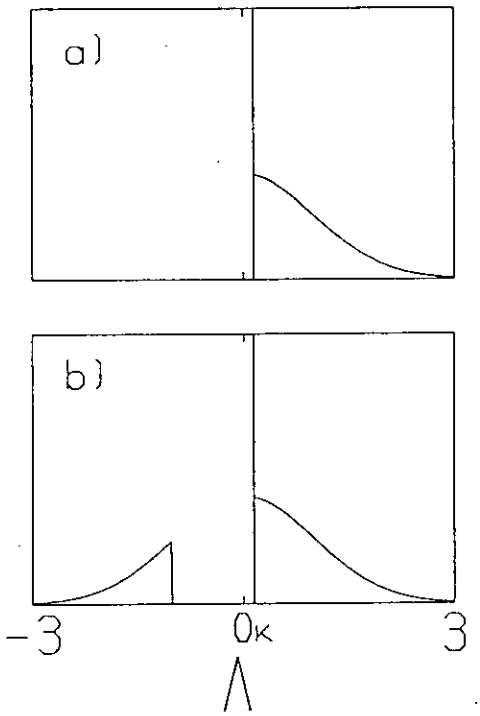

Figure 3.1: (a) The probability distribution of the stability parameter in a network optimised with $\kappa = 0.1$ and zero error fraction. (b) The probability distribution for a network optimised with $\kappa = 0.1$ and error fraction $f = 0.15$.

In order to evaluate these integrals one must identify certain representations of the dirac delta function.

$$\lim_{q\to 1} \frac{1}{\sqrt{2\pi(1-q)}} \exp -\frac{(z+\Lambda)^2}{2(1-q)} = \delta(z+\Lambda), \tag{3.51}$$

$$\lim_{q\to 1} \Theta(z+\kappa)\Theta(\Lambda-\kappa)\frac{(z+\kappa)}{(1-q)} \exp \frac{-(z+\Lambda)^2+(z+\kappa)^2}{2(1-q)} = \Theta(z+\kappa)\delta(\Lambda-\kappa). \tag{3.52}$$

Finally one obtains

$$\overline{\rho_\kappa(\Lambda)} = \frac{1}{\sqrt{2\pi}} \exp(-\frac{\Lambda^2}{2}) \left[\Theta(\Lambda-\kappa) + \Theta(\kappa-x-\Lambda)\right] + \delta(\Lambda-\kappa) \int_{\kappa-x}^{\kappa} Dz, \tag{3.53}$$

which is the field distribution function shown in Fig 3.1.

In this section a full derivation of the field distribution has been given. On the way to deriving this distribution the Gardner -Derrida results (3.12-3.14) have been obtained. The calculation is rather lengthy and involved. Wong and Sherrington[42,43] have recently presented a method of calculation that provides a simple recipe for deriving α_P and $\rho(\Lambda)$ for an arbitrary cost function. Using their recipe would shorten the calculation considerably, at the sacrifice (or relief) of not working from first principles.

3.5.1 Physical Significance of the Field Distribution

In Fig 3.1 there are two qualitatively different forms. The first is for $x \to \infty$. Here we have a delta function at $\Lambda = \kappa$ and a Gaussian tail for $\Lambda > \kappa$. Clearly this is the regime where the constraints 3.2 are all obeyed. Mathematically this has occurred because $x \to \infty$ indicates that the ground state of the cost function (which is an interaction configuration where all constraints are obeyed) has been singled out before the volume V_C shrinks to a point. Therefore when V_C shrinks to a point it indicates that we have reached an α where the constraints can no longer all be satisfied. When x is finite it indicates that the ground state has been singled out at a rate proportional to the shrinking of the volume V_C. The interaction configuration singled out is then one that stores the patterns with error fraction f_{min} given by (3.13).

The striking feature of fig 3.1b is the sharp gap, $\kappa - x < \Lambda < \kappa$, in the field distribution. It is a result of the particular cost function employed. Fig 3.1 illustrates that for a distribution of Gaussian form, the sites at which errors in the stability parameter occur violate the stability in a maximal way. In other words the fields at those sites, when the network is in a pattern, are as large as possible and opposite in sign to the pattern. The corresponding cost function (3.9) only penalises the number of stability errors, not the size. Conversely, if one desired a different field distribution at the same storage level α as figure 3.1, then the fraction of stability errors would have to increase.

3.6 Analysis of the Order Parameter Map

The order parameter map given by Eq.(3.26) is quite general and can be studied without specifying either the form or the parametrisation of $\rho(\Lambda)$. We first proceed to analyse the fixed point structure of the map without specifying $\rho(\Lambda)$. In particular, one can obtain conditions on $\rho(\Lambda)$ for various types of transitions from retrieval to no retrieval, which would, in turn, yield critical values of the parameters of the model. All these questions go back to the dependence of the fixed point structure of Eq. (3.26) and of the stability of these fixed points on the properties of $\rho(\Lambda)$ and on the noise T. Eq. (3.26) describes the parallel dynamics of a dilute

network. Here we are primarily concerned with the structure of the fixed points of (3.26) so that the dynamical features implied by this equation will not be considered in detail. The fixed points are the same as those of asynchronous dynamics, which is perhaps more realistic and more robust. The fixed points, $m^*(\kappa, T)$, of Eq. 3.26 are given by the roots of $g(m, \kappa, \beta)$ where

$$g(m,\kappa,\beta) = f(m,\kappa,\beta) - m \tag{3.54}$$

$$f(m,\kappa,\beta) = \int_{-\infty}^{\infty} Dy \int_{-\infty}^{\infty} d\Lambda \rho(\Lambda) \tanh\{\beta[m\Lambda + \sqrt{1-m^2(t)}y]\}, \tag{3.55}$$

with $T = 1/\beta$. This function is odd so we can focus our discussion on $m > 0$. For an attractor, we require $g'(m^*) < 0$ where $g(m^*) = 0$, and for an unstable fixed point $g'(m^*) > 0$.

There are three fixed point structures: $m = 0$ is always a fixed point, due to the antisymmetry of $g(m)$. It may be the only fixed point or there may be either one or two additional fixed points with $m > 0$.

1. When $m = 0$ is the only fixed point it must be stable (Fig. 3.2(1)). This is the situation of no retrieval.

2. When there is only one additional positive fixed point it must be stable and $m = 0$ must be unstable (Fig 3.2 (2)). The domain of attraction of the positive m fixed point is then unity and the retrieval is 'wide'.

3. Finally, when there are two additional positive fixed points, the $m = 0$ fixed point and the fixed point with the highest absolute value of m are stable and the fixed point in the middle is unstable (Fig. 3.2 (3)). The intermediate fixed point delimits the basin of attraction of the high-m (retrieval) fixed point and the retrieval is 'narrow'.

As the structure of g changes with the noise level, at fixed storage parameters (e.g. κ or α), the fixed point structure may change between any two of the three alternatives. For purposes of associative retrieval, we are primarily interested in changes between retrieval dynamics and no retrieval, namely (2)→(1) or (3)→(1). The first transition is continuous, in the sense that the finite m fixed point disappears as $m \to 0$ continuously. The second transition is discontinuous, as the two finite m fixed points coalesce and disappear at finite m. In addition there is

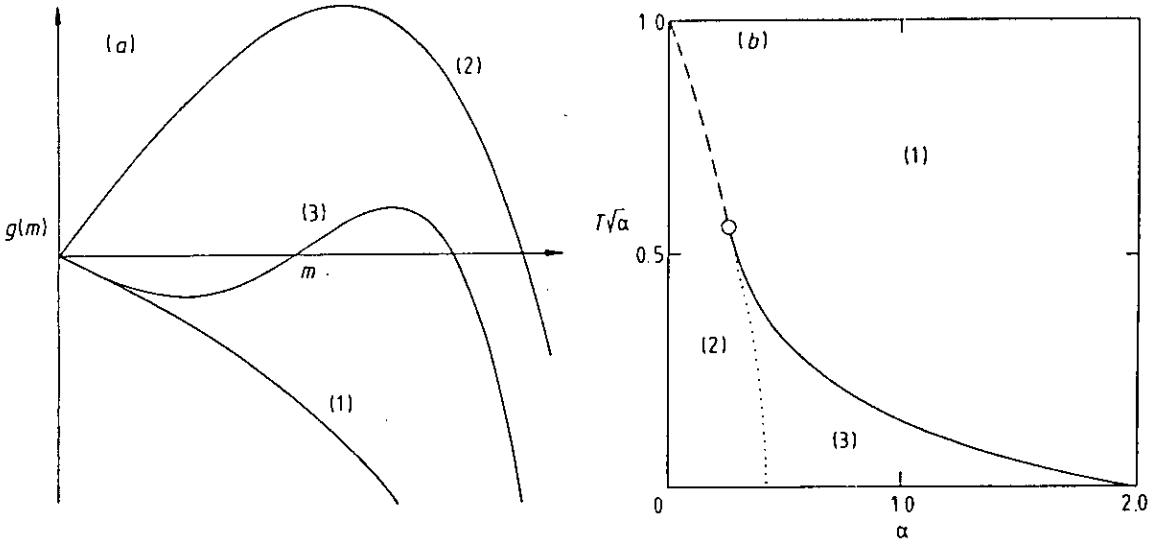

Figure 3.2: (a)Schematic sketch of the three possible fixed point structures in g: (1) no retrieval, (2) wide retrieval, (3) narrow retrieval. (b) Schematic representation of different retrieval regions in the phase diagram

a transition (2)→(3), through which the retrieval attractor survives, but its basin of attraction, which is unity in (2), is reduced in (3).

3.6.1 The Continuous Transition and the Tricritical Point

A necessary condition for a continuous transition is

$$g'(0) = 0, \tag{3.56}$$

which, using Eq. 3.55, reads:

$$\langle \Lambda \rangle = \frac{1}{\beta \int_{-\infty}^{\infty} Dy \ \mathrm{sech}^2(\beta y)} \tag{3.57}$$

where the angular brackets denote an average over the distribution $\rho(\Lambda)$:

$$\langle F(\Lambda) \rangle = \int_{-\infty}^{\infty} d\Lambda \rho(\Lambda) F(\Lambda). \tag{3.58}$$

However this condition is not sufficient since $g' = 0$ may also indicate a transition (2)→(3). This issue will be discussed fully below. The additional condition for a continuous transition is

$$g'''(m = 0) < 0, \tag{3.59}$$

to ensure a stable fixed point at arbitrarily small m as T is raised to the transition value. Whereas for

$$g'''(m=0) > 0, \tag{3.60}$$

$m = 0$ is a stable fixed point and the retrieval fixed point disappears discontinuously, when the two roots at finite positive m coalesce and disappear. The two cases are separated by the point at which

$$g'''(m=0) = 0, \tag{3.61}$$

when all five (positive and negative) roots of $g(m)$ coalesce at $m = 0$. The special point is determined by the two simultaneous equations (3.57) and (3.61). It is the analogue of a thermodynamic tricritical point. When (3.55) is substituted in (3.61), one has

$$3\langle\Lambda\rangle - \langle\Lambda^3\rangle = 0. \tag{3.62}$$

Note that this condition is temperature independent, which implies that it must reflect itself also in the $T = 0$ dynamics (see section 3.6.4).

3.6.2 Transitions near the 'Tricritical' Point

The line of discontinuous transitions is defined by the appearance of a double zero of g at non-zero m. This occurs when the maximum in Fig. 3.2(3) crosses $g = 0$. Thus, in addition to the equation $g(m) = 0$, $g'(m) = 0$ must be satisfied, both at $m \neq 0$. The two equations can be written as

$$\frac{f(m,\kappa,\beta)}{m} = 1 \tag{3.63}$$

$$\frac{df(m,\kappa,\beta)}{dm} = 1 \tag{3.64}$$

These are two equations for the three unknowns m, β and κ, whose solution is the equation for the line of discontinuous transitions, $\beta_c(\kappa)$, and for the discontinuity in the retrieval amplitude $\Delta m(\kappa)$.

To investigate the neighbourhood of the tricritical point, where m is small on both the continuous and the discontinuous sides, we expand Eq. 3.55 for small m, up to fifth order. Analysing, in the Appendix, the expanded map we find that the retrieval amplitude m grows as $\sqrt{\Delta T}$ as one goes below the line of continuous

transitions. The line given by Eq. 3.57 can be evaluated beyond the tricritical point and all the way to $T = 0$. In this region, where $g'''(0) > 0$, as one lowers T or raises β from the state of no retrieval, two non zero roots appear before (at higher temperature than) $g'(0)$ vanishes. This is the discontinuous transition. Thus, two lines start at the tricritical point. They are the dotted and the dashed lines in Fig. 3.2b. The first is the line of discontinuous transitions and the second is the continuation to $T = 0$ of the line of instability of the $m = 0$ fixed point. Near the tricritical point the two lines can be computed and compared, see the appendix. The result is that the two lines start with equal slopes and diverge at second order.

3.6.3 The Transition From Wide to Narrow Retrieval

The dotted curve in figure 3.2 is the continuation of the line of continuous transitions past the tricritical point. Its dynamical significance is that it separates two sub-regions: to its left there is 'wide retrieval', marked (2) in figure 3.2(b), and to its right is a region of 'narrow retrieval', marked(3).

The dotted curve is in a region where $g'''(0) > 0$. Thus above (and near) this curve $g'(0) < 0$ and the form of g is that of (3), with the middle unstable fixed point approaching $m = 0$, as one approaches the line $g'(0) = 0$. This unstable fixed point delimits the basin of attraction of stable high m fixed point. One may then say that the basin of attraction of this stable fixed point reaches the full interval (0-1) continuously as the dotted curve is crossed and the transition from 'narrow' to 'wide' retrieval is made.

3.6.4 Summary of Transitions

If one considers raising the temperature from $T = 0$ then then there are three possibilities for the transition sequence to no retrieval

1. (2) $\rightarrow$ (1) wide to no retrieval
2. (2) $\rightarrow$ (3) $\rightarrow$ (1) wide to narrow to no retrieval

3. (3) → (1) narrow to no retrieval

In this subsection the conditions for each of these transitions will be stated concisely.

If at $T = 0$, $g'(0) > 0$ is satisfied, which is equivalent to

$$\langle \Lambda \rangle > \sqrt{\frac{\pi}{2}}, \tag{3.65}$$

then one is guaranteed an attractor at $T = 0$ and the retrieval will be wide. As the temperature is raised the transition to no retrieval will occur by going directly from wide to no retrieval if (3.59) is satisfied at $T = 0$. This condition becomes

$$3\langle \Lambda \rangle - \langle \Lambda^3 \rangle < 0. \tag{3.66}$$

If (3.66) is not satisfied but (3.65) is then the transition sequence will be wide to narrow to no retrieval as the temperature is raised.

In the case that (3.65) is not satisfied then one is not guaranteed an attractor even at $T = 0$. If in addition (3.66) is satisfied then there are certainly no attractors. However when (3.65) and (3.66) are both not satisfied, to determine whether an attractor does exist at $T = 0$ one must seek numerically solutions of (3.63) and (3.64) as $\beta \to \infty$. If indeed one finds a solution then one will have narrow retrieval at $T = 0$ and the transition will be discontinuous from narrow to no retrieval as the temperature is raised. In this way all the qualitative information about the transitions can be obtained from the zero temperature map. Remarkably, nearly all this information, apart from the existence of a narrow retrieval attractor, is contained in the moments of $\rho(\Lambda)$.

3.7 Numerical Results for Overlap Dynamics

3.7.1 Noiseless Dynamics of Errorless Optimal Network

We now proceed to use the general theory of order parameter maps presented in the previous section to investigate the performance of optimal interaction matrices.

First we can recover the results of Gardner[41], by studying the structure of the basins of attraction in a noiseless ($T = 0$) errorless optimally connected network. In this case the order parameter map (3.26) is

$$m(t+1) = \frac{1}{2}\left(1+\mathrm{erf}\left(\frac{\kappa}{\sqrt{2}}\right)\right)\mathrm{erf}\left(\frac{m(t)\kappa}{\sqrt{2(1-m^2(t))}}\right) + \int_{\kappa}^{\infty} D\Lambda \mathrm{erf}\left(\frac{m(t)\Lambda}{\sqrt{2(1-m^2(t))}}\right). \quad (3.67)$$

For the noiseless network the temperature is fixed at zero, so that we can only consider the transitions effected by increasing α_P (decreasing κ). A transition from wide to narrow retrieval occurs when

$$\langle\Lambda\rangle = \frac{\kappa}{2}\left(1+\mathrm{erf}\left(\frac{\kappa}{\sqrt{2}}\right)\right) + \frac{\exp\left(-\frac{\kappa^2}{2}\right)}{\sqrt{2\pi}} = \sqrt{\frac{\pi}{2}}. \quad (3.68)$$

When this equation is solved for κ and the κ value converted into α_P one finds that the transition from wide to narrow retrieval occurs at $\alpha_P = 0.42$, which is the transition value of Gardner[41]. Above this value of α the m=0 fixed point becomes an attractor at the expense of the basin of retrieval. By searching for fixed points of (3.67) numerically one finds that the transition from narrow to no retrieval occurs at $\alpha_P = 2.0$ as expected.

3.7.2 Noisy of Errorless Optimal Network

Having recovered the results of Gardner[41], we can make the first new application of the general results for order parameter maps: to extend Gardner's study[41] to the case of the noisy network ($T > 0$). In this case the order parameter map (3.26) becomes, after suitable rotation in the plane of integration,

$$m(t+1) = \frac{1}{2}\int_{-\infty}^{\infty} Dy\left[1+\mathrm{erf}\left(\frac{\kappa+my}{\sqrt{2(1-m^2)}}\right)\right]\tanh(\beta y) + \frac{1}{2}\left[1+\mathrm{erf}\left(\frac{\kappa}{\sqrt{2}}\right)\right]\int_{-\infty}^{\infty} Dy\tanh\{\beta(m\kappa+\sqrt{1-m^2}y)\}. \quad (3.69)$$

The fixed points of (3.69) were studied numerically and the results are summarised in figures 3.3 and 3.4. In fig 3.3(a) we present the phase diagram on which are

drawn lines separating the different retrieval regimes. Recall that α is determined by the $q = 1$ condition as a function of the stability parameter κ. The lines are more fundamentally lines of $T_c(\kappa)$, which is why κ is given on the top horizontal axis. In order that the full range of temperature can be viewed in 3.3a the temperature axis has been rescaled as $T\sqrt{\alpha}$. The first curve of interest is the continuous transition line. Using equation (3.57) we find this line as

$$\frac{\kappa}{2}\left(1+\mathrm{erf}\left(\frac{\kappa}{\sqrt{2}}\right)\right)+\frac{\exp\left(-\frac{\kappa^2}{2}\right)}{\sqrt{2\pi}}=\frac{1}{\beta\int Dy\mathrm{sech}^2(\beta y)} \tag{3.70}$$

from which one can construct the dotted curve in Figure 3.3a. Moreover, expanding the left-hand side of (3.70) for large κ and the right hand side for small β one finds the relation $T_c = \kappa$. Since the correction to the left hand side is exponentially small and to the right-hand side it is of relative magnitude β^2, the relation holds over a wide region. In figure 3.3b the exact relationship is plotted. One can see that above the tricritical point, which is the region corresponding to the continuous transition line, the linear relationship is well obeyed. This relation vindicates the designation of κ as a stability parameter, the stability being with respect to temperature.

The condition given by (3.62) for the tricritical point becomes

$$\left(1+\mathrm{erf}\left(\frac{\kappa}{\sqrt{2}}\right)\right)(3\kappa-\kappa^3)+\sqrt{\frac{2}{\pi}}\exp\left(-\frac{\kappa^2}{2}\right)(1-\kappa^2)=0. \tag{3.71}$$

Equations (3.70) and (3.71) can be solved numerically to give the tricritical point as $\kappa_{tr} = 1.700, \beta_{tr} = 0.909$ which corresponds to $\alpha_{tr} = 0.258, T_{tr} = 1.100$ in the phase diagram, figure 3.3a. Beyond the tricritical point the line of continuous transitions becomes the phase separation curve between wide and narrow retrieval, and appears as the dotted line in fig 3.3a. It reaches $\alpha = 0.42$ at $T = 0$.

The discontinuous transition line (full curve in fig 3.3a) was calculated numerically by solving (3.64, 3.64). The discontinuous transition to no retrieval occurs in general at high α and low T values, whereas the continuous transition occurs at low α and high T values.

In Fig. 3.4 we show the retrieval quality m vs T for several values of $\alpha_c(\kappa)$, i.e. several values of κ. For high κ (low $\alpha_c(\kappa)$), m vanishes continuously at the transition. Hence the horizontal intercepts of the retrieval quality curves correspond

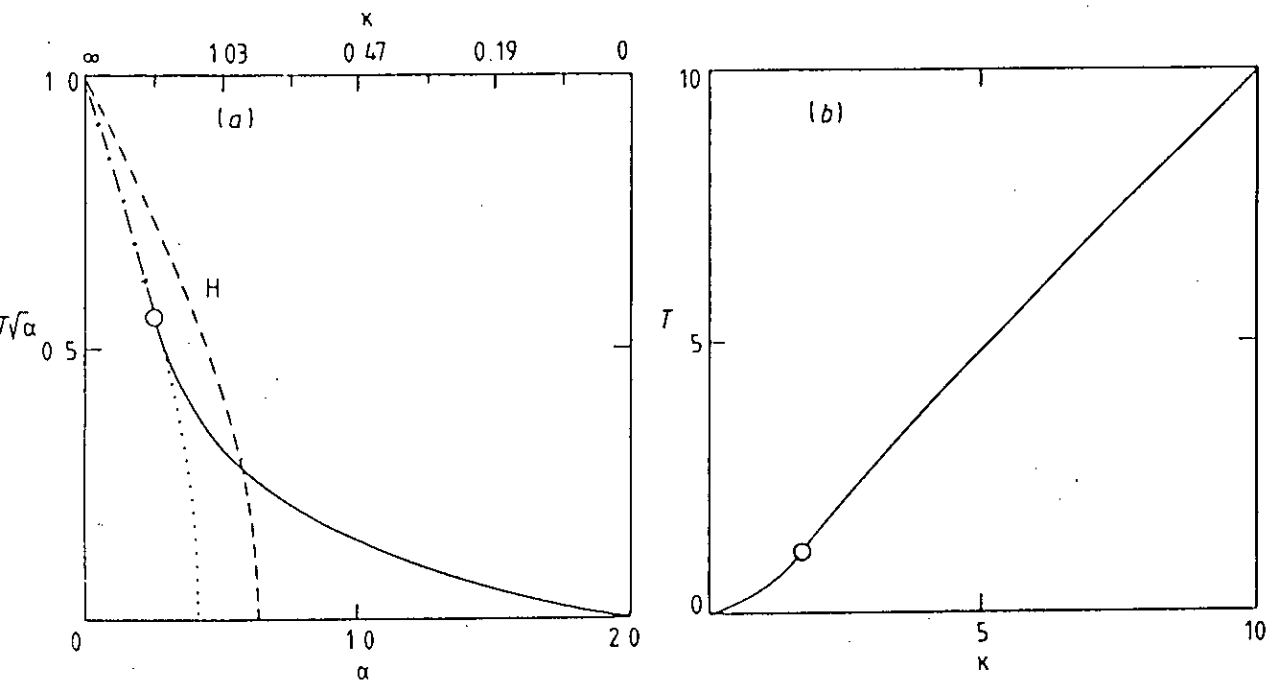

Figure 3.3: (a)Phase diagram, α–T plane, for dilute, optimally connected network. Full part of curve: discontinuous transition; dashed part: continuous transition; dotted part: transition from 'wide retrieval' (100 % basin of attraction) to 'narrow retrieval' (less than 100 %). The small circle is tricritical point. Curve (H) is phase separation curve for randomly dilute Hopfield network[40], with couplings normalised according to (3.4). The top horizontal axis gives the value of κ which corresponds to the α on the bottom axis.
(b)Transition temperature vs κ. Conventions as in Fig. 3.3(a). Note the relation $T=\kappa$ over almost the entire range of continuous transitions.

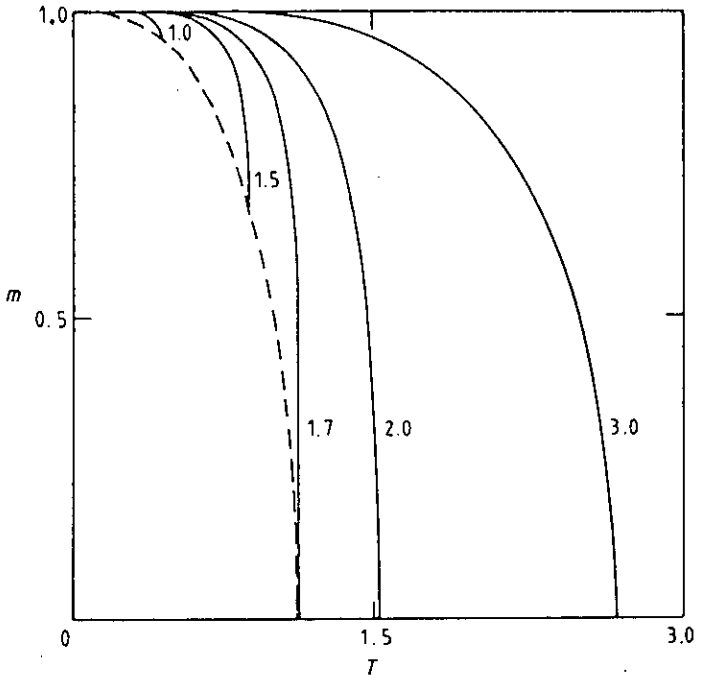

Figure 3.4: Retrieval quality m vs noise T for several values of κ (and hence of $\alpha_c(\kappa)$), marked on figure.

to points on the phase line separating wide retrieval from no retrieval. This transition is continuous up to the tricritical storage $\alpha_{tr} = 0.26(\kappa_{tr} = 1.7, T_{tr} = 1.1)$. For lower κ, m vanishes discontinuously, and the broken curve shows m at the discontinuous transition. The touching points of the retrieval quality curves with the broken curve correspond to points on the phase line separating narrow retrieval from no retrieval. It should be noted that while for low α_c (high κ) m vanishes continuously at the transition, wherever both networks retrieve, the high κ network gives better retrieval quality. For example at $T = 1.5$ the two curves that correspond to retrieval in the figure are $\kappa = 2.0$ and $\kappa = 3.0$. Their respective retrieval qualities are $m = 0.29$ and $m = 0.94$.

3.7.3 Comparison with the Hopfield Model

The phase line for the randomly diluted Hopfield network, first studied by Derrida, Gardner and Zippelius[40] is also plotted in fig 3.3a. The field distribution for this network can be obtained by taking the $a \to 0$ limit of that calculated in section 2.2:

$$\rho(\Lambda) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(\Lambda - \frac{1}{\sqrt{\alpha}})^2}{2}\right). \tag{3.72}$$

One sees that the field distribution for the Hopfield model is a broad Gaussian distribution. This is rather different in character from the distributions for the optimal networks in 3.1. In view of this Abbott and Kepler[44] have proposed that the Hopfield model and the optimal network belong to different universality classes. Wong and Sherrington[45] have also recently speculated that the mechanisms for recall may be fundamentally different.

The order parameter map for the randomly diluted Hopfield is given by equation (3.55) with

$$f(m) = \int Dy \tanh\left\{\beta\left[\sqrt{\alpha}y + m\right]\right\}. \tag{3.73}$$

The line of continuous transitions is given by equation 3.57 as

$$\frac{1}{\sqrt{\alpha}} = \frac{1}{\beta \int_{-\infty}^{\infty} Dy \, \mathrm{sech}^2(\beta y)} \tag{3.74}$$

For the randomly diluted Hopfield model there is in fact no tricritical point. This is because

$$3\langle\Lambda\rangle - \langle\Lambda^3\rangle = -\frac{5}{\alpha^{3/2}} \tag{3.75}$$

so 3.62 is never satisfied. According to the rules stated in section 3.6.4 this implies that for $\langle\Lambda\rangle > \sqrt{\pi/2}$ the retrieval will be wide at $T = 0$ and the transition as the temperature is raised will be continuous from wide to no retrieval. Whereas for $\langle\Lambda\rangle < \sqrt{\pi/2}$ there will be no retrieval even at $T = 0$. The $T = 0$ critical value of α is therefore $\alpha_c = 2/\pi = 0.64$.

In figure 3.3a this transition line is superimposed on the optimal phase diagram. One sees that while the Gardner prescription has a higher storage capacity for attractors at low temperatures, the Hopfield network stores more attractors at high temperatures. Recently it has been shown that the Hopfield network in fact has the largest storage capacity for attractors at sufficiently large temperatures [43]. At very high temperatures there is a common intercept in the transition lines at $T\sqrt{\alpha} = 1$.

3.7.4 Retrieval in Networks with Errors in Storage

We now consider the implications for attractors of a network of the Gardner-Derrida type discussed in the previous section. The formula for $\alpha_P(\kappa, f_{min})$ of

Gardner-Derrida shows a dramatic increase in the number of patterns that can be stored at each site if a fraction f_{min} of the patterns are stored incorrectly. For example, for f_{min}=0.1, $\alpha_P \approx 5.7$. An error at a site implies that the imposed stability condition Eq. 3.2 is violated at that site. From Fig. 3.1 we can see that for the Gardner-Derrida cost function the constraint violations ($\Lambda < \kappa$) always give stability violations ($\Lambda < 0$), so that we are not guaranteed an attractor, even at $T = 0$, as would be the case if all $0 < \Lambda < \kappa$. If the maximal number of patterns per neural connection stored in a perceptron with error fraction f_{min} is given by $\alpha_P(\kappa, f_{min})$, one would like to know what is α_c (see section 3.3.1) for field distributions of the form Fig. 3.1.

The order parameter map (3.26) becomes, after suitable rotation in the plane of integration

$$\begin{aligned} m(t+1) &= \frac{1}{2}\int_{-\infty}^{\infty} Dy \left[\mathrm{erf}\left(\frac{\kappa - x - my}{\sqrt{2(1-m^2)}}\right) - \mathrm{erf}\left(\frac{\kappa - my}{\sqrt{2(1-m^2)}}\right)\right] \tanh(\beta y) \\ &+ \frac{1}{2}\left[\mathrm{erf}\left(\frac{\kappa}{\sqrt{2}}\right) - \mathrm{erf}\left(\frac{\kappa - x}{\sqrt{2}}\right)\right] \int_{-\infty}^{\infty} Dy \tanh\{\beta(m\kappa + \sqrt{1-m^2}y)\}. \end{aligned} \tag{3.76}$$

Figure 3.5 presents the results of solving the fixed point equations of Eq. 3.76. Each point in the α–κ plane corresponds to a value of f_{min} where $\alpha = \alpha_P(\kappa, f_{min})$. Several lines of constant f_{min} are drawn in the figure. The dashed lines are lines of constant critical temperature. A point on such a line gives:

1. The temperature at which the retrieval attractor of the map Eq. 3.76 is destabilised for the particular values of α and κ .

2. The error fraction, $f_{min}(\kappa, \alpha)$, corresponding to optimal perceptron storage.

No distinction is made in the figure between continuous and discontinuous transitions to no retrieval. From Fig. 3.5 one can read off $\alpha_c(\kappa, T)$ which is the value of α beyond which storage with errors with κ fixed no longer gives attractors at temperature T. Note that to each $\alpha_c(\kappa, T)$ corresponds a value of f_{min}, so once the critical α has been crossed there are no attractors because at fixed κ an increase in α leads to an increase of f_{min}. To obtain $\alpha_c(\kappa, T)$ one draws a line parallel to the α axis, at the chosen value of κ, to find the intersection with the chosen

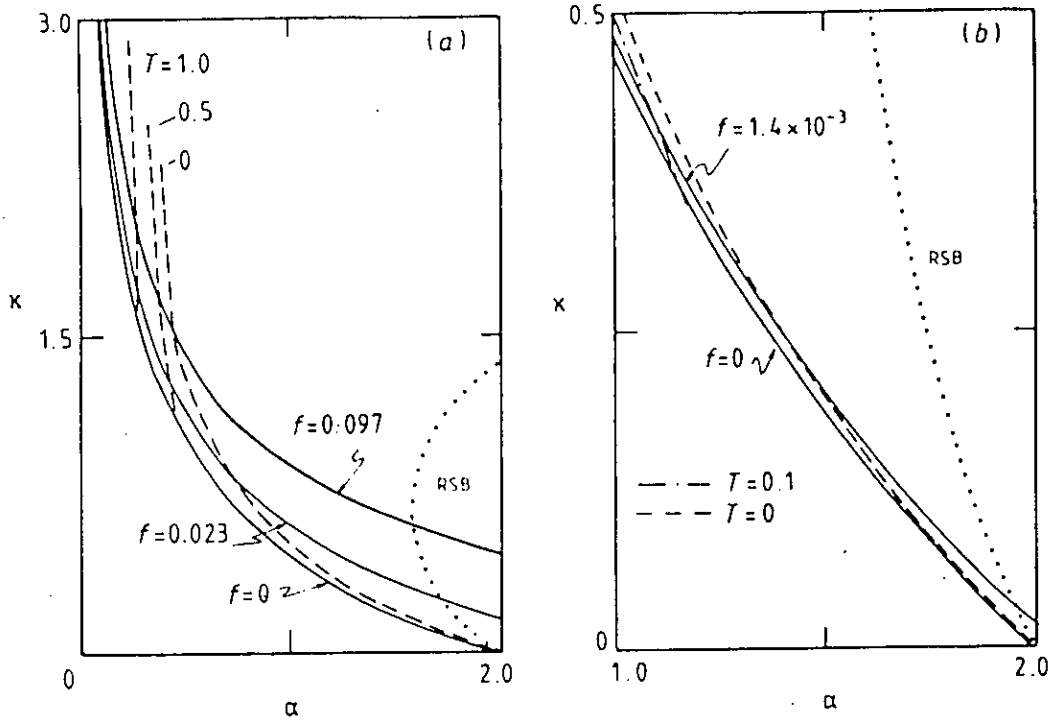

Figure 3.5: Phases of retrieval attractors for optimal storage with errors. Solid curves: perceptron storage capacities α_P vs training stability constraint κ for several values of the error fraction f_{min}. Dashed curves: retrieval storage capacities α_c vs κ for several values of the synaptic noise parameter T. The dotted curve delimits to its right the region in which replica symmetry is broken. (b) is an expansion of the large α small κ corner of (a), for very small error fractions.

temperature curve. One then projects down to the α axis to find α_c. To obtain the error fraction corresponding to α_c, one interpolates between neighbouring curves of constant f_{min}.

An alternative way of reading Fig. 3.5 is to select a value of f_{min} and a temperature and to find the minimum value of κ that can support this error fraction and still give an attractor at that temperature. This is done by finding the intersection between the relevant f_{min} and T curves.

3.7.5 The Noiseless Case

For $\kappa = 0$ the Gardner result, that the perceptron can store as many as $2C$ random patterns implies that an ANN with connectivity C can have as many as $2C$ attractors at zero temperature. It also implies that errors cannot possibly increase this number. The reason is that attractors at $T = 0$ are fixed points. Hence, every attractor is also a configuration stored by the perceptron. But the perceptron

cannot store more than $2C$ uncorrelated patterns. This rules out the possibility of the existence of attractors for $\alpha > 2$, even when the error fraction allows $\alpha_P > 2$. However, for storage at a fixed positive value of κ there is no reason, a priori, to exclude the existence of attractors for values of α with $2 > \alpha > \alpha_P(\kappa, 0)$.

At zero temperature and $\kappa = 0$ we see in Fig. 3.5 that $\alpha_c = 2$ at $f_{min} = 0$ which agrees with our a priori argument. In Fig. 3.5(b) the region of low κ is highlighted to show that for small positive κ, α_c is always less than 2 and attractors exist only for extremely low values of f_{min}. For example, in Fig. 3.5(b) we can take $f_{min} = 1.4 \times 10^{-3}$ and $T = 0$ to find that we must have $\kappa = 0.25$ to have retrieval fixed points. This shows that for low κ storage with errors will be detrimental to the possibility of attractors unless the error fraction is kept extremely small. On the other hand for $\kappa > 0, \alpha_c(\kappa, 0)$ is always greater than $\alpha_P(f_{min} = 0)$, so that at a *fixed* positive value of κ storage with errors always increases the storage capacity for retrieval.

3.7.6 The Presence of Noise

We now consider the effect of finite temperature. At $T = 0$ we have argued that the maximum value of α_c is fixed at 2 and this is given by the perceptron algorithm without errors. At finite T, α_c will be reduced due to the disordering nature of noise. For a spin configuration that is a fixed point at zero temperature to be stable against thermal noise, a finite value of κ is required (see e.g. Fig. 3.3b). Robustness against temperature demands a sacrifice of capacity for stability. Indeed, Fig. 3.3b shows that for the Gardner case ($f_{min} = 0$), in a wide region of κ values (namely those that give continuous transitions to no retrieval with temperature), $T_c \simeq \kappa$. Thus high temperature requires a high κ for retrieval. It is in this high κ regime that storage with errors gives a significant increase in $\alpha_c(\kappa, 0)$ over the Gardner case. For fixed α, storage errors allows a higher value of κ with the sacrifice of stability of a small fraction of sites. An interesting question is whether at finite temperature this trade-off will allow attractors to be retained at α values where otherwise (for no storage errors) there would be no attractors. In other words for a fixed finite temperature can the overall maximum of α_c be given by storage with errors?

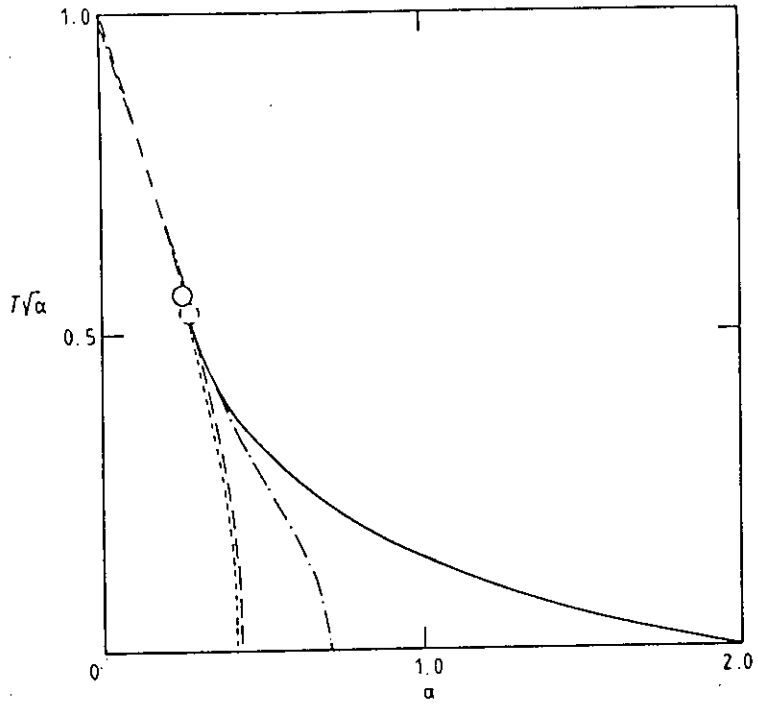

Figure 3.6: Phases of retrieval attractors for optimal storage with errors $f_{min} = 0.023$ and without errors ($f_{min} = 0$). Conventions for $f_{min} = 0$ are as for Fig. 3.3a. For $f_{min} = 0.023$ the long chain curve is the line of discontinuous transitions to no retrieval; the short chain curve is the line of continuous transitions to no retrieval; the tricritical point is marked by a dotted circle.

Returning to Fig. 3.5 one observes, inspecting the lines of constant critical temperature, that as T increases they become more vertical. This implies that as T increases, α_c becomes essentially independent of κ, over a wide range of κ. Equivalently, storage with no errors has approximately the same α_c as storage at a higher value of κ and allowing errors. For $T = 1.0$ the fixed temperature curve even bends back slightly on itself. The maximum α_c (the overall maximum storage capacity) for $T = 1.0$ is found to be 0.279 at $f_{min} = 0.037$, whereas for no errors $\alpha_c = 0.275$. Thus the improvement in α_c at finite temperature is only marginal. Finally, it is important to notice in Fig. 3.5 that in the region where replica symmetry breaking occurs, see (3.15), there are no attractors present. The point α=2, κ=0 is on the boundary. In the entire region where attractors exist replica symmetry is stable.

In order to compare retrieval phase diagrams of optimal perceptron storage networks with and without errors, we present in Fig. 3.6 a phase diagram for a network storing patterns with a fixed fraction of errors $f_{min} = 0.023$, superposed on the phase diagram Fig. 3.3a for which f_{min}=0. The two diagrams are qualitatively similar. The most significant difference between the two phase diagrams is

that at $T \lesssim 0.2$ the regime of narrow retrieval is sharply curtailed in the case of finite f_{min}, which reflects the sensitivity of the attractors to errors at low values of κ. It is also of interest to note that at high κ values (higher than that of the tricritical point) the regime of 'wide retrieval' is increased if errors are allowed. This is indicated by the continuous transition line for the network with errors rising marginally above the continuous transition line for the network without errors as the temperature is raised. Although this does reflect the enhanced robustness to noise of storage with errors over storage without errors at a fixed high value of κ, the effect is only marginal.

3.7.7 Summary of Results

The first aim of this chapter was to study the performance of the Gardner optimal interactions in a noisy attractor neural network. Figures 3.3a) and b) present the main results of this study. As the temperature is raised the storage capacity α_c for Gardner interactions decreases quite sharply. At low κ the transition to no retrieval is discontinuous whereas at high κ it is continuous. The position of the tricritical point where the discontinuous and continuous transition lines meet on the phase diagram can be calculated. The critical temperature is approximately equal to κ over a wide range of κ which validates the interpretation of κ as parameterising the stability to thermal noise. High values of κ also improve the retrieval quality. At high temperatures the randomly diluted Hopfield model has a larger storage capacity which underlines the fact that the Gardner optimal interactions are strictly optimal only at zero temperature. At low temperatures nevertheless α_c for the Gardner interactions is still relatively high.

Errors in storage, with interactions of the Gardner–Derrida optimal type, reduce α_c drastically at zero and low temperature. This reflects the sensitivity of the attractors to storage errors when the retrieval is narrow. At intermediate temperature, where the retrieval is wide, storage with a small error fraction may increase α_c over storage with no errors and a lower κ. However this effect is only marginal. One may conclude that storage errors of the Gardner–Derrida type are not particularly productive when one considers retrieval attractors. This has been demonstrated by Wong and Sherrington[43,45], who have recently studied optimal interactions that result from cost functions that are more relevant to retrieval in

a noisy environment.

Chapter 4

Towards Biology: Sparse Spatial Coding and Biased Patterns

4.1 Biased Patterns and the 1,0 representation

In the previous chapters the patterns to be stored have been been assumed random. This has meant that, to within $\sqrt{N}$ fluctuations, in a nominated pattern half the spins take value 1 and half the spins take value -1. This choice of patterns maximises the information stored in each pattern and facilitates the technical calculation of averages.

From a biological viewpoint random patterns are not acceptable. If the patterns are random, then a particular spin takes value 1 in half the patterns. This corresponds to a neuron in the neural assembly being involved in half of the computations that the assembly can perform. Biological evidence [46] suggests that neurons in the cortex have the spontaneous firing rate most of the time and have the elevated firing rate rather infrequently. This is interpreted to mean that a particular neuron is involved in only a small fraction of the computations that are carried out. Furthermore the fraction of neurons with an elevated firing rate at any one time is small [47].

These observations lead one to consider the case of biased patterns or as it has become more recently known — sparse spatial coding. In biased patterns the fraction f of neurons taking the active state in a given pattern is less than a half.

To set this down mathematically I shall change representation from the S_i an Ising spin (1,-1) neuron, to V_i a neuron that takes values (1,0). Although it has been shown that for unbiased patterns the V representation gives a storage capacity half that for the S representation [48], the V representation will prove to be more suitable for the models considered in this chapter. In fact initial investigations into the storage of biased patterns were discouraging because Ising spins were used [49]. In V representation η_i^μ denotes the value of the ith neuron in the μth nominated pattern. The exact definition of f is then given by the probability distribution of the quenched random variables η.

$$P(\eta) = f\,\delta(\eta - 1) + (1 - f)\delta(\eta) \tag{4.1}$$

Sparse coding is quite suitable for more general image recognition problems. A pattern or image, in any sense, involves a foreground and a background. The foreground, although it takes up only a small fraction of the field of view, contains the *features* of the pattern. The background is of less importance. A misrepresentation of the background of the pattern should not necessarily stop the pattern as a whole being recognised, whereas a misrepresentation of the foreground should be more serious. One can interpret the spins that take value 1 in a pattern as the foreground of the image and those that take value 0 as the background.

Gardner[37] extended her study of the interaction space volume to consider the case of sparse coding. She found that as the bias increases and f tends to zero the storage capacity α_c diverges as

$$\alpha_c \sim \frac{1}{f \mid \log f \mid}. \tag{4.2}$$

Although this result indicates that when sparse coding is used many more patterns can be stored, the total information stored within all the patterns does not increase.

In this chapter I shall study two connection rules that exhibit the divergence (4.2) in α_c. These are the Covariance Rule[50] and Willshaw's Rule[10]. The aim of this chapter is to give an overall understanding of the attractor structure in attractor neural networks that store biased patterns, and in particular to illustrate how the threshold controls the activity in the dynamic evolution of such networks. As was demonstrated in the previous chapter, a useful way to understand network

dynamics and consequently the attractor structure is to consider the randomly dilute lattice. This shall be carried out for the Covariance Rule in section 4.3. A remarkable result that shall be derived, is that for $f << 1$ the fixed point equations for the randomly diluted network are equivalent to the saddle point equations for the fully connected model. This indicates that randomly diluting the lattice may not change the characteristics of this model. In light of this, a detailed study of the attractor structure and trajectories in the space of the overlap parameters near to the memory attractor, will be pursued in sections 4.4–4.5. The order parameter maps that shall be derived always exhibit two attractors that are not correlated with the stored patterns. These are the "all zeros", or quiescent, fixed point and an high activity attractor. The latter is rather undesirable as it corresponds to a high spatial activity in the neural population.

The Willshaw connection scheme shall be considered in section 4.7. First Willshaw's original analysis is reviewed and its validity, physical interpretation and generalisations considered in 4.7. These analyses will then be used to show that the key features of biased pattern networks highlighted in the study of the Covariance rule, persist with the Willshaw rule. However Willshaw's connection rule was introduced in the context of pattern associators (section 1.4), thus section 4.8 aims to give an understanding of how the rule functions in the context of attractor neural networks. The most disappointing aspect of both connections schemes is the existence of a high activity attractor. A method of suppressing this attractor will be studied in 4.8.

4.2 The Covariance Rule

The Covariance Rule was first introduced by Sejnowski[50], and later reconsidered within attractor neural networks [51,52]. The connection strengths are given by

$$J_{ij} = \frac{1}{N}\sum_{\mu=1}^{p}(\eta_i^\mu - f)(\eta_j^\mu - f). \tag{4.3}$$

A variation of the mean-field theory of the Hopfield model discussed in chapter 1 [51,52] shows that at zero temperature and $f \ll 1$, the *optimal* value of threshold gives a first order phase transition to no memory at a value of α of the form (4.2).

When biased patterns are used the threshold θ becomes an important parameter. The expression (4.2) only holds for the covariance rule if the threshold is suitably chosen. The rôle of θ becomes more apparent when one considers the local field at a site i

$$h_i = \sum_{j \neq i} J_{ij} V_j - \theta \tag{4.4}$$

and the updating rule of the dynamics

$$\begin{aligned} V_i(t+1) = 1 \quad &\text{with probability} \quad \left[1 + \exp\left(-\frac{2h_i(t)}{T}\right)\right]^{-1} \\ V_i(t+1) = 0 \quad &\text{with probability} \quad \left[1 + \exp\left(\frac{2h_i(t)}{T}\right)\right]^{-1} \end{aligned} \tag{4.5}$$

where T is the temperature. The threshold acts to restrict the number of ones in a configuration. In the extreme case that θ is very large then all h_i will become negative and the configurational flow will be towards the all zeros configuration where all spins take value 0. This example of the configurational flow moving away from the region of configuration space with the same bias as the patterns shows that in order to describe the network fully the usual overlap parameter m needs to be complemented. To express the overlap of the configuration of the system with a single pattern (taken to be pattern 1) one requires two order parameters.

$$m_1^{(1)} = \frac{1}{Nf} \sum_{i=1}^{N} < \eta_i^{(1)} V_i > \tag{4.6}$$

$$m_2^{(1)} = \frac{1}{Nf} \sum_{i=1}^{N} < (1 - \eta_i^{(1)}) V_i > . \tag{4.7}$$

It is also useful to consider

$$y = m_1^{(1)} + m_2^{(1)}. \tag{4.8}$$

These parameters have straightforward physical interpretations: m_1 measures the number of "correct ones" (sites that are 1 in the configuration and pattern); m_2 measures the number of "incorrect ones" (sites that are 1 in the configuration and 0 in the pattern); y measures the "activity" (the number of ones in the configuration). The choice of m_1 and m_2 is convenient because they correspond to averages over 1-sites (sites where $\eta_i^{(1)} = 1$) and 0-sites (sites where $\eta_i^{(1)} = 0$) of the pattern respectively. We shall see that the asymmetry between the fields at the 2 types of site necessitates this separation.

Mean-field theory of the type discussed in chapter 1 gives no information on the dynamical behaviour of the model, which is the aspect we examine in this chapter. Specifically we consider flows in phase space of a the randomly diluted version (see 3.4) of the model in which the dynamics can be solved exactly. The points to be investigated are:

- To what extent are the flows restricted to the region of phase space that has the same bias as the patterns?
- What are the qualitative features of the basins of attraction of the patterns and how do they change as the system becomes saturated at α_c?
- How does varying the threshold change the flows?

In previous chapter (section 3.7.3) the results of randomly diluting the Hopfield model were discussed. We saw that the Hopfield model changes its characteristics considerably. In particular the transition at α_c becomes second rather than first order and α_c has a higher value. However we will show that when we randomly dilute the connection rule (4.3) the transition remains first order. Moreover in the limit $f \ll 1$ the fixed point equations of the randomly diluted model are identical to the saddle point equations of the fully connected model. The implications of this result will be discussed in section 4.6, first a derivation of the dynamic equations will be presented.

4.3 Derivation of Evolution Equations

The connection rules is randomly diluted in the usual manner (see section 3.4) and becomes

$$J_{ij} = \frac{C_{ij}}{C} \sum_{\mu=1}^{l} (\eta_i^\mu - f)(\eta_j^\mu - f), \tag{4.9}$$

where the distribution of C_{ij} is given by

$$P(C_{ij}) = \frac{C}{N} \delta(C_{ij} - 1) + \left(1 - \frac{C}{N}\right) \delta(C_{ij}). \tag{4.10}$$

C is the mean number of other neurons a particular neuron is connected to.

We now proceed to derive equations for the evolution of the order parameters. Consider the situation where $m_1^{(1)}$ is finite but $m_1^{(\mu)} \sim O(1/Nf)$, for $\mu > 1$. This requires $y \sim O(1)$. In this situation the configuration is near to pattern 1 and we need only consider the 2 order parameters associated with pattern 1. With this in mind we will drop the pattern 1 superscript from the order parameters.which may be expressed as

$$\begin{aligned} m_1(t+1) &= \ll \langle V_i(t+1) \rangle \gg_{1-\text{sites}} && (4.11) \\ &= \ll \left[1 + \exp\left(-\frac{2h_i(t)}{T}\right)\right]^{-1} \gg_{1-\text{sites}}, && (4.12) \\ m_2(t+1) &= \ll \langle V_i(t+1) \rangle \gg_{0-\text{sites}} && (4.13) \\ &= \ll \left[1 + \exp\left(-\frac{2h_i(t)}{T}\right)\right]^{-1} \gg_{0-\text{sites}} . && (4.14) \end{aligned}$$

The single angular bracket indicate a thermal average and the double angular brackets a composite average over sites and an ensemble of initial conditions. To perform the site averages we must construct expressions for the field distributions. As the configurations are near to pattern 1 we can split the field into a sum of signal terms from pattern 1 and a sum of noise terms from the other patterns.

$$h_i(t) = (\eta_i^1 - f) \sum_{k=1}^{K} \left(\eta_k^1 - f\right) V_k(t) + \sum_{\mu=2}^{l} \sum_{k=1}^{K} (\eta_i^\mu - f)(\eta_k^\mu - f) V_k(t) - \theta \qquad (4.15)$$

where the k index labels the K sites that are connected to site i ($C_{ik} = 1$). For a site k that is a 1-site

$$V_k(t) = 1 \text{ with probability } m_1(t), \qquad (4.16)$$

and for a site k that is a 0-site

$$V_k(t) = 1 \text{ with probability } \frac{f}{(1-f)} m_2(t), \qquad (4.17)$$

We can now write down the probability that

$$h_i(t) = (\eta_i^1 - f)(S_1(1-f) - S_2 f) + N_1(1-f)^2 - N_2 f(1-f) + N_3 f^2 - \theta \qquad (4.18)$$

as

$$\begin{aligned} & P(S_1, S_2, N_1, N_2, N_3) && (4.19) \\ &= \sum_{K=0}^{N} \frac{C^K e^{-C}}{K!} \frac{K!}{S_1! S_2! S_3!} (f m_1(t))^{S_1} (f m_2(t))^{S_2} (1 - f y(t))^{S_3} \end{aligned}$$

$$\times\delta_{S_1+S_2+S_3,K}$$
$$\times\frac{(K(P-1))!}{N_1!N_2!N_3!N_4!}(f^3y(t))^{N_1}(2(1-f)f^2y(t))^{N_2}$$
$$\times((1-f)^2fy(t))^{N_3}(1-fy(t))^{N_4}$$
$$\times\delta_{N_1+N_2+N_3+N_4,K(P-1)}. \tag{4.20}$$

Although this equation appears complicated it is rather easy to understand. The S terms are the signal, coming from the first term on the r.h.s of (4.15), and the N terms are noise. As the lattice is randomly diluted the spins are uncorrelated, therefore the probability distributions for the values which each of the terms in (4.15) take are independent. For example one of the K signal terms will be an S_1 term, which contributes $(\eta_i^1 - f)(1-f)$, if $\eta_k^1 = 1$ and $V_k = 1$. The probability that $\eta_k^1 = 1$ is f. Given that $\eta_k^1 = 1$ the probability that $V_k = 1$ is $m_1(t)$. Therefore the probability of an S_1 term is fm_1. Likewise one of the $K(p-1)$ noise terms will be a N_1 term, which contributes $(1-f)^2$, if $\eta_i^\mu = 1, \eta_k^\mu = 1$ and $V_k(t) = 1$. The probability of this happening is qf^3. The multinomial factors and Kronecker deltas come from the constraints in the number of terms in each sum in (4.15). Finally the sum over K, weighted with Poisson probability, averages over how many sites i is connected to.

As $C \to \infty$, bearing in mind that $C << \ln N$ the probability distribution h_i, $\rho(h_i)$ becomes a Gaussian:

$$\rho(h_i) = \frac{1}{\sqrt{2\pi\alpha y f^3(1-f)^2}} \exp -\frac{(h_i - (\eta_i^1 - f)\left(f(1-f)m_1 - f^2m_2\right))^2}{2\alpha y f^3(1-f)^2}. \tag{4.21}$$

The mean of this Gaussian is given by the mean of the summation of signal terms in (4.15), because the mean of the noise terms vanishes. The variance of the distribution comes from the variance of the summation of noise terms in (4.15), because the contribution from the signal terms to the variance is of order $1/C$.

In this chapter we are interested in the regime of highly biased patterns $f \ll 1$. In this regime one need only keep the lowest order terms in f and (4.21) becomes

$$\rho(h_i) = \frac{1}{\sqrt{2\pi\alpha y f^3}} \exp -\frac{(h_i - \eta_i^1 f m_1)^2}{2\alpha y f^3}. \tag{4.22}$$

Referring back to (4.20), equation (4.22) tell us that for $f \ll 1$ the dominant contribution to the signal is from S_1 terms and the dominant contribution to the noise is from N_1 terms. Equation (4.22) also shows that the mean of h_i depends

on whether site i is a 1 or 0-site. The two different Gaussian distributions that result from $\eta_i^1 = 1, 0$ can be used to perform the averaging over 1-sites in (4.12) and over 0-sites in (4.14) respectively. One finds

$$\begin{aligned} m_1(t+1) &= \int_{-\infty}^{\infty} Dz \left(1 + \exp\left[\frac{-2}{T_0}\left(m_1(t) + (\alpha f y(t))^{1/2} z - \theta_0\right)\right]\right)^{-1} \\ m_2(t+1) &= \frac{1}{f}\int_{-\infty}^{\infty} Dz \left(1 + \exp\left[\frac{-2}{T_0}\left((\alpha f y(t))^{1/2} z - \theta_0\right)\right]\right)^{-1}, \end{aligned} \tag{4.23}$$

$$\text{where} \quad T_0 = \frac{T}{f}; \quad \alpha = \frac{P}{C}; \quad \theta_0 = \frac{\theta}{f}. \tag{4.24}$$

At $T = 0$ the integrals appearing in (4.23) can be written as Gauss error functions to give

$$\begin{aligned} m_1(t+1) = &\quad \frac{1}{2}\left[1 + \text{ erf}\left(\frac{m_1(t) - \theta_0}{(2\alpha f y(t))^{1/2}}\right)\right] \\ m_2(t+1) = &\quad \frac{1}{2f}\left[1 - \text{ erf}\left(\frac{\theta_0}{(2\alpha f y(t))^{1/2}}\right)\right]. \end{aligned} \tag{4.25}$$

One could also consider the model defined by random sequential dynamics where at each time step a site is chosen randomly and updated by rule (4.5). This model would have, in place of the map (4.25), the flow

$$\begin{aligned} \frac{dm_1}{dt} &= \frac{1}{2}\left[1 + \text{ erf}\left(\frac{m_1 - \theta_0}{(2\alpha f y(t))^{1/2}}\right)\right] - m_1 \\ \frac{dm_2}{dt} &= \frac{1}{2f}\left[1 - \text{erf}\left(\frac{\theta_0}{(2\alpha f y(t))^{1/2}}\right)\right] - m_2, \end{aligned} \tag{4.26}$$

which is similar in form to the Wilson and Cowan[53] evolution equations for populations of interacting excitatory and inhibitory neurons. The fixed points of (4.25) and (4.26) are identical and one would expect that the trajectories in the m_1–m_2 plane near to these fixed points are qualitatively the same. For simplicity only the model using parallel dynamics will be considered further.

4.3.1 Discussion of Evolution Equations and their Fixed Point Structure

In studying the map (4.25) we are primarily interested in fixed points that are highly correlated with the patterns. These fixed points will be referred to as memory attractors to distinguish them from the nominated patterns. The fixed point

equations of (4.25) are equivalent to the saddle point equations derived for the fully connected model [51,52]. This is a remarkable result because it indicates that the replica method is unnecessary when $f \ll 1$. In order to understand why this is so we can compare the overlap saddle point equation for the fully connected Hopfield model, (2.78–2.79), and the order parameter map for the randomly diluted Hopfield model, (2.29–2.30). The difference between these equations lies in the order parameter r. Equation (2.60) gives the physical meaning of r: it is the sum of the squares of the overlaps from uncondensed patterns. However these microscopic overlaps are not independent and the replica method yields a self-consistency equation for r in terms of the Edwards–Anderson order parameter q. In the present case it appears that treating these overlaps as independent (which is what random dilution effectively allows one to do) gives correct results in the limit $f \ll 1$, even for the fully connected model.

This can be explained by recalling that it was N_1 terms in equation (4.20) that contributed to the noise. This meant that the noise at a site i only came from patterns μ where i was a 1-site. Furthermore the noise due to such a pattern μ came from sites k that had $V_k = 1$ and were 1-sites in pattern μ. For low activity y and for $f \ll 1$ the sites k that fulfill these criteria will tend to be different for different patterns μ. This suggests that the uncondensed overlaps are to a large extent independent and the Gaussian treatment of the noise due to uncondensed patterns is justified even for the fully connected model.

4.3.2 Analysis of Fixed Point Equations

The analysis of the map (4.25) is not as straightforward as the analysis of section(3.6) because the map is two dimensional. As shall be shown in the next section, this means that transitions where fixed points lose their stability or disappear can be rather complicated. In the next section we shall examine these transitions numerically. However we can make some qualitative analysis to determine the qualitative behaviour of some critical quantities.

The first feature to note of (4.25) is that there are two fixed points of the map that are not correlated with the patterns. These are the "all zeros" fixed point $m_1 = m_2 = 0$ and the high activity attractor where $y \sim O(1/f)$. The second of

these is outwith the region of validity of (4.25) because for such a large activity the configuration will have finite m_1 overlaps with more than one pattern. Although this attractor's position cannot be determined by (4.25) it does exist in the model and its effect on trajectories near to a pattern is evident in the next section. One would expect that at high thresholds the memory attractor will become unstable to the all zeros fixed point whereas at low thresholds the instability will be towards the high activity attractor.

For a memory attractor we desire a fixed point of (4.25) of the form

$$\begin{aligned} m_1^* &= 1-\epsilon \\ m_2^* &= \phi \end{aligned} \tag{4.27}$$

where ϵ and ϕ are small quantities. The starred values of the overlaps indicate that they are fixed point solution of (4.25). At low values of the threshold one would expect that the activity y of the fixed point would be larger than for high values of the threshold. Therefore the overriding condition for a memory attractor is that ϕ be small. We can make a first order approximation to ϕ by inserting $m_1^* = 1$, $m_2^* = 0$ into the right hand side of the fixed point equation for m_2^* and equating the left hand side to ϕ. One can then expand the error function for a large argument by using (3.48), to give

$$\phi = \frac{1}{\theta}\sqrt{\frac{\alpha}{2f\pi}}\exp -\frac{\theta_0^2}{2\alpha f}. \tag{4.28}$$

To determine whether this value of ϕ is less than unity the natural log of this equation is taken and the large terms kept to find the condition

$$\frac{1}{2}\ln\left[\frac{\alpha}{f}\right] - \frac{\theta_0^2}{2\alpha f} < 0. \tag{4.29}$$

If $\alpha \ll 1/f$ one can ignore the α in the logarithm and find a critical value of α at

$$\alpha_c(f) \approx \frac{\theta_0^2}{f \mid \log f \mid}. \tag{4.30}$$

This critical value of α is consistent with $\alpha \ll 1/f$ and it will give the qualitative form of α_c in the regime of 'small' θ_0.

For a high threshold , $\theta_0 \sim 1$, one would expect the activity of a memory attractor to be low. The overriding condition in (4.27) is that ϵ be small. Again we can a

first order approximation to find

$$\epsilon = \frac{1}{(1-\theta_0)}\sqrt{\frac{\alpha f}{2\pi}}\exp -\frac{(1-\theta_0)^2}{2\alpha f}. \tag{4.31}$$

When the natural logarithm of this equation is taken and only the dominant terms retained one finds

$$\alpha_c(f) \approx \frac{(1-\theta_0)^2}{2f \mid \ln(1-\theta_0) \mid}. \tag{4.32}$$

This is the qualitative form of the critical value of α for 'high' threshold.

At some intermediate value of θ_0 there must be a crossover between the two forms of α_c. As θ_0 increases this will occur when (4.30) becomes equal to(4.32). This condition gives

$$\frac{2 \mid \ln(1-\theta_0) \mid \theta_0^2}{(1-\theta_0)^2} = \ln f. \tag{4.33}$$

For $\theta_0 \approx 1$ the dominant term on the l.h.s of (4.33) is $1/(1-\theta_0)^2$ so that the qualitative form of the crossover value of θ_0 is

$$\theta_0 = 1 - (\mid \ln f \mid)^{-\frac{1}{2}}, \tag{4.34}$$

which is very close to unity for $f \ll 1$. Overall the maximum value of α_c will be given when θ_0 is just below the crossover value and so

$$\alpha_c \approx \frac{1}{f \mid \ln f \mid} \tag{4.35}$$

which is the same as the Gardner bound (4.2). Of course the Gardner bound is for memory attractors to be identical to the nominated patterns so that $m_1^* = 1$ and $m_2^* = 0$. Therefore the interactions given by (4.3) are not the Gardner optimal interactions.

The existence of the 2 fixed points uncorrelated with the patterns reflects the fact that the activity is not rigidly constrained by the dynamics. Although the stored patterns have a set low level of activity the configurational flow may explore regions of configuration space that do not have the same level of activity. The main parameter that controls the level of activity is θ and we have used the dependence of the fixed points on θ in the preceding analysis. However one can also get a qualitative understanding of the directions of configurational flows by looking at how the level of activity y comes into the map (4.25).

For both 1 and 0-sites the variance of the noise distribution is $2\alpha y$. An increase in the activity will increase the noise which in general will decrease m_1 but increase m_2. This can be seen by noting that in (4.25) the derivative of m_1 with respect to y is negative whereas the derivative of m_2 is positive. Therefore the activity may continue to increase or decrease depending on the threshold, storage and initial level of activity. This constitutes a feedback-loop in the activity. that may be delicately balanced.

In order to analyse the flows in the neighbourhood of the memory attractor one could perform a linear stability analysis of the fixed points of (4.25). However such an analysis is difficult to carry out in more detail than that of the initial part of this section because one cannot obtain simple exact expressions for the position of the memory attractor. In the following section we shall resort to a numerical investigation of the configurational flows.

4.4 Numerical Study of Evolution Equations

In this section we quantify numerically the qualitative analysis and discussion of the previous section. In order to reduce the number of parameters we will keep a fixed value of $f = 10^{-4}$ and only consider the zero temperature equations. This leaves two parameters α and θ_0.

The analysis performed in the previous section to determine the form of the critical value of α rested on the assumption that θ_0 controlled the level of activity y in the memory fixed point. Figure 4.1 validates this assumption. The figure plots the position of the memory attractor at fixed α as θ_0 increases. At the lowest θ_0 value shown the activity and both m_1^* and m_2^* are greatest. One sees that as θ_0 increase both m_1^* and m_2^* decrease. This means that the number of incorrect zeros increases but the number of incorrect ones decreases. The transition to no memory occurs at $\theta = 0.73$ when m_1^* drops discontinuously to zero whereas m_2^* vanishes continuously. In this case we see a transition to no memory at fixed α as θ_0 increases.

We now turn to investigating the transition to no memory at fixed θ_0 as α increases. In chapter 3 one could also obtain information about the nature of the

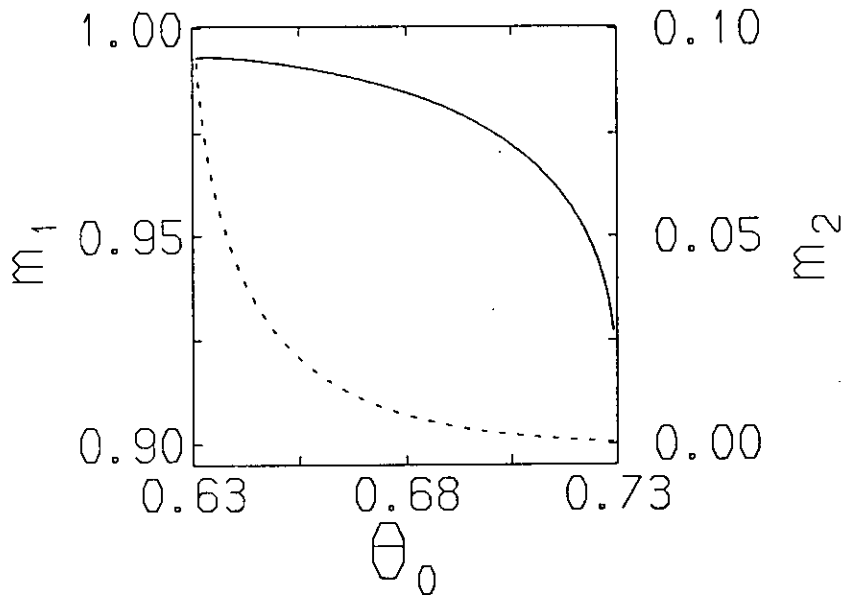

Figure 4.1: A plot of the m_1 (full curve) and m_2 (broken curve) values of the memory fixed point against θ_0 for $f = 10^{-4}$ and $\alpha = 200$.

basin of attraction at the transition, namely whether there was narrow or wide retrieval. However in the present case the nature of the basins of attraction is rather more complicated. The distance a configuration is from the pattern could be be parameterised by the Hamming distance. However this obscures the nature of the incorrect spins within the configuration. The use of the m_1 and m_2, which parameterise the number of incorrect zeros and incorrect ones respectively, allows the basin of attraction to be viewed in a plane. The value of θ_0 should influence the shape of the basin of attraction in this plane. As α increases one might expect the basin of attraction to vanish in a manner dependent on this shape. The vanishing of the basin of attraction is intrinsically linked to the nature of the transition to no memory, therefore the value of θ_0 should effect the nature of the transition to no memory.

Figures 4.2 and 4.3 show trajectories in the m_1–m_2 plane at selected values of θ_0 and α. They show sequences in which α increases past some critical value. In fig.4.2 at $\alpha = 260$ the memory is an attractive node and there is a saddle point to its lower left which limits the memory's basin of attraction. The term attractive node means qualitatively a fixed point of the map to which there are definite directions of entry. A saddle point is an unstable fixed point which has definite directions of departure. There is also a saddle point, at higher value of m_2 out of the frame, which acts as a watershed for trajectories reaching the high activity

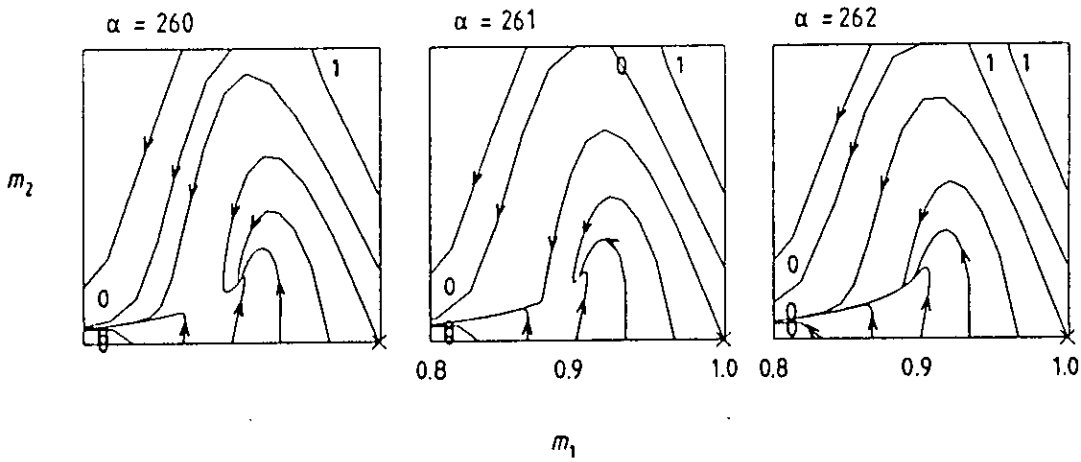

Figure 4.2: A Sequence of frames showing trajectories of (4.25) near to the pattern which is marked by a cross. In the sequence α passes through its critical values. When a trajectory leaves the frame, a 1 indicates that it continues to the high activity attractor and a 0 indicates that it continues to the all zeros fixed point. The fixed parameter values are $f = 10^{-4}$ and $\theta_0 = 0.699$.

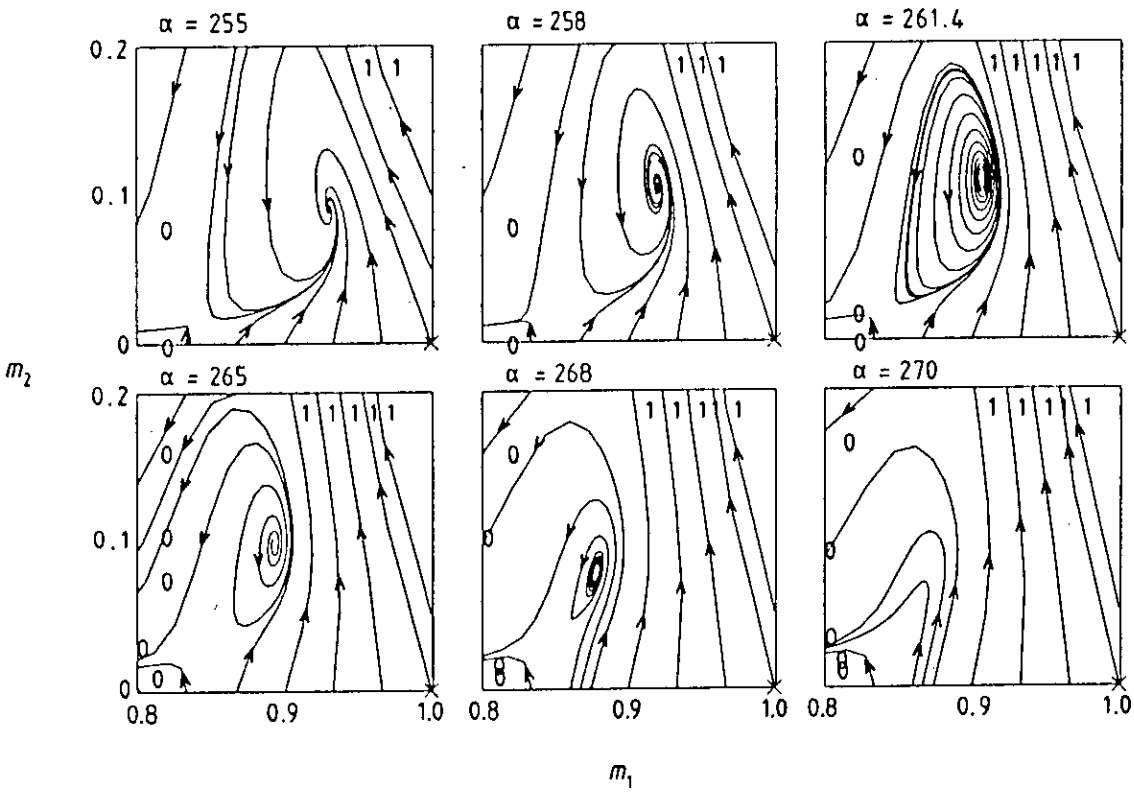

Figure 4.3: A sequence of trajectories as in fig.4.2. The fixed parameter values are $f = 10^{-4}$ and $\theta_0 = 0.6917$.

attractor. At $\alpha = 261$ the qualitative structure remains the same. However a trajectory starting at the pattern (marked by a cross) is no longer within the memory's basin of attraction. When α reaches 262 the memory and lower saddle point have annihilated. This sequence of memory loss can be classified as a saddle-node bifurcation. At the transition value of α the trajectories that were contained in the basin of attraction of the memory attractor are captured by the all zeros fixed point. We shall describe this as the attractor destabilising towards the all zeros fixed point.

In fig. 4.3 the memory starts as a spiral. Qualitatively an attractive (unstable) spiral is a fixed point with no definite directions of entry (departure). The 2 saddle points mentioned above are again present. At $\alpha = 258$ the pattern is no longer within the memory's basin of attraction. The memory destabilises at $\alpha = 261.4$ and trajectories starting from it spiral out to an attractive limit cycle. At $\alpha = 265$ this limit-cycle is no longer present and all trajectories leave the frame. However at $\alpha = 268$ the spiral restabilises and so a memory is again present. When α reaches 270 the lower saddle and memory have annihilated and there are no fixed points present in the frame. The destabilisation and restabilisation of the memory, of which this sequence is an example, will be referred to as *intermittancy*

These two figures illustrate the 2 types of memory loss that are present in the model — saddle node bifurcations and spiral destabilisations. At lower θ_0 values the same mechanisms are present but spirals destabilise to the high activity attractor and the higher saddle point is involved in the bifurcations. One may summarise these observations by the following generalisations of the features of the memory near to the transition.

- At high θ_0 the attractor is a node and which destabilises with increasing α with respect to the all zeros fixed point.
- As θ_0 is decreased slightly. The feed-back loop in activity, mentioned in the previous section, becomes delicately balanced and the trajectories begin to spiral. The memory is then an attractive spiral. The transition to no memory will occur through complicated mechanisms such as limit cycles and intermittancy. However the end result is destabilisation towards the all zeros fixed point.

- As θ_0 is decreased still more the memory attractor is again a spiral. However the overall destabilisation is towards the high activity attractor.
- For low θ_0 the memory is a node that destabilises towards the high activity attractor.

One can interpret these various memory losses as the basins of attraction of the all zeros and high activity attractors becoming large enough to encroach on the memory. The value of θ_0 then determines which of these basins of attraction reaches the memory first. Figure 4.3 is in the intermediate θ_0 regime where both non-memory attractors are strongly affecting the memory. This gives rise to the spiraling, limit cycles and intermittancy.

Figures 4.2 and 4.3 again show that the transition to no memory is discontinuous because m_1^* and m_2^* do not vanish continuously at the transition. In addition we see that the basin of attraction of the memory attractor does not take up the whole of the m_1–m_2 plane. In the language of chapter 3 the basins of attraction are always narrow. However in chapter 3 we saw that when the retrieval was narrow then the basin of attraction vanished continuously at the transition to no memory which was discontinuous in m. In the present case the basins of attraction do not vanish continuously.

The figures also illustrate the need for a more careful definition of the critical value of α. When considering neural networks as associative memories one would like the the pattern to be within the basin of attraction of the memory attractor so that there is a clear association between them. The value of α at which this is no longer so will be called α_1. However α_1 is not a parameter directly relevant to stability analysis of the fixed points of the map, thus unless one can find formulae for the basins of attractions, it is only from numerical studies of flows as performed in this work that α_1 can be determined. A second α value of importance is α_2, which we define to be the lowest value at which there is no memory. By using α_1 and α_2 we avoid dealing with any intermittancy that may be present. α_2 corresponds more closely to the definition of α_c used in previous chapters. However α_1 is an easier parameter than α_2 to determine numerically. One need only insert $m_1 = 1$ and $m_2 = 0$ as the initial condition for the map (4.25) and then iterate to a fixed point. If this fixed point is the all zeros fixed point or high activity attractor then α is greater than α_1.

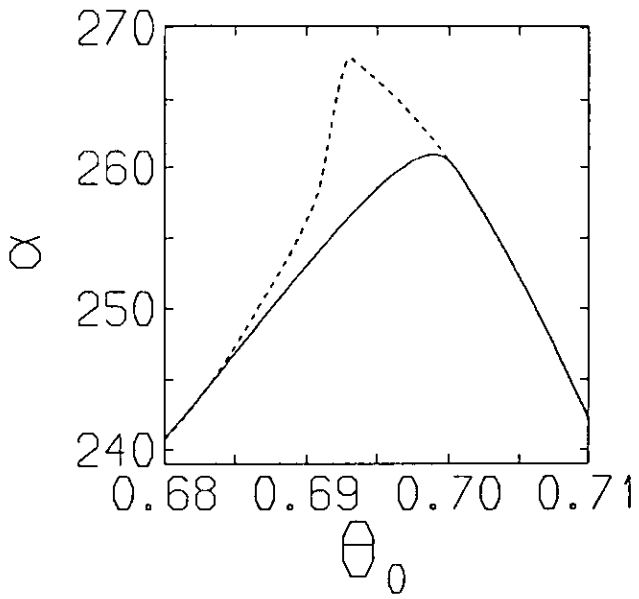

Figure 4.4: A plot of α_1 (full curve) and α_2 (broken curve) against θ_0 for $f = 10^{-4}$. The range of θ_0 is that within which both α values are maximised.

Fig.4.4 shows that α_1 and α_2 differ only in the region of θ_0 values that give the highest storage. This also illustrates that in the region of highest storage capacity the basins of attraction are small. The slight kink on the left of α_2 curve is where the attractor becomes a spiral. The ease of calculation of α_1 and the similarity of the $\alpha_1(\theta_0)$ and $\alpha_2(\theta_0)$ curves make it convenient to use α_1 to quantify the storage capacity throughout the rest of this chapter.

Fig.4.5 shows how α_1 has a θ_0 dependence similar to (4.30) and (4.32). Namely for low θ_0 the curve appears quadratic in θ_0 whereas for the high theta values it appears approximately quadratic in $1 - \theta_0$. The position of the memory at α_c, shown by the broken curves, moves sharply in the transition region between the 2 forms.

4.5 Higher than Random Overlap between Patterns

In the Hopfield model it has been found that multiple storage of a pattern improves its recall [54]. This as an example of a certain memory being reinforced or stored more strongly than others [55]. Storing a pattern twice is the extreme case of

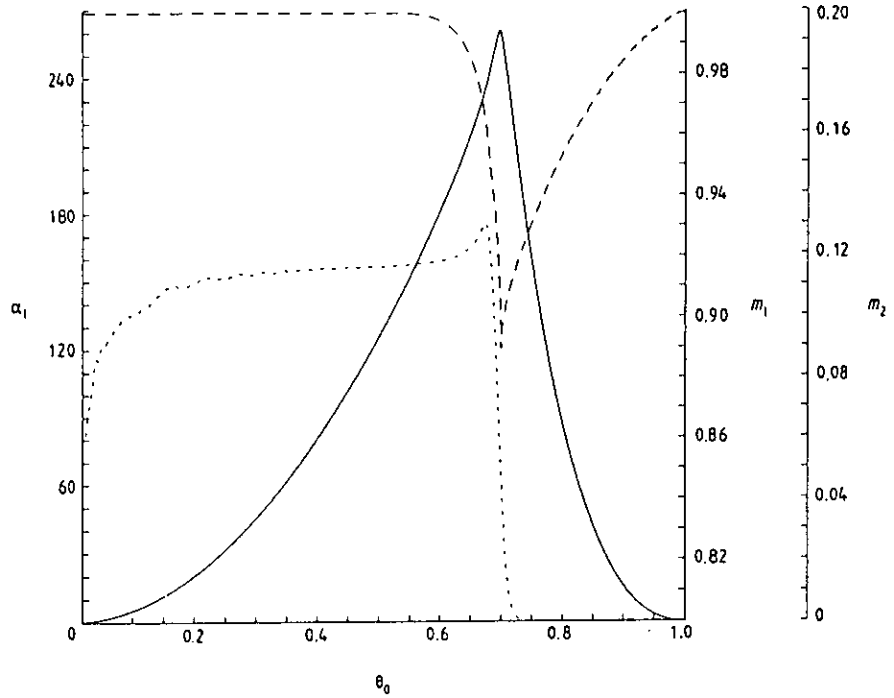

Figure 4.5: A plot of α_1 against θ_0 at $f = 10^{-4}$ with the values of the order parameter memory fixed points at α_1 superimposed. Full curve, α_1; broken curve m_1; dotted curve m_2.

storing two patterns that have higher than random overlaps with each other. The latter situation has been studied in both the random diluted [40] and fully connected [56] Hopfield model. In both cases it gives rise to 3 distinct phases: at low α there is a separate memory for each pattern; as α is increased there is a single memory equally correlated with both patterns and at higher α there is no memory. In [40] the range of α values for the undistinguishing memory includes values that are greater than α_c for the recall of the random patterns. This also applies at the higher values of the correlation between the two patterns in [56]. This implies that as the storage becomes large the network will forget patterns that have not been reinforced but still remember those that have. These ideas have been extended within the Gardner framework, discussed in chapter 3, by Virasoro[57].

However all these studies focussed on the case of random unbiased patterns. In the case of biased patterns the patterns are already correlated by the bias so that any two stored patterns share a large number of 0-sites. One may interpret the increase in storage capacity as a result of this correlation [57]. In this section we shall see investigate whether the recall of a pattern (taken to be pattern 1) can be further improved by storing an additional pattern (taken to be pattern 2) that

has a large number of 1-sites in common with it. The overlap within the 1-sites can be quantified by introducing a parameter Q,

$$Q = \frac{1}{Nf}\sum_{i=1}^{N}\eta_i^{(1)}\eta_i^{(2)}. \tag{4.36}$$

If patterns 1 and 2 are identical we find $Q = 1$ whereas if the one sites are uncorrelated we find $Q = f$. The method used is to derive evolution equations for the overlaps with the two patterns of interest. These overlaps,

$$\begin{aligned} m_1^{(1)} &= \frac{1}{Nf}\sum_{i=1}^{N}\eta_i^{(1)}V_i \\ m_1^{(2)} &= \frac{1}{Nf}\sum_{i=1}^{N}\eta_i^{(2)}V_i \\ m_2^{(1)} &= \frac{1}{Nf}\sum_{i=1}^{N}(1-\eta_i^{(1)})V_i \end{aligned}$$

are defined in the same way as before but have the pattern index reinstated. Again it is convenient to use the activity ,

$$y(t) = m_1^{(1)} + m_2^{(1)}, \tag{4.37}$$

as a parameter. The parameter $m_2^{(2)}$ is not used explicitly because

$$\begin{aligned} m_2^{(2)} &= y(t) - m_1^{(2)} \\ &= m_1^{(1)} + m_2^{(1)} - m_1^{(2)}. \end{aligned}$$

We let patterns 1 and 2 have overlap $Q \sim O(1)$ and all other pattern pairs be uncorrelated. In deriving evolution equations for the 3 order parameters required, one uses the same techniques outlined in the previous sections and so here we will just write down the zero temperature result in the regime $f \ll 1$.

$$\begin{aligned} m_1^{(1)}(t+1) &= \frac{Q}{2}\left[1 + \text{ erf}\left(\frac{m_1^{(1)}(t) + m_1^{(2)}(t) - \theta_0}{\sqrt{2\alpha f y(t)}}\right)\right] \\ &\quad + \frac{1-Q}{2}\left[1 + \text{ erf}\left(\frac{m_1^{1}(t) - \theta_0}{\sqrt{2\alpha f y(t)}}\right)\right] \\ m_1^{(2)}(t+1) &= \frac{Q}{2}\left[1 + \text{ erf}\left(\frac{m_1^{1}(t) + m_1^{2}(t) - \theta_0}{\sqrt{2\alpha f y(t)}}\right)\right] \end{aligned}$$

$$+\frac{1-Q}{2}\left[1+\ \mathrm{erf}\left(\frac{m_1^2(t)-\theta_0}{\sqrt{2\alpha f y(t)}}\right)\right] \qquad (4.38)$$

$$m_2^{(1)}(t+1) = \frac{1}{2f}\left[1-\ \mathrm{erf}\left(\frac{\theta_0}{\sqrt{2\alpha f y(t)}}\right)\right]$$
$$+\frac{1-Q}{2}\left[1+\ \mathrm{erf}\left(\frac{m_1^{(2)}(t)-\theta_0}{\sqrt{2\alpha f y(t)}}\right)\right]$$

The equations for $m_1^{(1)}(t+1)$ and $m_1^{(2)}(t+1)$ are transformed into each other when $m_2^{(1)}$ and $m_2^{(2)}$ are interchanged. This symmetry means that any solution of the fixed point equations of the three dimensional map (4.38) will remain a solution when the fixed point values of $m_2^{(1)}$ and $m_2^{(1)}$ are interchanged.

To construct a phase diagram the criterion used is that starting from one of the patterns the fixed point to which the map (4.38) iterates classifies the phase to which the α, Q values belong. This is equivalent to using an α_1 definition for the critical value of α. The distinguishing memory phase is characterised by a fixed point of the form

$$\begin{aligned} m_1^{(1)} &\simeq 1 \\ m_1^{(2)} &\simeq Q \\ m_2^{(2)} &\ll 1, \end{aligned}$$

the undistinguishing memory phase is characterised by

$$\begin{aligned} m_1^{(1)} = m_1^{(2)} &\simeq Q \\ m_2^{(2)} &\ll 1 \end{aligned}$$

and the no memory phase is characterised by the high activity attractor or all zeros fixed point. Fig.4.6 shows that only at comparatively high Q values is the undistinguishing memory phase present. In fact there is only a small range of Q values where all 3 phases are present. Except at very low Q values the storage capacity is *reduced* compared with Section 2.

These observations reflect the importance of the threshold parameter and activity of a configuration. A memory that does not distinguish between the patterns ($m_1 = Q$) will have too low an activity to be a fixed point at low Q. Although a Q overlap between patterns will strengthen the connections between shared 1's,

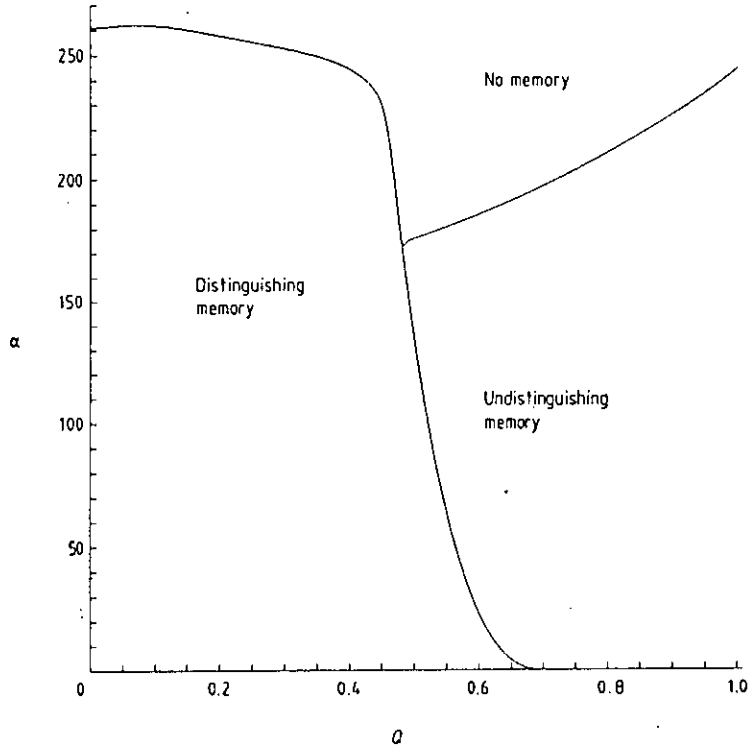

Figure 4.6: A phase diagram depicting the three phases discussed in the text.

this will lead to higher activities in trajectories near to the patterns because m_1 will be increased. This may destabilise, through the resulting increase in y and therefore noise, the memories that distinguish between the patterns at lower α than before. Consider the case of storing a pattern twice ($Q = 1$). Equations (4.38) reduce to (4.25) if

$$\begin{aligned} \theta_0 &= 2\theta_0' \\ \alpha &= 4\alpha' \end{aligned}$$

where the primed values are those for storing the pattern once. Hence an increase in the maximum α for recall of the pattern is possible only if θ_0 is suitably increased.

4.6 Discussion of Results of Covariance Rule

So far in this chapter several interesting results that involve biased pattern networks been obtained. Although the calculations have centered on the covariance rule (4.3), the results may give some general properties for models that use biased patterns and 1,0 spins.

Firstly we have seen that as well as memory fixed points there are two other attractors to which the configurational flow may lead. These are the all zeros fixed point

an high activity attractor. The all zeros fixed point may be considered useful for, as Shinomoto[58] and Buhmann *et al* [52] have pointed out, if the configurational flow is towards this fixed point then it is a signal that initial configuration was not recognised by the network. I consider this argument appealing in the case of the all zeros fixed point because the signal of non-recognition is the activity within the network dying out. When the activity has died out the network is then ready for processing a new input. Whereas for the high activity attractor the signal of non-recognition is the activity proliferating. This leads to very high spatial activity which for a signal of non-recognition appears to be much ado about nothing. Although the high activity fixed point can co-exist with the memory attractors, its presence is not biologically satisfactory as it indicates that the spatial activity is not constrained in the model. A suitable modification to rid the model of such a fixed point could be to introduce an activity dependent threshold. This might also improve content addressable memory because the region of the m_1–m_2 plane that comprises the basin of attraction of the high activity attractor would become within the basin of attraction of either the all zeros or memory attractor. Implementations of such a proposal will be discussed later in this chapter and chapter 5.

The mathematical structure of the memory fixed point changes with the threshold. This was initially noted in the qualitative from of $\alpha_c(\theta_0)$, equations (4.30) and (4.32). Later we saw, in figures 4.2 and 4.3, that the nature of the entry of trajectories to the attractor changes with θ_0. Indeed in the intermediate range, between low and high θ_0, the attractor could become a limit cycle rather than a fixed point in the m_1–m_2 plane. These limit cycles are rather appealing because within them the activity y oscillates somewhat and oscillations in spatial activity within the brain are well known [3]. However there is some doubt as to whether these limit cycles would sustain themselves even in the randomly diluted model. This is because the system does not reach a fixed point in finite time t therefore the condition

$$C^t \ll \sqrt{N}, \tag{4.39}$$

cannot hold as $t \to \infty$. On the other hand the mean-field equations for the fully connected model are equivalent to the map (4.25) so the condition (4.39) may not in fact be necessary. In order to determine if the limit cycles do exist numerical simulations would have to be carried out.

A parameter that has attained a high importance in the present chapter is the threshold. The threshold must be carefully chosen in order to gain the maximal storage capacity. Near to this optimal threshold value small changes in the threshold change the characteristics of the trajectories in the order parameter space considerably. The characteristics of the trajectories define the basin of attraction of the memory, thus the threshold determines both the storage capacity and associativity. In addition section 4.5 demonstrated that storing two highly correlated patterns is only only fruitful when the threshold is suitably adjusted. This would then rule out the recall of the other independent patterns. All of these points suggest that in future the threshold may have to be modelled with more sophisticated properties such as site or time dependence.

Several surprising features of the model have also been illustrated. Firstly the fixed point equations of the map (4.25) are equivalent to the mean-field equations of the fully connected model, implying that random dilution has not changed the features of the model to a great extent. Tsodyks[59] also noted this "universality" for the continuous-time model. Secondly in considering the storage capacity there is a need for careful definition of the critical value of α, due to the possibilities of intermittancy and the pattern being outwith the basin of attraction of the memory.

In the next section we shall discuss another connection rule for storing biased patterns and investigate which of the features discussed here in the context of the covariance rule persist when a different connection rule is used. In particular the rôle of the threshold and existence of an all zeros fixed point and high activity attractor shall be highlighted.

4.7 The Willshaw Rule

4.7.1 Willshaw's Rule and Dale's Principle

Willshaw's Associative Network [10] was first introduced as a device that learned a set of associations, each between an input pattern and an output pattern. In contrast to the models discussed previously in this work, the synaptic connections

can only take values 1 or 0. Willshaw *et al*[10] found that the network is most efficient when the patterns are extremely biased. It was found that at the optimal value of f

$$f = \frac{1}{\ln 2}\frac{\ln N}{N} \tag{4.40}$$

the network has an information efficiency of 69%. This remarkably high information capacity alone justifies further study of the Willshaw rule. However the rule has several other features of merit from a biological viewpoint.

Let us first review the Willshaw rule and how Willshaw's Associative Net functions. We will consider without loss of generality the case where the Network is auto-associative. Basically a synaptic connection between 2 neurons is made if both neurons are active in at least one pattern that has been stored. Willshaw's rule is

$$J_{ij} = \Theta\left(\sum_{\mu=1}^{P} \eta_i^\mu \eta_j^\mu\right) \tag{4.41}$$

where P is the number of nominated configurations and

$$\begin{aligned} \Theta(x) &= 1 \quad \text{when} \quad x > 0 \\ &= 0 \quad \text{when} \quad x \leq 0 \end{aligned}$$

One may note that all the synaptic connections are positive which is in contrast to the covariance rule where approximately equal numbers of connections were positive and negative. This feature of the Willshaw rule has some biological appeal due to a biological observation known as Dale's Principle [60]. Dale's principle states that the synaptic connections emanating from a particular neuron are either all excitatory or all inhibitory. Interpreting this in terms of model synapses and neurons one can say that for a neuron i the set of synapses $\{J_{ki}\}$, where the index k runs through all the neurons k that i has a synaptic connection with, must all be of the same sign. The Willshaw rule satisfies this condition because all the synaptic connections emanating from any neuron are positive. One can say that the Willshaw rule creates a neural network in which all the neurons are excitatory. Dobson[61] has shown that a simple variation of the Willshaw rule creates a network where all the neurons are inhibitory, without deterioration of performance.

4.7.2 Willshaw's Analysis and its Validity

If one considers the setting up of the synaptic connections by storing the patterns in sequence as a dynamic learning process, then before learning has begun the synaptic connection matrix has all elements set to zero and as the learning proceeds more and more connections are switched on. This process can continue until the network is saturated at which point the network can no longer make the correct associations.

Willshaw[10] gave a simple analysis of when this saturation occurs. We denote the probability of a connection being on (i.e $J_{ij} = 1$) by c which is given by

$$\begin{aligned} c &= 1 - (1 - f^2)^P \\ &= 1 - \exp(-\gamma) \text{ as } P \to \infty \text{ and } f \to 0 \end{aligned} \tag{4.42}$$

where

$$\gamma = Pf^2. \tag{4.43}$$

It is convenient to use γ as the loading parameter rather than α because at maximum loading the number of patterns is not proportional to the number of neurons.

The field at a site i is defined as

$$h_i(t) = \sum_{j \neq i} J_{ij} V_j(t) - \theta \tag{4.44}$$

The original Willshaw model worked by performing one parallel update according to the rule:

$$V_i(t+1) = \mathrm{L}(h_i), \tag{4.45}$$

where

$$\begin{aligned} \mathrm{L}(x) &= 1 \quad \text{when} \quad x \geq 0, \\ &= 0 \quad \text{when} \quad x < 0. \end{aligned} \tag{4.46}$$

If the configuration of the network is one of the nominated patterns, then we shall say the network is " in a pattern ". The connection rule ensures that connections are on between sites that are both 1 in a pattern. When the network is in a pattern the 1-sites will therefore receive a field of exactly Nf. This allows the threshold to be set at

$$\theta = Nf \tag{4.47}$$

which is the maximum it can be whilst ensuring that the 1-sites are correctly recalled.

If we turn to the sites that are 0-sites in a pattern the situation is somewhat different. These passive sites have not caused any modification to the synapses (4.41) and so when the network is in a pattern there is no signal component to the local field. One can say that the field these sites do receive is composed entirely of noise due to synaptic interference. However this noise has a slightly different character than the noise component of the local field in the previous section (4.15), because the mean is not zero.

Willshaw defined the capacity of the network as the number of patterns stored when this noise will cause a 0-site to be incorrectly recalled as a 1. For this to happen when the threshold is given by (4.47), all the 1-sites must be connected to a particular 0-site. The probability of this happening if the connections are independent is c^{Nf}. The capacity of the network is then given by

$$N(1-f)c^{Nf} = 1. \tag{4.48}$$

The value of c that solves (4.48) can be converted into a storage level γ_1 through (4.42). γ_1 gives the critical loading implied by Willshaw's analysis. The l.h.s. of (4.48) is the mean number of 0-sites which are connected to all Nf 1-sites assuming that the connections are independent. If the conditions $f \ll 1$ and $N \rightarrow \infty$ are assumed (4.48) gives

$$f = -\frac{\ln N}{N \ln c} \tag{4.49}$$

When c the fraction of connections on is finite, equation (4.49) shows that the pattern associator only functions well in the extremely sparse coding limit of

$$f = f'\frac{\ln N}{N} \tag{4.50}$$

To measure the efficiency of the network Willshaw considered the information stored I. For perfect recall this is given by

$$I = P \ln \left(\begin{array}{c} N \\ Nf \end{array} \right) \tag{4.51}$$

$$\simeq PNf\left[\ln\left(\frac{1}{f}\right)\right] \quad \text{For } N \gg Nf \tag{4.52}$$

where the binomial coefficient

$$\left(\begin{array}{c} N \\ M \end{array} \right) = \frac{N!}{M!(N-M)!}. \tag{4.53}$$

Utilising (4.42) and (4.49) one finds the information at the limit of capacity as:

$$I = N^2 \ln(1-c) \ln c$$

which has a maximum at $c = 1/2$ at which the information stored per synapse is $(\ln 2)^2$. The maximum information that can be stored per synapse is $\ln 2$ so the network functions at $\ln 2$ or 69% efficiency. At this value of c and at the capacity of the network

$$P = (\ln 2)^3 \left(\frac{N}{\ln N} \right)^2 \tag{4.54}$$

To show that this obeys the Gardner bound one can insert $f \sim \ln(N)/N$ into (4.2) to obtain

$$\alpha_c \sim \frac{N}{(\ln N)^2}. \tag{4.55}$$

This analysis relies on the assumption that the synaptic connections are independent of each other. Goloumb *et al* have shown that the correlation between two synapses J_{ij} and J_{ik} feeding into the same post synaptic neuron i is given by

$$\langle (J_{ij} - < J >)(J_{ik} - < J >) \rangle = \exp\left(-2Pf^2\right) Pf^3 \tag{4.56}$$

so that the correlation increases with f.

The analysis concentrates on the perfect recall of patterns. More general memory attractors will be considered in the next section. The case of exact storage can be investigated by simulation in a relatively straightforward manner. The simulation amounts to presenting a network with a stored pattern and checking whether the spins at each site are aligned to the thresholded fields. The programme is structured in the following way.

1. A number of patterns that obey the activity constraint $\sum_{i=1}^{N} V_i = Nf$ are generated and stored according to rule (4.41).

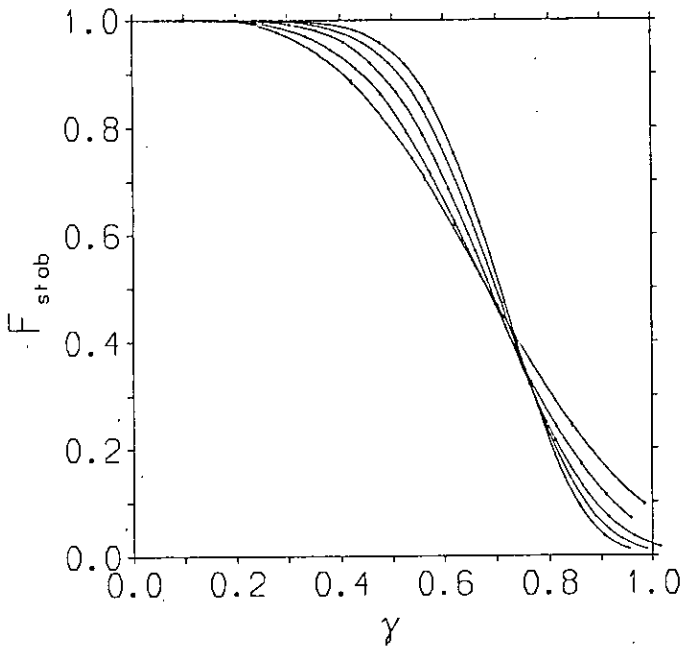

Figure 4.7: The fraction of patterns stored exactly in the net vs. γ for $f' = 1/\ln(2)$ and $\theta_0 = 1$.
System sizes are: 64; 128; 256; 512; 1024 where the curves steepen as system size increases. The curves are best fit polynomials through the points.

2. All the patterns so far stored are tested for stability by presenting them to the network and testing each site to see if it is correctly aligned to the thresholded molecular field, i.e whether or not

$$(2\eta_i^\mu - 1)\,(2\mathrm{L}(h_i) - 1) > 0, \tag{4.57}$$

where $\mathrm{L}(x)$ is defined by (4.46). If a site is encountered that is incorrectly aligned then the pattern μ is unstable. After testing all the patterns in this manner one can calculate the fraction of a patterns that are stable F_{stab} as for the present value of γ.

3. More patterns are then stored and the calculation of F_{stab} is repeated. In this way one obtains the value of F_{stab} at several points in the learning process for the particular realisation of the patterns.

4. All the previous steps are repeated 20 times so that averaged values of $F_{stab}(\gamma)$ are obtained.

This computer experiment was repeated for different system sizes as shown in fig(4.7) In fig(4.7) the transition through which F_{stab} changes from 1 to 0 with increasing γ can be seen. The transition sharpens slowly as N increases. $\gamma_1 =$

0.693, given by the analysis of section 4.7.2 is within the transition range of γ. It appears to very close to the values of γ at which the different system size curves intersect. In order to understand the full meaning of γ_1 Willshaw's analysis has to be extended a little. This will also allow a finite size scaling form for the curves in Fig. 4.7 to be developed.

4.7.3 Finite Size Scaling

Eq.(4.48) specifies that the mean number of unstable 0-sites is equal to 1. The l.h.s gives the mean number of unstable zero sites in a pattern. We can model the actual number of such sites with a Poisson distribution so that $P_s(k)$, the probability of k sites being unstable in a pattern is given by

$$P_s(k) = \frac{\lambda^k e^{-\lambda}}{k!}, \tag{4.58}$$

where λ is the l.h.s of Eq. 4.66. The probability of a pattern being stable is then P_p which is given by

$$P_p = P_s(0) = e^{-\lambda} \tag{4.59}$$

At γ_1 the probability of a pattern being stable is $e^{-1} = 0.378$ and is independent of system size. This explains the position of γ_1 as the point where the different system size curves intersect in Fig 4.7. However in Fig 4.7, for $N = 1024$, γ_1 has $F_{stab} \approx 0.5$ rather than 0.378 which suggests that the assumed independence of the stability of patterns and of stability of sites within patterns, used in the derivation of (4.59), is not accurate. The general expression for P_p is obtained by inserting the l.h.s of (4.48) into (4.59):

$$P_p = \exp - \left(N^{(1+f'\ln(1-\exp-\gamma))}\right). \tag{4.60}$$

Eq. (4.60) implies that as $N \rightarrow \infty$, values of γ less than γ_1 will yield $P_p = 1$ whereas values greater than γ_1 will yield $P_p = 0$. This is because γ_1 is the value at which the power that N is raised to on the r.h.s. of (4.60) is equal to zero. Therefore as $N \rightarrow \infty$, the transition from $F_{stab} = 1 \rightarrow 0$ should become discontinuous. There is no direct evidence for this in Fig 4.7, and it appears that extremely large values of N would have to be simulated in order to verify the discontinuity. However the finite size scaling forms of the curve may be investigated to verify

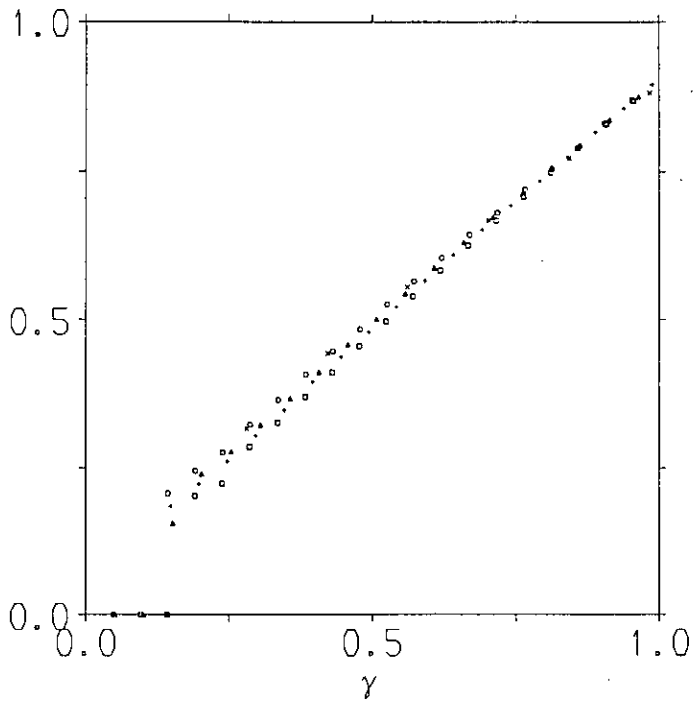

Figure 4.8: The data of Fig 4.7 is plotted according to the finite size scaling form (4.60). The system sizes are: 64 (crosses); 128(circles); 256(triangles); 512(plus signs); 1024(squares).

the validity of (4.60). This is carried out by arranging (4.60) to give

$$\gamma = -\ln\left[1 - \exp\left\{\frac{1}{f'}\left(\frac{\ln(1/P_p)}{\ln N} - 1\right)\right\}\right]. \tag{4.61}$$

In Fig. 4.8 the numerical values of P_p from Fig.4.8 are inserted into the r.h.s. of 4.61 and plotted against γ. The different system sizes do indeed collapse onto the same line as γ increases but the line has gradient 1.4 rather than 1. This suggests that

$$P_p = \exp - \left(N^{(1+f'\ln(1-\exp(-1.4\gamma)))}\right) \tag{4.62}$$

which is less than the expression Eq. 4.60. This indicates that the actual value of P_p is smaller than that predicted by the analysis. This discrepancy can be attributed to the correlations between synapses. As f increases equation (4.56) suggests that the effect of such correlations should become more important in calculating critical storage capacities.

4.7.4 The Effect of Lowering the Threshold

In the analysis of the covariance rule we saw that to achieve the maximum storage capacity the threshold must be suitably chosen. With the Willshaw rule a very specific choice of threshold has been made namely equation (4.47). If the threshold is chosen to be higher than this then the field received in a pattern by a 1-site will be less than the threshold, so that the pattern will not be correctly recalled. However if the threshold is chosen as

$$\theta = \theta_0 N f \tag{4.63}$$

where θ_0 is less than unity then the 1-sites will still be correctly recalled. The processing of noisy inputs, where some 1-sites have been flipped to zero and the signal at 1-sites brought below Nf, could then be achieved. In order that the 1-sites be correctly recalled one would have to choose $\theta_0 < 1$. This illustrates how changing the threshold alters the basin of attraction of the memories. In order to determine how changing the threshold alters the storage capacity we will generalise Willshaw's analysis to calculate the condition for a pattern to be a recalled exactly when $\theta_0 < 1$. Again we shall consider the limit of stability of the pattern as the loading at which on average a 0-site will become unstable. This occurs when:

$$N \sum_{x=\theta_0 Nf}^{Nf} \left(\begin{array}{c} Nf \\ x \end{array} \right) c^x (1-c)^{(Nf-x)} = 1. \tag{4.64}$$

The l.h.s. of Equation (4.64) is the probability that a 0-site is connected to more than $\theta_0 Nf$ 1-sites multiplied by N. For $f \ll 1$ and with the assumption that the connections are uncorrelated the l.h.s is equal to the mean number of 0-sites that are unstable when the network is in a pattern.

If the $x = \theta_0 Nf$ term of the sum in Eq. (4.64) is taken as the dominant contribution then Eq. 4.64 becomes

$$N \left(\begin{array}{c} Nf \\ \theta_0 Nf \end{array} \right) c^{\theta_0 Nf} (1-c)^{(1-\theta_0)Nf} = 1. \tag{4.65}$$

Taking the natural log of this equation and using Stirling's approximation gives

$$1 - \theta_0 f' \ln\left(\frac{\theta_0}{c}\right) - (1-\theta_0) f' \ln\left(\frac{1-\theta_0}{1-c}\right) = 0 \tag{4.66}$$

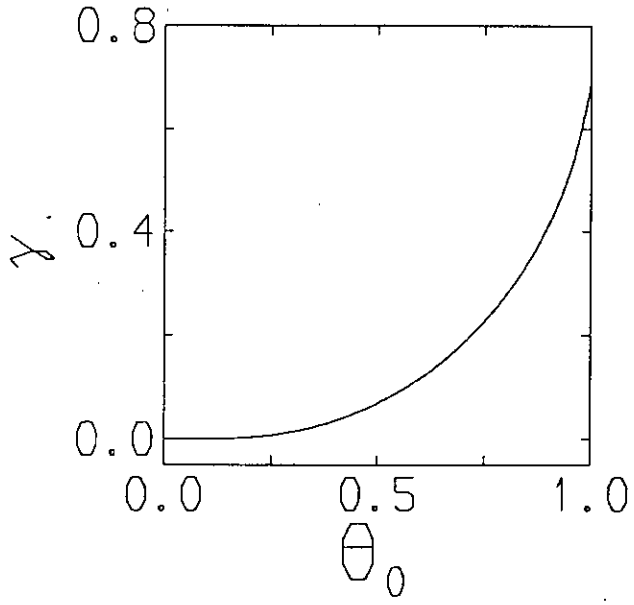

Figure 4.9: A plot of γ_1 against θ_0 for $f = 1/\ln(2)$. The curve was obtained from the solution of (4.66).

Eq(4.66) gives a relationship between θ_0, f', c for the limit of stability of patterns. If l.h.s > 0 then the pattern is unstable to 0-sites flipping to 1's. Equation (4.66) can be solved to give γ_1 through (4.42). Fig. 4.9 illustrates the dependence of γ_1 on the threshold. A dramatic decrease in storage can be seen as the threshold is lowered. For example when θ_0 is reduced to 0.9, γ_1 is reduced from 0.693 to 0.412.

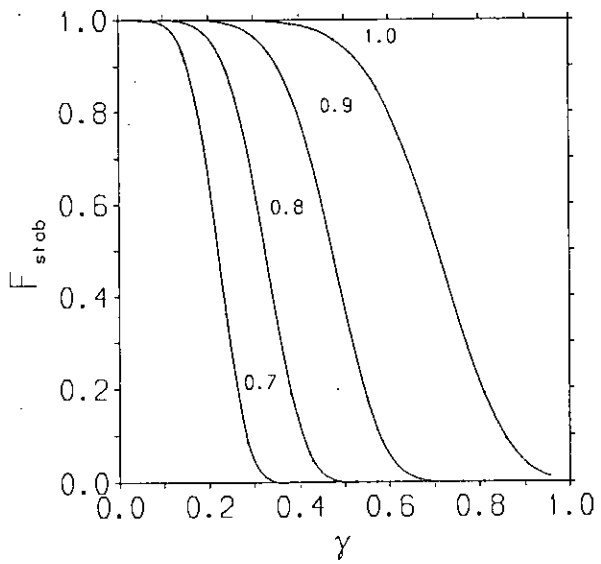

Figure 4.10: Simulation results for different thresholds with the system size constant at $N = 1024$. The fraction of patterns stored exactly in the net vs. γ for $f' = 1/\ln(2)$ and for the different θ_0 marked on the figure. The curves are best fit polynomials through the points.

where f' is given by (4.50).

Figure 4.10 presents simulations of the same type as figure 4.7 but with $\theta_0 < 1$. The simulation results show, as did equation (4.66), that the storage capacity decreases sharply with θ_0.

4.7.5 Gaussian Noise Analysis

If we examine the above analysis we can see that the probability of the noise at a 0-site is a binomial distribution. Eq. (4.66) shows that the probability that a 0-site is unstable in a pattern is given by the area under the tail of the distribution cut off at $\theta_0 Nf$. In the previous chapters we have represented synaptic noise as a Gaussian distribution by taking the cumulant expansion to the second cumulant. If we do this in the present case then we obtain instead of (4.64)

$$N \int_{\theta_0 Nf}^{\infty} \frac{\mathrm{d}y}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\overline{R})^2}{2\sigma^2}\right) = 1 \tag{4.67}$$

where the mean of the noise $\overline{R}$ is given by

$$\overline{R} = Nfc. \tag{4.68}$$

and the variance of the noise σ^2 is given by

$$\sigma^2 = Nfc(1-c)\,. \tag{4.69}$$

We may write the l.h.s of (4.67) as a Gauss error function and expand it for a large argument, using (3.48), to give

$$\frac{\sqrt{\sigma^2}}{\sqrt{\pi}(\theta_0 - \overline{R})} N^b \tag{4.70}$$

$$\text{where} \qquad b = 1 - \frac{f'(\theta_0 - c)^2}{2c(1-c)} \tag{4.71}$$

For $b < 0$, (4.70) vanishes in the thermodynamic limit and for $b > 0$ it diverges. (4.70) represents the number of one sites that have flipped to ones after 1 parallel update, thus $b < 0$ indicates that the pattern is stable. The transition to no memory then occurs at $b = 0$.

To judge the accuracy of this signal and Gaussian noise analysis one can take $\theta_0 = 1$ and find $\gamma = \ln(1 + f'/2)$. Using the Willshaw analysis one finds $\gamma = 1/f'$.

This error highlights the difficulties in studying the model analytically. In a mean field theory fluctuations are usually approximated by a Gaussian distribution, but here we see that this approximation gives an inaccurate result. This is because it is the tail of the noise distribution that is important for the stability of 0-sites, and the tail of a binomial distribution is not accurately approximated by a Gaussian.

4.8 Attractors in the Willshaw Network

In dealing with the Willshaw rule so far we have only investigated whether the patterns are stable. To analyse this point we have considered one parallel update of the sites. In general the configuration of the network after such an update is not a fixed point of the dynamics. It is only when the configuration has remained unchanged, as is the case when a pattern is stable, that an attractor has been reached. It may be the case that the patterns are themselves unstable but there are attractors near to the patterns. This has usually been the case in the models discussed in the previous chapters. If the pattern is not stored exactly we wish to know if the network configuration will evolve to such a memory attractor when a pattern is presented as the initial configuration. We also have to consider the case of noisy inputs and to which attractor they will evolve. The main aim of this section are to determine how Willshaw's rule will perform in the context of ANN's. More precisely how will the new features of recurrency, the need for attractors and asynchronous dynamics affect the performance? Will the model need to be modified in anyway to deal with these new features?

In the previous section we in fact considered $m_1(1)$ and $m_2(1)$ the order parameters after one time step. The condition for stability was that after presenting $m_1(0) = 1$ and $m_2(0) = 0$ we demanded $m_1(1) = 1$ and $m_2(1) = 0$. For an ANN we are interested in $m_1(\infty)$ and $m_2(\infty)$ that result from more general inputs.

The attractors uncorrelated with the patterns that were present when the covariance rule was used are also present when we study the Willshaw rule. If we consider the all zeros configuration where each site takes value 0 then the local field given by (4.44) is zero at all sites. Therefore this configuration is stable if $\theta_0 \geq 0$. We can also see that the an high activity attractor is present by examining

(4.44) for the case where all sites take value 1. All the local fields are greater than zero, thus the configuration is stable, provided that $Nc \gg \theta$. This is the case when c is finite and f is given by (4.50).

The remaining possibility is for memory attractors that are not the exact patterns. In this case the errors in the memory attractor must be incorrect ones. This is because if the threshold is chosen correctly, $\theta_0 \leq 1$, the 1-sites will all be recalled correctly. However if one considers iterating from a pattern, then if some 0-sites flip to 1s the activity has increased and the probability of other 0-sites flipping to 1s will increase. In this way a finite number of zero sites flipping to 1s will cause the activity to proliferate and the high activity attractor should be reached.

4.8.1 Suppressing the High Activity Attractor

In order to rid the network of any high activity fixed point one needs a mechanism that inhibits the production of 1s in the dynamics. A constant threshold is already present but we require an additional activity dependent threshold that will vary in the configurational flow produced by the dynamics:

$$\theta = (\theta_0 N f + \theta_1(y))$$

The mean noise at 0-sites increases linearly with the activity and we wish θ_1 to exhibit the same behaviour:

$$\begin{aligned} \theta_1 &= byNf \\ &= b\sum_i V_i. \end{aligned}$$

One can consider this inhibition as resulting from an auxiliary network of inhibitory neurons, the activity within which is directly proportional to the activity of the excitatory network. However this second interpretation, used by Golomb *et al* [62], requires that the inhibitory neurons respond instantaneously to changes in the activity of the excitatory network. Alternatively the activity dependent threshold is equivalent to changing the Willshaw rule to make the synapses that have not been switched on inhibitory:

$$J_{ij} = (1+b)\Theta\left(\sum_{\mu=1}^{P} \eta_i^\mu \eta_j^\mu\right) - b \qquad (4.72)$$

Before any learning, when all synapses take value $-b$, the only fixed point is the all zeros configuration. As the learning proceeds and more and more synapses are switched on the inhibition is gradually removed. If we take the learning to an extreme where all synapses are on then the the domain of attraction of the high activity attractor comprises nearly all the configuration space except the portion with $y < \theta_0$. At some point in the learning, or equivalently at some value of c, a high activity attractor must become stable. If the average connection strength is positive, a high activity attractor with $\sum_i V_i \sim N$ exists. For $b = 1$ the average interaction strength is positive when $c > 0.5$. However high activity attractors with $\sum_i V_i \sim N^g$, where $0 < g < 1$, may occur in the learning at an earlier time.

To evaluate when this occurs we consider an initial configuration of activity $y(0)$ and calculate the activity at the next time step $y(1)$. If

$$y(1) > y(0), \tag{4.73}$$

the activity is increasing and we assume that a high activity attractor exists. Conversely if we cannot find a value $y(0)$ where (4.73), holds then we conclude that no high activity attractor exists. The analysis of section 4.7.4 can be extended in a straightforward manner to give

$$y(1) \sim N^{g(y(0))} \tag{4.74}$$

where

$$g(y) = 1 - uf' \ln\left(\frac{u}{yc}\right) - (y-u)f' \ln\left(\frac{y-u}{y(1-c)}\right) \tag{4.75}$$

and

$$u = \frac{\theta_0 + by}{1+b}. \tag{4.76}$$

The activity y^* that maximises $g(y)$ is given by the equation

$$\left(\frac{y^* c(1+b)}{u_0 + by^*}\right)^b \left(\frac{y^*(1-c)(1+b)}{y^* - u_0}\right) = 1. \tag{4.77}$$

This value of y^* in fact maximises the probability that $h_i = \theta_0$. For $b = 1$, (4.77) has solution

$$y^* = \frac{\theta_0}{\sqrt{1 - 4c(1-c)}}. \tag{4.78}$$

This is the value of the activity that maximises the activity at the next time step. $g(y^*(c)) = 0$, then gives the connectivity value c_{haa} at which a high activity

attractor appears. This equation can be solved numerically to give, for $b = 1, \theta_0 = 1$

$$c_{haa} = 0.333, \tag{4.79}$$

$$y^* = 3.001. \tag{4.80}$$

The value of c_{haa} given by (4.79) yields a value of γ through equation (4.42). This value of γ is less than γ_1, the storage at which the patterns become unstable. Therefore a high activity fixed point occurs before the Willshaw critical loading γ_1 is reached. However, (4.80) shows that for iteration to this high activity fixed point at c_{haa}, an activity thrice that of the patterns is required. In view of this one would not expect the high activity attractor to affect the dynamics initially, when iterating from a stored pattern.

The upshot of this calculation is that the high activity attractor has been suppressed. This is demonstrated most simply by the result that for $b = 1$, the high activity attractor appears only when $C \geq 0.333$. This is in contrast to $b = 0$ when the high activity attractor is always stable. When $b = 1$ the previous paragraph has hinted that configurational flows that start to move away from a stored pattern may not reach an high activity attractor even when one is present. To understand why this is so let us consider again the case $b = 0$. If the pattern is unstable, 0-sites flip to 1s and the activity proliferates until the high activity attractor is reached. However in the case where $b > 0$ the flipping of a 0-site to 1 may cause a decrease in the local field at other sites because the connections from the 0-site to these other sites may be inhibitory. Therefore the proliferation of activity may be halted. This argument suggests that when $b > 1$, memory attractors which do not coincide exactly with the stored pattern may be present. In order to ascertain whether this actually occurs I will present some numerical simulations.

4.8.2 Numerical Simulations of Attractor Structure

In order to investigate the effect on the attractor structure of the introduction of inhibitory interactions a similar method to that used to calculate the fraction of exactly stable patterns is employed. The computation is performed as follows.

1. A number of pattern are stored by the connection rule (4.72).

2. A pattern, μ, is presented to the network which is then allowed to evolve according to the serial dynamics with updating rule (4.45) until a stable configuration is reached. The foreground overlap with the pattern $m_1^\mu(\infty)$ and the activity $y^\mu(\infty)$ of this stable configuration are then computed.

3. This process is repeated for all patterns stored so far in order that the average of the overlap between a configuration associated with a pattern by the dynamics and that pattern can be computed.

$$< m_1(\infty) > = \frac{1}{P}\sum_{\mu=1}^{P} m_1^\mu(\infty) \tag{4.81}$$

$$< y(\infty) > = \frac{1}{P}\sum_{\mu=1}^{P} y^\mu(\infty) \tag{4.82}$$

4. This scheme is repeated for increasing system size and the appearance of a discontinuity in one of the averaged parameters is taken to indicate a critical storage level γ_2.

γ_2 in fact measures the storage level at which trajectories in the m_1–m_2 plane near to the pattern begin to be captured by a high activity attractor. This is because a single trajectory that reaches a high activity fixed point will make a large contribution to $< y(\infty) >$. Therefore γ_2 measures the storage level at which a finite number of patterns become unstable to the high activity attractor. This is contrast to γ_1, the capacity as given by Willshaw's analysis, which theoretically should predict the storage level when 0.38 of the patterns remain exactly stable.

Figure 4.11 shows a sudden change in the value of the associated activity. Whereas the averaged foreground overlap remains large until a higher loading has been reached. Of particular interest is the fact for $\theta_0 = 1$ the 2 levels of inhibition simulated show different behaviours which contrasts with the pattern associator where the behaviours are identical. This indicates that the use of $b = 1$ has suppressed the onset of the high activity attractor until higher storage levels, as was predicted by the analysis of section 4.8.1. We also note that the transition value γ_2 suggested by the simulations, is in both inhibition cases below the γ_1 value.

In order to examine the attractor structure in more detail we can analyse the computer data in a slightly different way. Figure 4.12 presents histograms of

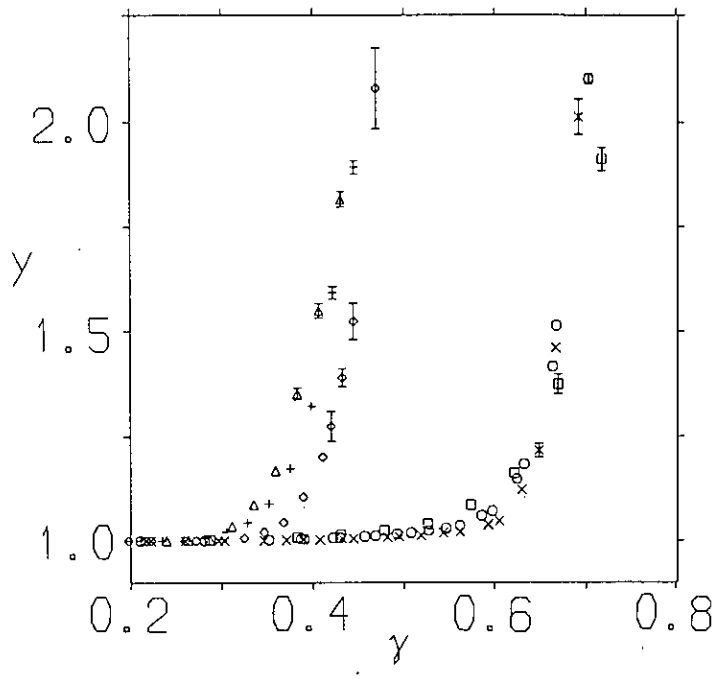

Figure 4.11: $< y(\infty) >$ (see (4.82)) vs. γ for $f' = 1/\ln(2)$ and $\theta_0 = 1$.
For $b = 1$ the system sizes are: 128(squares); 256(circles); 512(crosses)
For $b = 0$ the system sizes are: 128(triangles); 256(plus signs); 512(diamonds)
Errorbars are shown when they are larger than symbol size.

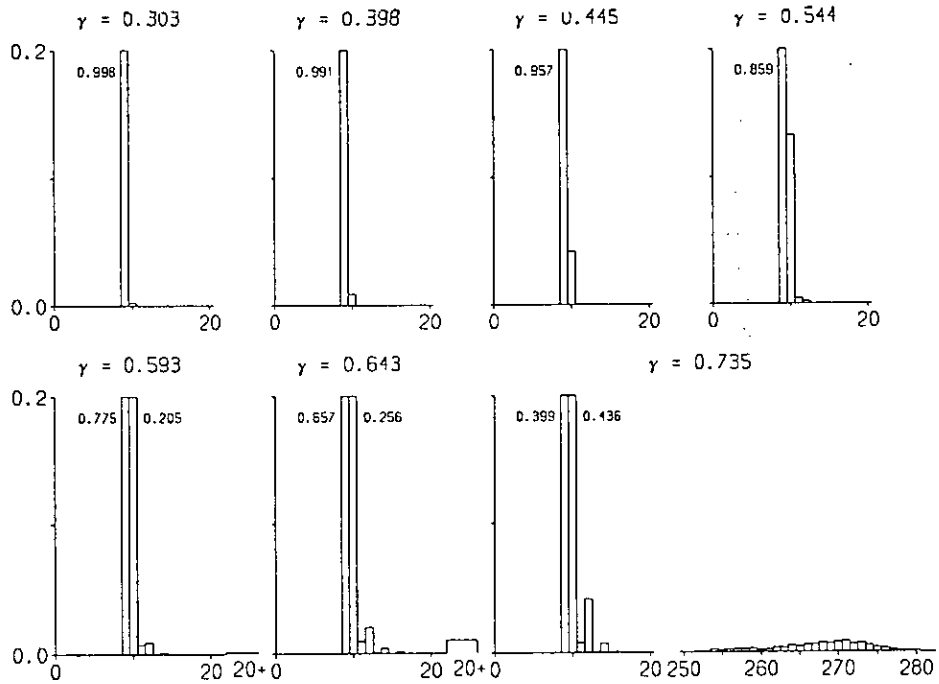

Figure 4.12: Histograms of number of 1s in configurations iterated to from patterns when $b = 1$. Each bar of the histogram is the fraction of configurations with the corresponding activity. When a bar leaves the frame the value it takes is written adjacent to the bar. The number of systems averaged over was 10. $N = 512$ and $f' = 1/\ln 2$ so that each pattern contains 9 1-sites; $\theta_0 = 1$.

the activities of the attractors that have been reached for the case of $b = 1$. At $\gamma = 0.303, c = 0.261$, a small fraction of the patterns has become unstable however instead of iterating to a high activity fixed point, as would happen for $b = 0$, a stable configuration is reached after a few spin flips. As γ increases more patterns become unstable and the fixed points reached have a greater range of activities. At $\gamma = 0.593, c = 0.447$ some trajectories reach high activity fixed points. The distribution of these high activity fixed points is fairly even until $\gamma = 0.693, c = 0.5$ where a peak appears. This high activity peak is illustrated for $\gamma = 0.735, c = 0.52$

The introduction of inhibition has allowed slightly noisy versions of the patterns to become fixed points at high loading. In this way the capacity of the network is increased compared with the case of no inhibitory interactions. The incorrect background ones that are present in these fixed points are the 1s common to the large number of patterns that have small overlaps with recalled pattern. Thus a form of generalisation, be it welcome or not, is taking place.

4.9 Discussion of Willshaw Rule

The Willshaw rule is a connection scheme that is rather contrary in that its formulation (4.41) is very simple, yet its performance is difficult to analyse exactly. It is only for the case of extremely biased patterns that Willshaw's original analysis appears accurate. However, within this regime of f a thorough understanding of the rule can be obtained. When the original Willshaw rule is used, so that Dale's Principle is obeyed, the memory attractors are the exact patterns. Increasing the storage level results in a higher connectivity within the network, and the patterns eventually become unstable to a high activity attractor. When the connections are modified through (4.72), so that connections are inhibitory before they are switched on and become excitatory, the high activity attractor is suppressed. The high activity attractor that occurs with this version of the Willshaw rule is, to an extent, within biological acceptability. This is because it is only appears for high connectivity levels that are not realised in biological neural networks. However there is a forfeit in biological reality because Dale's principle is violated. In order to avoid this forfeit, Golomb *et al* have argued that the inhibitory interactions can be considered an effective interaction arising from an auxiliary network of

inhibitory neurons. This idea will be pursued in the next chapter.

The introduction of inhibitory interactions into the model also changed the attractor structure of the memories. This is most clearly demonstrated by the simulations presented in Figures 4.11 and 4.12. When inhibitory interactions are introduced patterns that are not exactly stable may still have attractors associated with them. The simulations also illustrated that for the Willshaw rule in attractor neural networks γ_2 as opposed to γ_1 given by Willshaw's analysis, is the more relevant saturation storage.

For the two connection rules, Covariance and Willshaw, studied in this chapter the storage capacity for extremely biased patterns $f \ll 1$ is very high and in both cases has the form of the Gardner bound (4.2). However to obtain this large storage capacity the threshold must be carefully chosen. The threshold also determines the basins of attractions of the memory attractors. This is particularly apparent with the Willshaw rule, where the overlap has to be greater than the threshold for a noisy pattern to be retrieved.

To summarise, the main biological problem with biased pattern networks is obeying with Dale's principle whilst avoiding the existence of high activity attractors. In addition there is the problem of low rates which shall be discussed in the next chapter.

Chapter 5

A Biologically Acceptable Neural Network Model

5.1 Low and High Rates

In the previous chapters, features of attractors that exist in neural network models have been analysed and discussed. In the preceding chapter the task of bringing these attractors closer to biological reality was initiated by examining the spatial activity of the network configuration. In this chapter this task will be pursued more vigorously by considering the temporal activity or rates of the model neurons.

5.1.1 Biological Rates

In order to bring Neural Network models closer to biology biological experiments must be considered so that one knows the behaviour that the models are trying to reproduce. Many Neurophysiological studies are based on microelectrode recordings from cortical areas of animals trained to perform certain tasks. As an example we shall discuss the experiments of Miyashita and Chang[63].

In these experiments a monkey was positioned in front of a video monitor and the following trial sequence carried out:

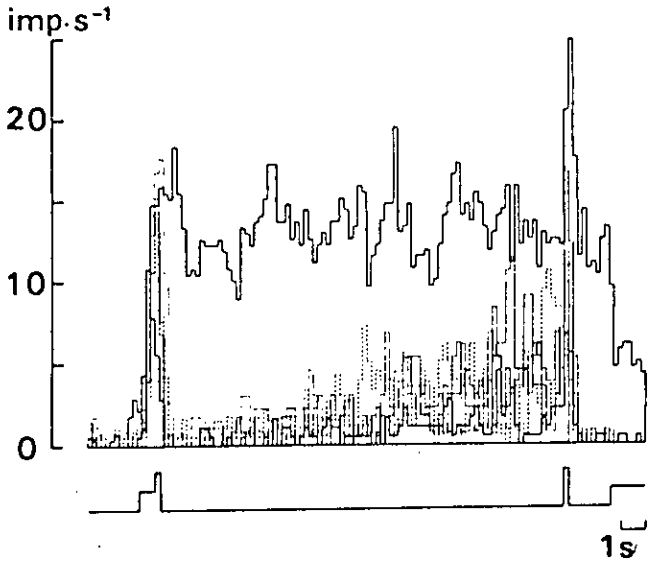

Figure 5.1: Activity histograms of a neuron upon presentation of 7 visual stimuli. One pattern provokes a high rate while the other 6 lead to the low rates. The stimuli is presented for $200ms$, as is indicated by the schedule marked under the graph. The elevated rate persists for 16 seconds. (Taken from Miyashita and Chang[63])

- The monkey depressed a lever.
- A green warning light appeared to signal that a stimulus was imminent.
- A stimulus pattern appeared on the screen for 0.2 seconds.
- A delay lasting 16 seconds occurred in which the video screen was blank.
- A second stimulus, that was either identical to, or different from, the first, appeared for 0.2 seconds.
- The monkey made a choice, by touching the video screen or not, as to whether the second stimulus was identical to or different from the first.

Before the trials were carried out the monkey had been trained to make the correct choices and any errors it made during the trials were discarded from the results. Throughout the trial the monkey had a microelectrode that recorded from the anterior ventral part of its temporal cortex. This microelectrode recorded the emission of spikes from a single neuron. Each monkey underwent a series of these trials with many different stimuli.

The experiments showed that when the monkey received the initial stimulus the firing rate of the recorded neurons, in most cases, became elevated (the firing rate

was greater than 10 spikes/s). Of the 144, out of a total of 188, neurons where this occurred, 77 showed a selective response in the delay period. This means that only for certain stimuli did the firing rate of the neuron remained elevated during the delay period. Most of these neurons only gave a response in the delay period to a small number of the stimuli. These stimuli to which the neurons showed a selective response differed from neuron to neuron. Figure 5.1 shows the results of 7 trials with different stimuli on the same neuron. The figure clearly shows that 1 stimulus out of the seven elicited a high firing rate from the neuron that continued during the delay period to be terminated by the second stimulus.

The significant résult of the experiment was that during the 16 second delay the neurons could sustain an elevated firing rate. The elevated firing rate was greater than 10 but less than 20 spikes per second. This is firm support for the concept of attractors where a high firing rate of a single neuron can be maintained by the cooperative action of an assembly of neurons. It is difficult to see how a single neuron could sustain an elevated firing rate independently, in the absence of a stimulus.

5.1.2 Rates and Glauber Dynamics

The problem in relating the attractors discussed in the previous chapters more precisely to the elevated rates noted in the experiments of Mayashita and Chang is in the interpretation of rate in the neural network model. If we wish the Glauber dynamics used for random sequential updating, to represent the dynamics of a system of real neurons, then the time taken for an updating sweep must be related to a biological time scale. In an updating sweep each spin is visited once on average once, thus only gets a chance to take the active value 1 once every time step. It seems natural to equate the time step with some refractory period of the neuron. However this raises a serious problem. The models so far discussed are designed so that when the network configuration is near to a stored pattern, the spins active in a pattern have positive local field. This was revealed most simply by the signal and noise analyses (see for example section 2.2). A spin that has a positive local field will take the active value 1 whenever it is updated. In the biological comparison this amounts to emitting a spike after every refractory period. A neuron active in the recall then spikes at the maximum possible rate, that is at several hundred

spikes per second. Although the effects of temperature and synaptic asymmetry may mean that the attractor is not a single fixed spin configuration, so that some spins may not be frozen, a spin with a local field that is, on average, positive will still spend more than half the time in the active state. This firing rate is still far too high for fruitful comparison with biology.

5.1.3 Neural Network Models that Produce Low Rates

In the context of ANN's there have been three models proposed that can produce firing rates of a satisfactory magnitude. These models all use a bath of inhibitory neurons to control the overall spatial activity. The patterns are stored in an excitatory network to which the inhibitory network is coupled. All three models basically work by having the inhibition strong enough that the spatial activity generated when all the neurons in a pattern fire simultaneously is too high to be stable. When the excitatory network is retrieving, at each time step only a fraction of the excitatory neurons in the pattern can emit spikes. If, at each time step, this fraction is a random subset of the the excitatory neurons in the pattern, then each neuron in the pattern will emit on average a spike every few time steps. In order for this idea to work, the local field of each neuron involved in recall must, on average, be below threshold so that the neuron will emit spikes less than half the time.

However all three models that have implemented this idea have serious drawbacks. The model proposed by Amit and Treves[64,65] has the inhibitory network working through an effective inhibition so that the activity in the inhibitory network reacts instantaneously to changes in the activity of the excitatory network. In addition the effective inhibition has a very specific pattern dependent structure that is necessary to to ensure that a single pattern, rather than a mixture of patterns, is recalled.

The model introduced by Rubin and Sompolinsky[66] used the Willshaw synaptic structure with an effective inhibition of the same type discussed in the previous chapter, section 4.8.1. However the low rate mechanism described above requires that the complete patterns are not stable so that a large value of the parameter b, discussed in section 4.8.1, must be employed. If the threshold were positive, as

was the case in the previous chapter, then the only attractor would be the all zeros configuration. In order to destabilise the all zeros attractor a negative threshold is introduced. Although this negative threshold is not immediately biologically acceptable it can be rationalised by construing it as uniform external stimulus to the network from other cortical regions, or *attention*. In the resulting recall mode the fields of the active neurons are below threshold. It is then noise, parameterised by temperature T that allows these active neurons to emit spikes every few time-steps.

However, two additional problems appeared with this model. Firstly, the model relies on the fact that when a pattern is being recalled, the Willshaw structure gives the same field to all 1-sites. When this homogeneity is relaxed, for example by introducing non-uniform thresholds, the spins tend to freeze and the low rates are lost. Secondly, when the inhibition was simulated in full by a separate inhibitory network, it was found that a recalled pattern could not be stabilised for times greater than 10 updating sweeps of the system or Monte Carlo Steps (MCS). This instability can be understood by realising that the activity of the excitatory population is not rigidly constrained. The modulation is by feedback from the inhibitory network. The delay in the feedback and the fact that the network is stochastic, due to the finite temperature, means that fluctuations in the activity of the excitatory network will occur. These fluctuations can instantaneously reduce the activity of the network to zero. The spatial activity in the network will regenerate itself but the network has no information as to which pattern was being recalled. Therefore the network will typically wander from pattern to pattern.

Buhmann[67] has studied a network similar to [66] but with several differences. Firstly, the thresholds are positive which is a more comfortable situation biologically. The inhibition is effected in a quadratic manner rather than the linear manner of [66] and equation (4.72). Thirdly, the connection scheme from the inhibitory to excitatory network is not fully connected. This network appeared to be the first model, with an explicit inhibitory network, to stabilise a recalled pattern, in which the active neurons fired with low rates, for extensive periods of time (greater than 30 M.C.S).

However a closer inspection of the simulations of [67] revealed that the crucial feature necessary to bring about this stabilisation was none of the features men-

tioned in the previous paragraph, but the specific implementation of the random sequential dynamics. In the dynamics used, the neuron to be updated was chosen randomly at each update. This means that, in an time step of N updates, the number of times each individual neuron is updated is given by a Poisson distribution with mean $N \times 1/N = 1$. The probability that a neuron is not updated is then $e^{-1} \simeq 0.38$. This implies that 0.38 of neurons, that emit a spike, will remain in the active state without being updated for more than 1 time-step. Moreover 0.14 of neurons that emit a spike will remain in the active state for more than two time steps. If in one time-step, a fluctuation in the activity occurs and no neurons emit spikes, then the network has some record of the neurons that spiked in the previous time steps. This memory allows the same pattern to be recalled after a temporary drop in activity. When the network was run with the updating of the neurons in a fixed sequence, or alternatively when in each time-step every neuron was updated but the order of updating was randomly chosen, the long term instability of the patterns returned.

From the critique of these works some important principles can be extracted:

- The observance of Dale's principle which separates the neural network into two sub-assemblies of excitatory and inhibitory neurons serves as useful framework in which the two sub-assemblies are functionally distinct. The excitatory neurons store and recall information whereas the inhibitory neurons monitor the activity of the network. Biological evidence[3] shows that only 15-20 % of cortical neurons are inhibitory which makes the less direct deployment of these neurons for information processing within the models appear more reasonable.

- Noise (parameterised by the temperature T) is an essential feature as it allows neurons depolarised below threshold to emit spikes. On the updating of such neurons the probability of emitting a spike is less than a half so that low rates may result.

- In order to stabilise the recall of patterns for significant lengths of time, there must be some short term memory of the spiking within the network.

5.2 Integrate and Fire Neurons

5.2.1 Neurons that Integrate over Time

The third principle extracted in the last section indicates that a there must be some memory of the spike history of the network. In the model of Buhmann[67] this was realised by the phenomenon of neurons remaining in the active state for more than 1 time-step. However this was an artifact of a particular dynamics. One way to introduce this memory into a neural network by more direct means is to allow the model neurons to integrate their inputs over time as well as space. The model neurons studied so far in this thesis have only integrated their inputs over space, by virtue of the summed weights rule (1.8).

If we return to biology it is well known that the cell membrane integrates its inputs over time as well as space. This is performed by the membrane having a time constant τ that is due to the capacitance of the cell membrane produced by its insulating properties. The changes in membrane potential due to incoming action potentials are then smeared out over time and the effects of non-simultaneous incoming action potentials can add up. This effect can be partially modelled by introducing the time constant τ through an exponential decay of the membrane potential to its resting level.

$$h_i(t+1) = h_i(t)\exp(-\frac{1}{\tau}) + \sum_j J_{ij}V_j(t) \tag{5.1}$$

$$= h_i(0)\exp\left(-\frac{t+1}{\tau}\right) + \sum_{t'=1}^{t}\sum_j J_{ij}V_j(t')\exp\left(-\frac{(t-t')}{\tau}\right) \tag{5.2}$$

In (5.2), h_i denotes the membrane depolarisation rather than the absolute membrane polarisation so that $h_i = 0$ indicates that the membrane potential is at the resting level. The term membrane potential will be used to refer to h from now on, to distinguish it from the local field (1.3). The initial condition in (5.2) is given at $t = 0$. In biology this corresponds to the post-synaptic potential at the end of the absolute refractory period due to the last spike emitted by the neuron. The values $h(0)$ may take will be discussed in section 5.3.2.

In (5.2) the sum over t' shows that such a model neuron is integrating its inputs over time as well as space. What equation (5.2) amounts is keeping the membrane

potential as a dynamical variable of each neuron, rather than calculating it only when a neuron is updated. This is the essence of an integrate-and-fire neuron. The study of such model neurons has developed separately from the study of the model neurons considered in the previous chapters[68] . Integrate and fire neurons are usually considered on the single neuron level as a unit that integrates a constant afferent stimulus which decays exponentially. When the potential reaches threshold the neuron emits a spike and the membrane potential is set to some negative (hyper-polarised) value. Continuous time is usually considered and the membrane potential dynamics written as a differential equation

$$\frac{dh_i(t)}{dt} = -\frac{h_i}{\tau} + I(t), \tag{5.3}$$

where $I(t)$ is the afferent current. If the afferent current is constant then the time for the neuron to reach the threshold, which, excluding the ARP, amounts to the period of the firing frequency, is given by [68]

$$\tau \ln \left[\frac{I\tau - h_0}{I\tau - \theta} \right]. \tag{5.4}$$

If $I\tau < \theta$, the potential does not reach threshold and the firing frequency is zero.

5.2.2 New Considerations Introduced by Integrate and Fire Dynamics

The basic aim of the present chapter is to incorporate integrate and fire neurons into a network with attractors and associative memory properties. In practice this amounts to a complete revision of network dynamics. The first step of this revision is to move towards continuous time so that the exponential decay of the membrane potentials can be implemented. With regards to computer simulation one still has to discretise the time so that a programme that carries out the updating may be written. When Glauber dynamics are used, the shortest time scale is one updating sweep which corresponds to a refractory period of approximately 5 ms. When time scales shorter than this are considered, one must also consider the biological mechanisms that operate at these shorter time scales. As the time scale becomes shorter and shorter then more biological complexity must be taken into account. For example the duration of an action potential is about 1 ms. If one considers a basic time scale much smaller than this then the effect of an action

potential must be spread over more than one unit of time. In order to avoid an unreasonable amount of complexity we will choose the basic time scale to be of order $1ms$. This time scale is additionally convenient as it is approximately the resolution at which biological experiments are carried out. This will then allow a direct comparison between the results of computer simulation and the results of biological experiment.

At this time scale there is one obvious biological feature that must be modelled. This is the time for an action potential emitted by the pre-synaptic neuron to reach the post-synaptic neuron. This delay time is a summation of the time for the action potential to propagate along the axon, the time for the synaptic transmitters to cross the synaptic cleft and the time for electrotonic propagation of the effect of the action potential up the dendritic tree. The modelling of the delay time will be explained in detail in section 5.3.6.

A feature of network modelling that must be reconsidered when the basic time scale is shortened is the implementation of noise. In Glauber dynamics the noise is implemented through the probabilities of the outcome of the update (4.5). The probabilities require a random number to be generated. For two consecutive updates of a neuron these random numbers are, of course, uncorrelated which means that the noise is uncorrelated at consecutive updates. As discussed in Chapter 1, noise may be interpreted in a variety of different ways. The main features it has been used to capture are variations in the properties of neurons over the neural population; variations in time of an individual neurons properties; and the variations in stimuli afferent upon on individual neuron from extra-network sources. The latter two features both involve fluctuations in time at an individual neuron. As the time scale becomes shorter these fluctuations must be correlated from one update of the neuron to the next update.

In the present chapter we will primarily consider noise as extra network afferents. These afferents directly affect the membrane potential of a neuron so that the most convenient way of implementing the noise and allowing correlations in time to develop is to write the equation for the membrane potential as

$$h(t+1) = h(t)\exp(-\frac{1}{\tau}) + I(t) + \xi. \tag{5.5}$$

Here I(t) represents the coherent afferents on the neuron whereas ξ is a random number which represents the uncorrelated extra-network afferents. This imple-

mentation of the noise should allow the mechanism of low rates to occur: the accumulation of noise components ξ can allow the membrane potential of a neuron left below threshold by the coherent input $I(t)$ to cross the threshold. Recall that the noiseless integrate and fire neuron would only spike if the stimulus was above a certain strength. This was demonstrated by equation (5.4).

In equation (5.5), $I(t)$ may have two components: the first is the feedback from action potentials within the network; the second is a coherent stimulus from outside the network. In the dynamics discussed in the previous chapters, recall of a stored pattern could be initiated by setting the spin configuration of the network to be a noisy version of a stored pattern and then iterating the dynamics. However in the dynamics presently being formulated it is the membrane potentials that are the important dynamical variables rather than the states of neurons. In order to stimulate a network towards recall of a stored pattern it is the membrane potentials that must be initialised. The most realistic way of doing this is to impose a stimulus component of $I(t)$ at a certain set of neurons for an initialisation period.

Another feature that must be reconsidered when one moves to shorter timescales is the refractory periods. With Glauber dynamics the refractory period did not have to be modelled because the basic unit of time was some refractory period. However we now have to implement a refractory period explicitly. In addition the distinct phenomena of absolute and relative refractory periods must be modelled separately. An absolute refractory period can be modelled by simply not updating a neuron for a short time after it has fired. This means that the neuron will not spike but may still receive afferents. In addition, if immediately after a neuron has spiked, the membrane potential is reset to a negative value, the neuron will emerge from the absolute refractory period hyper-polarised. The time taken for the neuron to reach its resting potential and depolarise to somewhere near the threshold value will then represent a relative refractory period.

5.3 The Model Network in Detail

5.3.1 The Parameters of the Network Model

The network is composed of N_e excitatory neurons and N_i inhibitory neurons. An acceptable biological figure of $N_e \sim 0.2N_i$ will be observed. Each neuron is described by two time dependent variables: its membrane potential, h_i^e (or h_i^i) and the time of the last spike emitted by it t_i^e (or t_i^i), where the superscript differentiates the excitatory from inhibitory neurons.

In addition, each neuron has several other parameters associated with it:

1. A threshold $\theta^{e,i}$
2. An exponential decay time constant, $\tau^{e,i}$, of the neuron's membrane potential.
3. The absolute refractory period (ARP) of duration ρ. Within this period, which immediately follows the emission of a spike a neuron cannot emit a second spike.
4. The value of the post-spike membrane potential, μ, which may effectively represent a relative refractory period.

The synaptic couplings between pairs of neurons fall into four categories: excitatory-excitatory, excitatory-inhibitory, inhibitory-excitatory and inhibitory-inhibitory. The connections within each of the last three categories are taken to be uniform, for example any two inhibitory-inhibitory connections take the same value J^{ii}/Nf. The excitatory-excitatory connections are assigned values according to the Willshaw connection scheme discussed in section 4.7 with a different normalisation. To be specific

$$J_{ij}^{ee} = \frac{J^{ee}}{N_e f} \quad \text{if in at least one pattern} \quad \eta_i = \eta_j = 1 \tag{5.6}$$

$$= 0 \qquad \text{otherwise.} \tag{5.7}$$

Altogether there are four coupling parameters J^{ei}, J^{ie}, J^{ii} and J^{ee},each of which are of approximate magnitude unity.

Each synapse also has a delay parameter, δ_{ij} associated with it. This delay is the time taken for a spike emitted by the pre-synaptic neuron j to reach the post-synaptic neuron i. Finally there are two global noise levels, T_e and T_i, for the excitatory and inhibitory networks respectively.

5.3.2 The Dynamics

The basic time scale is determined by the ARP which corresponds to about $2ms$. The ARP is of duration ρ time steps. These time steps are the resolution of the model. In general we will consider $\rho = 4$. Within each time step, each neuron is updated. The updating is basically parallel within each time step so that the order of updating is of no importance. An exception to this statement is when there are no delays in which case the updating should be random sequential. The updating procedure for a particular neuron is as follows.

1. The neuron's membrane potential is reset according to the equation

$$h_i^{e,i}(t) = h_i^{e,i}(t-1)\exp(-\frac{1}{\tau^{e,i}}) + I_i + \sum_j J_{ij}\delta(t - t_j - \delta_{ij}) + \xi, \qquad (5.8)$$

   where I_i is the stimulus (if any) and ξ is the incoming noise. The third term on the r.h.s. of Equation (5.8) represents the integration of action potentials, from within both the excitatory and inhibitory sub-networks, that arrive at neuron i at time t.

2. The last time the neuron spiked is examined to see if the neuron is in an ARP. This is so if $t \leq t_i + \rho$. If this is the case the neuron is refractory and cannot spike so the updating is complete.

3. If the neuron is not refractory and if $h_i^{e,i}(t) \geq \theta_i^{e,i}$ then the neuron spikes at time t. In this case t_i is set to t and $h_i^{e,i}$ is reset to μ. If $\mu < 0$ the neuron will eventually emerge from the refractory period with a hyper-polarisation of around $\mu\exp(-\rho/\tau)$, which will lead to an effective refractory period as discussed in section 5.2.2.

5.3.3 Initial Stimulus and Associative Recall

A stimulus, at a neuron i, is represented by an afferent current that persists, for an initialisation period $\mathcal{T}_S$ and generates an excitatory post synaptic potential of magnitude I_i per time step. The associative memory properties of the network are determined by how such stimuli are distributed across the combined network and how the network responds. The distribution of the stimuli is described by:

x_p the fraction of neurons in the pattern receiving I_i

x_b the fraction of neurons in the background receiving I_i

x_I the fraction of inhibitory neurons receiving I_i.

The first two groups are both excitatory. The first group is composed of neurons which have $\eta_i^1 = 1$, whereas the second group have $\eta_i^1 = 0$ (the pattern being recalled is taken to be pattern 1). The inhibitory neurons are not included in the pattern structure but may still be subject to stimuli.

5.3.4 The Noise Distribution

So far the exact distribution of the noise ξ has not been specified. When the noise is interpreted as random afferents from extra-network sources then the distribution will have a finite positive mean. This because it is excitatory neurons that make the long range connections within the cortex. So in general the afferents from extra-cortical sources will be excitatory. The mean of the noise must vary as the size of the time step $1/\rho$, is altered. To show this explicitly we write the mean as

$$\overline{\xi} = \frac{T}{\rho}. \tag{5.9}$$

Equation (5.9) guarantees that whatever value ρ takes, the mean noisy contribution to the membrane potential in one refractory period will be T (excluding the effect of decay of membrane potential). T then becomes a measure of the extra-network activity. With regard to simulation the easiest distribution to use is a Gaussian. A Gaussian distribution can be justified if the mean number of random inputs to a neuron in each time step is large. This should be the case if ρ is not too large so that the discretisation of time is not too fine. One still

needs to specify the variance of the Gaussian distribution. The variance represents the stochastic part of the extra-network afferents for if zero variance was chosen, each network neuron would just be receiving a constant stimulus of T/ρ every time-step. Although this variance can be chosen arbitrarily, we choose to make a correspondence with the noise implementations of the previous chapters. In these chapters a neuron was updated once every refractory period, so that a random number to calculate the probability of firing was generated once every refractory period. We can make a correspondence between the random number generated for Glauber dynamics and a noisy afferent to the neuron. Although strictly a tanh function should be used for Glauber dynamics, this is qualitatively the same as the use of a Gauss error function for which the noisy afferent has a Gaussian distribution with variance T^2, where T is the temperature parameter of Glauber dynamics. The noisy afferent could have arrived at any time during the refractory period. In the present dynamics we update the neuron ρ times in a refractory period so that one can approximate Glauber dynamics in the present framework as a binomial process for the arrival of a Gaussian random variable. Mathematically,

$$\begin{aligned} \xi = 0 \quad &\text{with probability} \quad 1 - \frac{1}{\rho} \\ \xi = z \quad &\text{with probability} \quad \frac{1}{\rho}\frac{1}{\sqrt{2\pi T^2}} \exp - \left(\frac{z^2}{2T^2}\right) \end{aligned} \tag{5.10}$$

However the point of formulating the noise as a stochastic contribution to the membrane potential was to introduce the possibility of correlations within the noise. The stochastic process realised by (5.10) will generate a stochastic contribution to the membrane potential on average once every ARP. This allows time for the different stochastic contributions to decay and so suppress the build up of correlations. In order to avoid this effect we must have a stochastic contribution at each update. In order to do this and still have some correspondence with the temperature of Glauber dynamics we can use a Gaussian distribution for ξ that has the same variance, σ^2, as (5.10):

$$\sigma^2 = T^2\frac{1}{\rho}, \tag{5.11}$$

and a mean given by T/ρ.

5.3.5 Shunting and Noise

In chapter 4 the inhibition was introduced in a linear manner so that the effective inhibition term used in (4.8.1), $\frac{a}{Nf}\sum_i V_i$, was simply subtracted from the fields at the sites, in the same way that excitatory contributions to the fields were added. This linear effect of inhibition is hyper-polarisation. However in reality inhibition occurs through shunting as well as hyper-polarisation. Shunting basically decreases the depolarisation caused by excitatory afferents. This is because inhibitory afferents increase the overall conductivity of the membrane. One can visualise a simple circuit to understand that an excitatory current will cause a smaller voltage drop across the shunted membrane because the resistance is lower. Inhibition in cortical neurons is a mixture of increased conductivity for K^+ and Cl^- ions. Increased K^+ conductivity has both hyper-polarising and shunting effects whereas increased Cl^- conductivity is mostly shunting. In this chapter the aim is to present simulations which can be compared with biological experiments. In order to do this shunting must be modelled.

In this chapter shunting is modelled by expressing the membrane potential h as

$$h_i = \frac{U_i^{\text{ex}} - (1-s)U_i^{\text{in}}}{1 + \frac{sU_i^{\text{in}}}{\theta}}. \tag{5.12}$$

where U^{ex} is the temporal summation (linear summation with exponential decay) of excitatory afferents and U^{in} is the temporal summation of the inhibitory afferents. The inhibitory term with coefficient $1-s$ represents the hyper-polarising effect of the inhibition whereas the term in the denominator with coefficient s/θ represents the shunting effect. The reason for the choice of these coefficients will become apparent in the next paragraph. This manner of modelling inhibition is a simplification because it does not distinguish an inhibitory afferent that increases Cl^- conduction, and so results primarily in shunting, from one that causes a high degree of hyper-polarisation. Instead all inhibitory afferents cause the same ratio of hyper-polarisation to shunting. This ratio is determined by the parameter s. As s is varied from 0 to 1 one goes from a fully hyper-polarising implementation of inhibition to a fully shunting implementation.

When the noise in the network is implemented as random afferents the value of s does not alter the the time at which a neuron's membrane potential crosses

threshold, although the actual value of the membrane potential at a general time may be different. In order to see this it is convenient to consider the noiseless case first. The condition for the emission of a spike is

$$h_i = \frac{U_i^{\text{ex}} - (1-s)U_i^{\text{in}}}{1 + \frac{sU_i^{\text{in}}}{\theta}} > \theta, \tag{5.13}$$

which yields, after multiplying out the denominator

$$U_i^{\text{ex}} - U_i^{\text{in}} > \theta. \tag{5.14}$$

This condition is independent of s and is equivalent to the condition obtained for a fully hyper-polarising implementation of shunting. In the presence of noise the condition for the emission of a spike becomes

$$\frac{U_i^{\text{ex}} + \xi - (1-s)U_i^{\text{in}}}{1 + \frac{sU_i^{\text{in}}}{\theta}} > \theta. \tag{5.15}$$

The noise term ξ is placed alongside U_i^{ex} because we are viewing noise as an excitatory afferent. What this implies is that the noisy afferents are shunted as well. One then finds

$$U_i^{\text{ex}} - U_i^{\text{in}} + \xi > \theta \tag{5.16}$$

which again is independent of s and equivalent to fully hyper-polarising inhibition.

Practically one may take advantage of the fact that running the network with or without shunting is mathematically equivalent, to produce several representations of the membrane potential from the same run. It is most convenient to initially run a simulation with the inhibition fully hyper-polarising (to which the quoted synaptic efficacies in the figures refer) and then vary the amount of shunting to a biologically realistic level. The desired level may be determined by taking into account the current biological knowledge as to how much shunting occurs, or more pragmatically by comparing the appearance of the representation to experimental results. The fully hyper-polarising representation of the membrane potential may be described as "effective" because the amount the potential is below threshold determines the the amount of excitatory afferents the neuron must receive to emit a spike. In contrast representation of the membrane potential that involves some shunting may be described as "biological".

5.3.6 Delays

Synaptic transmission delays may have several origins. They may reflect the length of the pre-synaptic axonal distance; the time taken for transmitters to cross the synaptic cleft; or the post-synaptic dendritic distance. We model delays by the parameter δ_{ij} associated with each synaptic connection. In the updating scheme described in section 5.3.2 a spike emitted by the pre-synaptic neuron j at time t affects the membrane potential of the post-synaptic neuron i at time $t + \delta_{ij}$. The δ_{ij}s are quenched random variables that are selected according to some distribution $Pr(\delta_{ij})$. The true nature of this distribution is not known. One could make relatively complex choices for the distribution: for example one could consider different distributions for excitatory and inhibitory neurons which would reflect the tendency for the inhibitory neurons' dendritic trees to be more localised. However one would hope that the details of the distribution do not qualitatively affect the performance of the network. We will consider two contrasting but relatively simple distributions which shall be referred to as *uniform* and *random*.

For uniform delays δ_{ij} is independent of i and j:

$$\delta_{ij} = D^{e,i} \tag{5.17}$$

where $D^{e,i}$ is the fixed delay length, which may assume different values according to whether the pre-synaptic neuron is excitatory or inhibitory.

For random delays we choose a maximal delay ($D^{\max}$) and a minimal delay ($D^{\min}$) and then set each δ_{ij} to a value inclusively between $D^{\max}$ and $D^{\min}$ with equal probability:

$$\begin{aligned} Pr(\delta_{ij}) &= \frac{1}{D^{\max} - D^{\min} + 1} \quad \text{if} \quad D^{\min} \leq \delta_{ij} \leq D^{\max} \\ &= 0 \qquad\qquad\qquad\qquad \text{otherwise} \end{aligned} \tag{5.18}$$

In general $D^{\min}$ will be taken to be 1.

The introduction of delays into the network forms a test of robustness. The network is most robust when the effect of spikes on membrane potentials is instantaneous. This is because fluctuations in activity within the network are more quickly damped out. If associative memory can be found in the presence of delays then one would be optimistic that associative memory would endure in the presence of other disruptive influences.

Our specific choice of distributions is motivated by a desire to keep the model simple, which in turn facilitates simulation, hence the use of uniform delays. Random delays are used to compare with uniform delays and check whether the simplicity of the latter has not obscured any important phenomena. Random delays are also a partial realisation of structural inhomogeneities. The introduction of such inhomogeneities into the model is important because the connection strengths are uniform for each of the four types of synapses (see Sec 5.3.1). We will show that in many parameter regimes the two delay distributions give qualitatively similar results, demonstrating the insensitivity of the basic substrate to model details, though random delays provide a more robust network and higher storage capacity.

5.4 Computer Experiments and Graphic Displays

We shall investigate the model described in the previous sections solely by computer simulations. There are two reasons for this. Firstly, as it stands the model is too complex to perform an analysis on and in addition the Willshaw connection rule used is only amenable to analysis in the regime of extreme bias of patterns $f \sim \ln(N)/N$ as discussed in chapter 4. Secondly, the reason the model is complex is because it has been designed to be reasonably faithful to biological reality and this has necessitated the introduction of a large number of new parameters and a revision of the dynamics. In order to test whether this complexity has served its purpose one would like to see the manner in which the model functions rather than obtain theoretical bounds on, for example, storage capacity. Of course, the underlying motivation is to produce results that, whilst resembling those of biological experiments, also manifest the fundamental principles of associative memory, thus vindicating the more abstract studies of the last thirty years. For these purposes a computer simulation is ideal. I shall present fairly sophisticated graphic displays produced by simulation runs that may be compared directly with the results of cortical experiments. The displays are of three types that shall now be discussed in turn, referring to Fig. 5.2

5.4.1 Membrane Potentials

These are displayed in the three top windows in the figure, each is equivalent to an intra-cellular electrode measuring the time course of the membrane depolarisation of a particular neuron. Time runs from left to right. One probe neuron is selected from each of the three classes (from top): foreground (excitatory neurons in the pattern being recalled); inhibitory neurons; background (excitatory neurons not in the pattern being recalled). The resting potential and threshold depolarisation levels are marked by horizontal lines across each window, with the threshold drawn above the resting potential. The potential plotted may be the "effective" membrane potential in which all the inhibition acts as hyper-polarising current, or a "biological" membrane potential in which some some fraction of the inhibition has been converted into shunting (see section 5.3.5). In both cases, whenever the potential reaches the threshold a spike is emitted. This can be checked by comparing the times at which the threshold is crossed with the corresponding spike raster.

5.4.2 Spike Rasters

Three spike rasters form the central window of each display. In each raster there are seven rows of dots. Each row of dots represents the spike emission times of a particular neuron. Each raster represents seven randomly chosen neurons from a particular class: inhibitory (top raster); background (middle raster) and foreground (bottom raster). The raster representing background neurons uses large dots, for ease of viewing, because the emission of a spike from one of these neurons is a less frequent occurrence than the emission of a spikes from a neuron in one of the other two classes. Time runs from right to left. The time at which the initial stimulus is turned is marked by a vertical line. The small horizontal bar at the top of the raster window sets the time scale and is 2 ARP in length.

This type of plot is the most useful for comparison with real cortical recordings where spike rasters are often presented (see for example [47]). It also gives a useful visual impression of effects such as "freezing" (when some neurons spike with high rates while others are quiescent). Freezing has a tendency to occur amongst the background neurons due to the use of the Willshaw matrix that leaves some

background neurons highly interconnected with neurons in the pattern whilst other background neurons are disjoint.

5.4.3 Average Spike Rates and the Edwards-Anderson Parameter

The three bottom windows display a quantitative picture of the spike activity within the three classes of neurons during a run. The quantity plotted is a running *time* average of the *population* average of the spike rate over each of the classes of neurons. It is calculated by counting the total number of spikes that have been emitted by neurons within the class in a time bin of fixed length B ARPs. This number is then divided by the number of neurons within the class and by B to give $\nu^{p,i,b}(t)$ the short time average of the number of spikes per neuron per ARP. The superscript corresponds to the class of neurons being considered so that

$$\nu^{p}(t) = \frac{1}{BN^{e}f} \sum_{t'=t-B\rho}^{t} \sum_{i} \eta_i V_i^{e}(t') \tag{5.19}$$

$$\nu^{b}(t) = \frac{1}{BN^{e}(1-f)} \sum_{t'=t-B\rho}^{t} \sum_{i} (1-\eta_i) V_i^{e}(t') \tag{5.20}$$

$$\nu^{i}(t) = \frac{1}{BN^{i}} \sum_{t'=t-B\rho}^{t} \sum_{i} V_i^{i}(t') \tag{5.21}$$

where $V_i(t) = 1$ if neuron i spiked at time t and is 0 otherwise. However because the background spike rate is very low $(1-f)\nu^b/f$ is displayed in the window. This implies that when the displays of ν^b and ν^p are of the same height it indicates that equal absolute numbers of foreground and background neurons have spiked in the bin period.

To quantify the activity within the different classes over the length of an individual run one calculates the average activity over time and population of a neuron within a class. This simply amounts to counting the total number of spikes emitted from neurons within a class, dividing by the duration of the run in ARPs and the number of neurons in the class to give $\langle \nu^{p,i,b} \rangle_t$ the average number of spikes per ARP. Retrieval then manifests itself through $\langle \nu^{p} \rangle_t$ being significantly higher than $\langle \nu^{b} \rangle_t$. The value of $\langle \nu^{i} \rangle_t$ is not directly relevant to retrieval but we shall see that

$\langle \nu^{p,\cdot,i} \rangle_t \simeq \langle \nu^{\cdot,i} \rangle_t$. This is acceptable because the inhibitory neurons then do not have an unreasonably high spike rate.

In order to quantify the distribution of rates over a class of neurons we shall use the Edwards-Anderson[34] order parameter. This parameter is in fact the mean square of the rates over the class:

$$q^{p,b,i} = \frac{1}{N^{p,b,i}} \sum_i (\langle \nu_i^{p,b,i} \rangle_t)^2. \tag{5.22}$$

The variance of the rate distribution over a class is then

$$\sigma^2 = q^{p,b,i} - (\langle \nu^{p,b,i} \rangle_t)^2. \tag{5.23}$$

When this variance is zero, then each neuron in the class has spiked with the same average frequency. The opposite extreme is when the neurons are frozen so that some do not spike at all ($\langle \nu_i^{p,b,i} \rangle_t = 0$) and some spike at maximal rate ($\langle \nu_i^{p,b,i} \rangle_t = 1$). This would give $q^{p,b,i}$ its maximal value $q^{p,b,i} = \langle \nu^{p,b,i} \rangle_t$.

5.5 Results from the Computer Simulations

In order to focus the simulations on specific points we shall present graphic displays that illustrate:

1. The manner in which associative memory and retrieval manifests itself.
2. The effect of varying the amount of thermal noise.
3. How the simulations with the different delay distributions compare with each other.

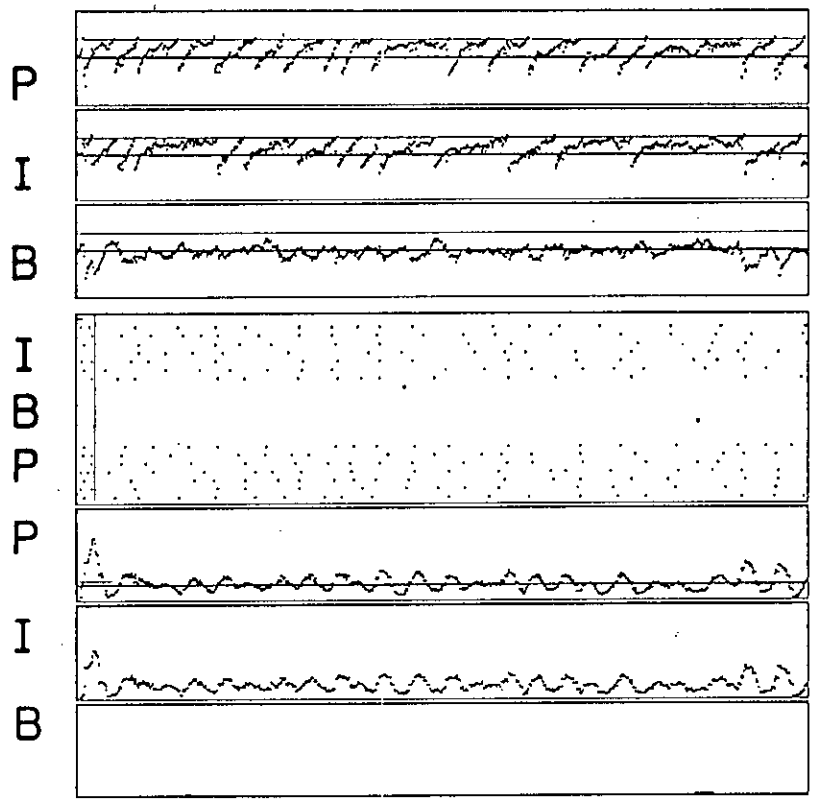

Figure 5.2: see p.144 for caption

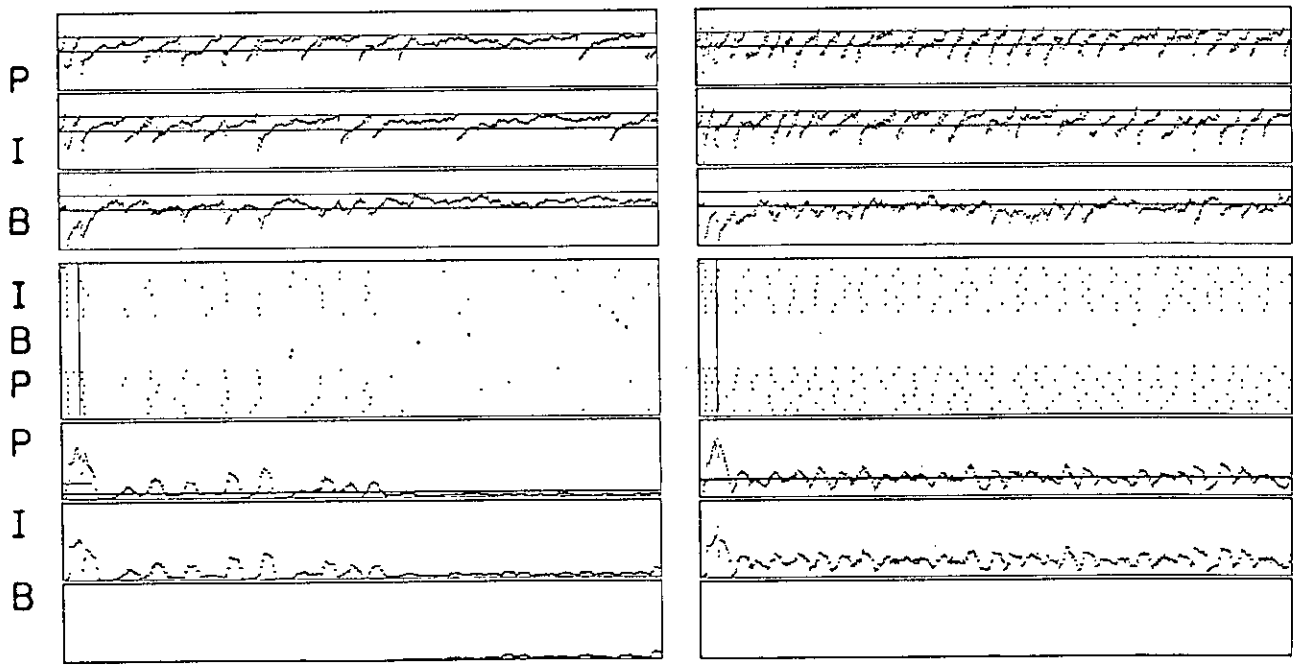

Figure 5.3: see p.146 for caption

Figure 5.2: Retrieval illustrated by graphic display produced during a run of the network. The top three windows are the time course of the membrane potential from three selected neurons. From top: pattern, inhibitory, background. These neurons' spike times are recorded in the bottom row of the corresponding spike raster. The shunting representation used fully hyper-polarising ($s = 0$). The horizontal lines mark the resting potential (lower line) and threshold.
The large middle window contains the spike rasters: the top 7 rows (small dots) contain the spike times of 7 inhibitory neurons; the middle rows (big dots) contain the spike times of 7 background neurons; the bottom rows (small dots) contain the spike times of 7 pattern neurons. The vertical line marks the end of the stimulus period. The short horizontal line at the top left of the window sets the time scale and corresponds to 2 ARPs.
The bottom windows display the time course of the average activity rates. From top: neurons in a pattern, inhibitory neurons, background neurons. The thin horizontal bar marks the position of an activity level of 10 % of the maximal rate. The thicker horizontal line marks the average level of the activity within the current run.
All windows share a common time axis.
Parameters:
Neurons: $N^e = 1000, N^i = 200$. Patterns: $P = 3, f = 0.2$. Synaptic strengths: $J^{ee} = 0.87$, $J^{ie} = 0.87$, $J^{ei} = 0.9$, $J^{ii} = 0.9$. ARP: $\rho = 4$. Delays: uniform $D^e = 3, D^i = 2$. Thresholds: $\theta^e = \theta^i = 0.25$. Noise: $T_e = 0.05, T_i = 0.045$. Post spike hyperpolarisation: $\mu = -0.2$ Decay constants: $\tau^e = \tau^i = 4$. Stimulus: $x_p = 0.4$, $x_b = 0.0$, $x_i = 0.1$; amplitude $I = 0.04$; duration 5 ARP. Duration of run = 200 ARP $\approx 400ms$.
Results:
Foreground $\langle \nu^p \rangle_t = 0.086$, $q^p = 0.008$; background $\langle \nu^b \rangle_t \approx 10^{-4}$, $q^b \approx 10^{-6}$; inhibitory $\langle \nu^i \rangle_t = 0.075$, $q^i = 0.006$.

5.5.1 Associative Retrieval

Associative retrieval manifests itself as an elevated firing rate in the pattern neurons that survives for an extended period of time after the initial stimulus has ceased. The elevated firing rate should extend to those pattern neurons that did not receive an initial stimulus. The firing rate is elevated compared to the firing

rate of the background neurons. The firing rate of the inhibitory neurons is not directly relevant to associative retrieval. However it is important that they still exhibit low rates. This is assured by demanding that the firing rate of the inhibitory neurons is not higher than that of the pattern neurons. This assumes that the pattern neurons have low (but elevated) rates. There is no sharp criterion for what is a biologically acceptable low rate. Nevertheless we can arbitrarily choose a rate of 50 spikes per second as the threshold for acceptability. This translates, with an ARP of 2 ms, into $\nu = 0.1$. The simulation in Fig. 5.2 fulfills these requirements and is thus an example of associative retrieval. One may note that the rates die down to the steady level of $\nu = 0.1$ (for the foreground neurons) after a transient burst of actvity driven by the initial stimulus. The is reminiscent of the data from real biological experiment shown in Fig. 5.1.

In Fig. 5.2 it can be seen that for the foreground and inhibitory neurons q is very near $(\langle\nu\rangle_t)^2$ so that the rates are evenly distributed within these two classes. However for the background neurons this is not true and $q^b \gg \langle\nu^b\rangle_t$. This implies that the some background neurons are spiking with higher frequency than others. The large value of the EA parameter for the background neurons may well be particular to the Willshaw connection scheme used for the J^{ee} connections. At any rate its source is easily understandable: the background neurons with the higher rates are those connected to a large proportion of the neurons in the pattern being recalled. In Fig.5.2 two of the background neurons emit spikes, but one can see that their rates are very much less than those of the foreground neurons. The more serious implication of the high EA value for background neurons is that many background neurons are spiking with very low rates or not all. This can be seen in the spike raster of Fig. 5.2 where five out of the seven background neurons probed do not emit spikes. The average background rate $\langle\nu^b\rangle_t$ converts into something less than 0.5 spikes per second. Instead one would like to see all background neurons spiking with the biological spontaneous firing rate of around 2 spikes per second.

Apart from the reservation about the distribution of background firing rates, Fig: 5.2 displays convincingly the the manifestation of associative memory that the model was developed with a view towards. This type of retrieval occurs with a wide range of parameters as long as the level of rates demanded is not too low. The level of rates is controlled by the relative strength of the inhibitory synapses,

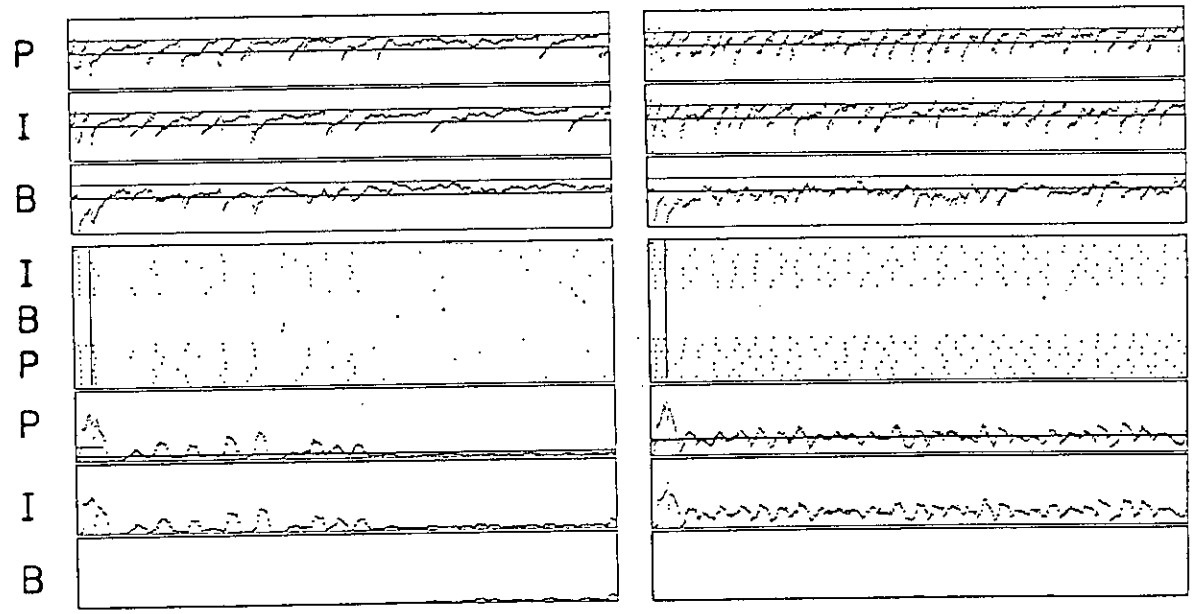

Figure 5.3: Two displays illustrating the temperature dependence of retrieval. Parameters are as for Fig. 5.2 except for temperatures.
left hand display $T_e = 0.04, T_i = 0.036$; right hand display $T_e = 0.06, T_i = 0.054$. Results: left hand display: Foreground $\langle \nu^p \rangle_t = 0.038, q^p = 0.001$; background $\langle \nu^b \rangle_t = 0.002, q^b \approx 0$; inhibitory $\langle \nu^i \rangle_t = 0.037, q^i = 0.001$.
right hand display: Foreground $\langle \nu^p \rangle_t = 0.115, q^p = 0.013$; background $\langle \nu^b \rangle_t \approx 0, q^b \approx 0$; inhibitory $\langle \nu^i \rangle_t = 0.105, q^i = 0.011$.

J^{ii} to J^{ie}, and more directly by the temperature.

5.5.2 The Dependence on Temperature

In general it was found that to achieve associative recall the temperature of the inhibitory network needed to be lower than that of the excitatory network. A qualitative explanation for this is that if, in the recall process, no neurons have fired for a short period of time so that the intra-network contributions to the membrane potentials have decayed below their equilibrium levels, then the excitatory neurons will have spikes generated through noise at a higher rate then the inhibitory neurons. This excitation will cause the activity in the network to regenerate itself and recall to continue.

Figure 5.3 displays two runs with identical parameters to figure 5.2 except for the temperatures. In the left hand frame, which is at low temperatures, the pattern remains stable for only a short time after the initial stimulus. The rates quickly drop to a very low level and the recall of the pattern is lost. After the recall is lost, the activity does not regenerate itself to the level it attained after the initial

stimulus. In the second frame which is at higher temperatures than Fig 5.2, the recall of the pattern continues after the initial stimulus has been turned off. The display is similar to Fig 5.2 except that the rates are higher. Figure 5.3 illustrates how the rates in recall increase with temperature and that the temperature must be must be above a certain level for recall to occur reliably. Of course, at the extreme of very high temperature one would expect recall to breakdown aswell. The left hand side frame of the figure is of additional interest as it suggests a mechanism through which the network could be removed from an attractor once a neural computation was complete. Starting from a recall mode similar to that of Fig. 5.2, the temperature could be lowered and the activity in the network will simply die down, as in the left hand frame of Fig. 5.3. The interpretation of temperature as extra-network afferents can then be construed further as attention.

5.5.3 Oscillatory Behaviour with Uniform and Random Delays

The difference between the behaviours of nets with uniform or random delays is best exposed through an unexpected but interesting phenomenon that appeared during the simulations. This is oscillatory behaviour in the rates.

An example of this oscillatory behaviour is shown in the left hand frame of Fig. 5.4. One can identify the oscillations visually in all three types of graphic display but most easily in the average rates window. The oscillation is not perfect, which may be explained by the stochasticity in the model, and the average period is about 8 ARP. These oscillations are very reminiscent of those first reported by Gray and Singer[69,70]. In order to make a more definite connection between the oscillations seen in the present simulations and these cortical recordings, more detailed experiments are required to define the true nature of the biological oscillations, and a systematic analysis of the simulation data, in the manner of [71], would need to be carried out.

In general there are two simple ways to produce oscillations in the simulations with uniform delays:

1. Increase the strength of the excitatory efficacies relative to the inhibitory

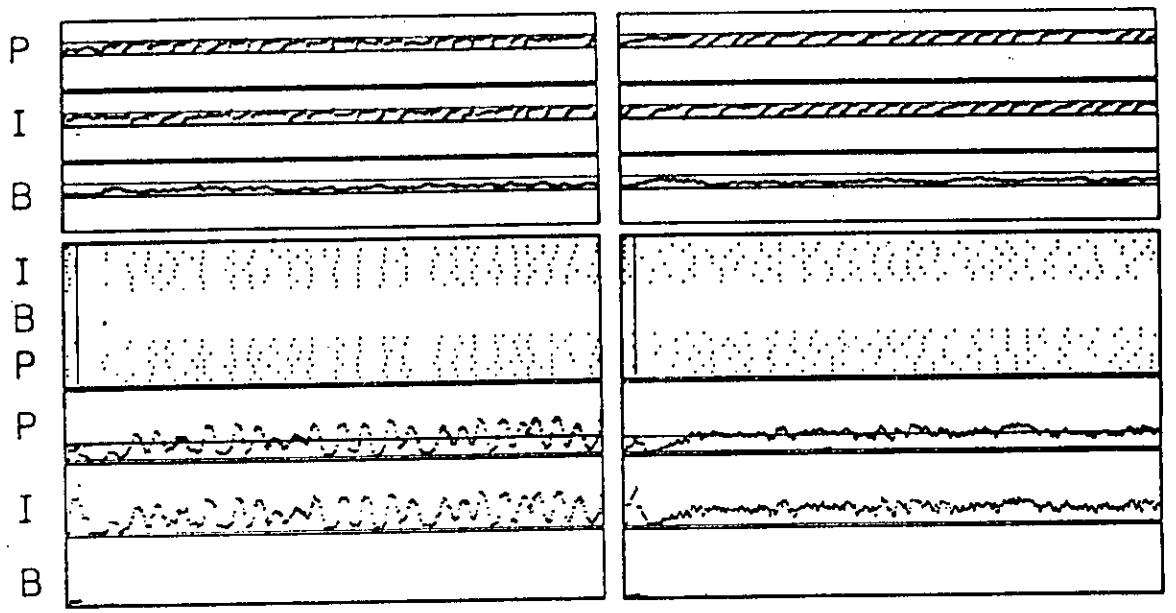

Figure 5.4: Displays illustrating oscillatory behaviour with uniform delays (left hand frame) and the absence of such behaviour with random delays (right hand frame). Parameters are as in Fig. 5.2 except that the excitatory synaptic strengths have been increased to $J^{ee} = J^{ei} = 1.1$.

ones (e.g. Fig. 5.4).

2. Keep a small stimulus present throughout the run.

The latter method is perhaps more relevant to the Gray-Singer oscillations. The first method is illustrated in Fig. 5.4. where the parameters are identical to Fig. 5.3 except that the strength of the excitatory synaptic efficacies has been increased. In the right hand display of the figure the parameters are identical to the left hand side except that random delays have been used. The mean delay times are the same in both frames. In the presence of random delays the oscillatory behaviour has been lost, although retrieval is still good. Oscillatory behaviour is attainable with random delays but one must search harder in the parameter space, than for uniform delays, to find it. One may say that although oscillatory behaviour is not an artifact of uniform delays, its presence is enhanced by them.

In order to compare the robustness of uniform and random delays we resort to testing the storage capacity. In Fig 5.5 the left hand display shows a network with uniform delays that has been overloaded by the storage of too many patterns. The network is identical to that of Fig. 5.3 except that the number of patterns stored has been increased from 3 to 7. The difference between the two displays is an

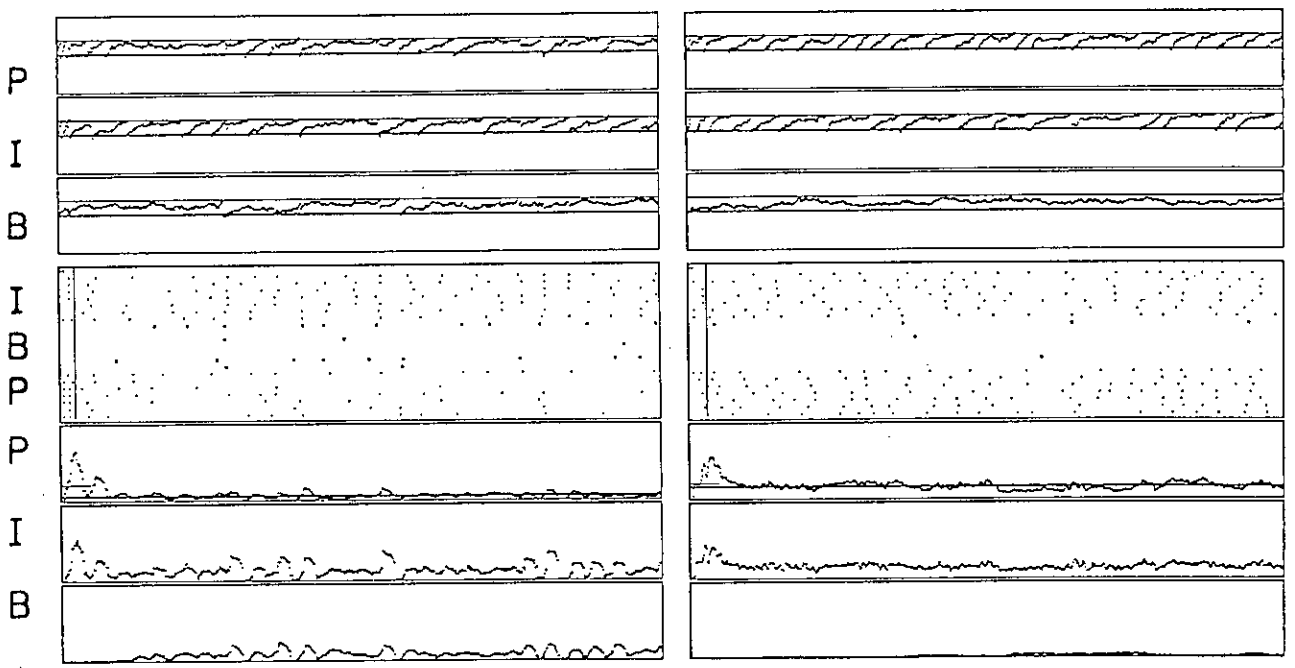

Figure 5.5: Displays illustrating breakdown of retrieval at high storage for uniform delays (left display) but retention of retrieval for random delays (right display). Parameters are the same as Fig: 5.2 except for $P = 7$ and $J^{ee} = J^{ei} = 0.9$. The representation of the membrane potential uses $s = 0.7$. Right hand display (uniform delays): Foreground $\langle \nu^p \rangle_t = 0.032, q^p = 0.001$; background $\langle \nu^b \rangle_t = 0.011, q^b = 0.0$; inhibitory $\langle \nu^i \rangle_t = 0.069, q^i = 0.005$; left hand display (random delays): Foreground $\langle \nu^p \rangle_t = 0.083, q^p = 0.007$; background $\langle \nu^b \rangle_t = 0.002, q^b \approx 0$; inhibitory $\langle \nu^i \rangle_t = 0.079, q^i = 0.006$.

increase in background activity, in Fig. 5.3 over Fig. 5.5, after the stimulus has been switched off. This results in decreased foreground activity in Fig. 5.5 which leads to the foreground neurons no longer having a significantly higher rate than those of the background. This sequence of pattern destabilisation is of a different appearance and origin from that shown on the left hand of Fig. 5.3, but is to be expected from a Willshaw connection scheme. At too high a loading some background neurons will become almost fully connected to the foreground neurons. These background neurons will then fire at nearly same rate as the foreground neurons and the increased spatial activity will cause other background neurons to fire. The inhibitory network will suppress the overall activity in after a slight delay, by which time the recall of the initial pattern has been lost.

In the Willshaw net with inhibition discussed in chapter 4, recall could still take place when some background neurons became active. This was because the inhibition acted instantaneously. When we use random delays then effects of fluctuations in activity within the pattern being recalled or inhibitory network are transmitted in a manner more smeared over time than with uniform delays. This is because some delays are shorter than average and some are longer. This should allow the network to react in a quicker but less forceful manner to fluctuations in activity. The network may then be able to suppress the activity of those background neurons highly connected to the foreground neurons, without endangering the recall of the pattern. That this is indeed the case is illustrated in the right hand frame of Fig.5.5. Here the parameters are identical to the left hand frame, except that random delays, with the same mean, are used. One can see that recall still takes place.

The storage levels illustrated in the previous paragraph do not at first glance seem very impressive. However one must remember that a Willshaw connection scheme is being used for values of f far from the optimal regime of $f \sim \ln(N)/N$ for the Willshaw network. In Fig. 5.6 the value of f is reduced to 0.1 and one sees that when 25 patterns are stored, recall can still occur when random delays are employed.

Figure 5.7 makes a comparison between two different membrane potential representations of the same run. In the upper windows s, defined in (5.12), is 0.7 so that a large proportion of the inhibition acts through shunting. This is a "bio-

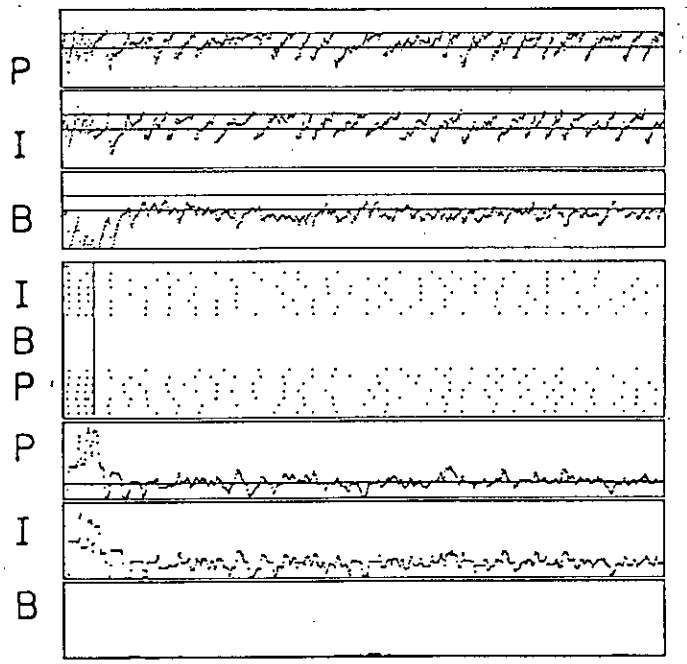

Figure 5.6: Display illustrating the enhanced storage capacity of a network with random delays and lower spatial coding f. The parameters are the same as Fig: 5.2 except for $P = 25, f = 0.1$, $N^i = 100$, $J^{ee} = J^{ei} = J^{ie} = J^{ii} = 1.1$, $T^i = 0.04$, and random delays are used.

Results:

Foreground $\langle \nu^p \rangle_t = 0.103, q^p = 0.011$; inhibitory $\langle \nu^i \rangle_t = 0.094, q^i = 0.009$; background $\langle \nu^b \rangle_t \approx 0$

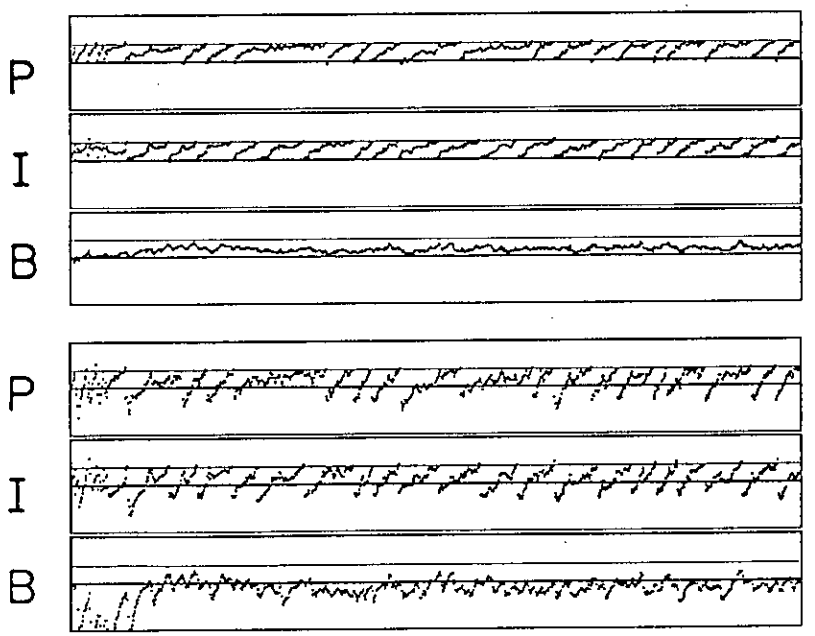

Figure 5.7: Displays illustrating different shunting representations of same run. The run is the same as that shown in Fig: 5.6 The top three windows show the membrane potential when $s = 0.7$, whereas the bottom three membrane potential windows show fully hyper-polarising inhibition (s=0). One should note that the spike times are identical.

logical" representatión of the membrane potential. At the neurons displayed the membrane potential is seldom less than zero, so that overall hyper-polarisation of the cell membrane is rare. The times when overall hyper-polarisation does occur is immediately after a spike. In contrast the lower displays show the same three neurons' membrane potential when $s = 0$, so that the inhibition is fully hyper-polarising and the "effective" membrane potential is shown . Here the membrane potentials are often less than zero. This is especially true for the background neuron. Overall the "biological" membrane potential fluctuates less violently than the "effective" membrane potential. This is to be expected as the "biological" case is modelling the effects of shunting and therefore capacitance of the cell membrane. Finally one should note that the threshold is always crossed at the same time for both representations as proven in section 5.3.5.

5.6 Discussion and Conclusion

The graphic displays illustrating the performance of the neural network model presented in the is chapter are encouraging. They are the first simulations from a neural network model, functioning as an associative memory, that may be compared directly with biological experiments. In this way the gulf between neural network theory and biological reality has been bridged. In future work the model could be explored further in order to understand its capabilities and limitations. Once this information has been gathered, the task of comparing with biological experiment would be on a surer footing. The origins of biological phenomena which are at present not well understood, such as the Gray–Singer oscillations could then be elucidated.

At present the only means of investigation of the model is by computer simulation. Although this yields visually impressive results, the understanding one achieves is superficial compared with insight given by the analyses of chapters 2–4. For example the graphic displays and the corresponding rate and Edward-Anderson order parameters, show the existence of a retrieval phase and non-ergodicity. Yet one has little understanding of the mechanisms by which the retrieval phase is destabilised, for example by decreasing temperature (Fig. 5.3). In order to develop analytic techniques, the model studied in this chapter may well have to be

simplified. In that case simulations of more detailed formulations of the model, as presented in this chapter, would take on the crucial rôle of bridge between abstract theory and experiment.

Chapters 2–4 of this thesis each present analytic studies which one would like to effect on the model of this chapter. Chapter 2 investigated certain associativity properties of the Hopfield model. The mechanism of associativity of the present model (see section 5.3.3) is novel, and its functioning has yet to be fully explored. The equivalent of a first-time step equation (2.2) would be a useful tool for this exploration. As discussed in the previous paragraph, the equivalent of the static mean-field theory presented in section 2.4, would allow the mechanism of retrieval to be understood. In chapter 3 the functioning of interactions optimal for one environment (the perceptron) was investigated in a different environment (attractor neural network with noise). One could consider the model of the present chapter as a new network environment. The functioning of interactions, different from or more complicated than the Willshaw interactions, in the new environment could be analysed in a straightforward manner if one had the equivalent of the dynamic equations of section 3.4.

The development of analyses to explain the results of this chapter would undoubtedly lead to a deeper understanding of the present model, and would perhaps generate other models and methods of general interest. However the question remains as to the model's worth. That can only be determined by the scientific community at large through cross-disciplinary appraisal and constructive criticism. I hope that the response will be positive so that the present work will repay something to the tradition of research from which it has borrowed.

Appendix
Expansions about the Tricritical Point

The function g will be parameterised by β and another parameter κ, which determines the loading level α and may even be α itself. For example, in Gardner's optimal network[41], κ is the site stability parameter. Or, it may be related to the number of errors allowed in the process of storage, etc. Ultimately we wish to consider a phase diagram in the κ–T or α–T plane.

The condition for a fixed point reads:

$$m = f^{(1)}(\beta,\kappa)m + f^{(3)}(\beta,\kappa)m^3 + f^{(5)}(\beta,\kappa)m^5. \tag{1}$$

The line of continuous transitions, determined by (3.56) together with $g'''(m=0) < 0$, gives β, or T, as a function of κ. As T is lowered below this line, at fixed κ, one has the usual mean field result for the developing retrieval amplitude m:

$$m^2 = \frac{f^{(1)}_\beta(c)}{f^{(3)}(c)}\triangle\beta, \tag{2}$$

where the subscript indicates partial differentiation; the variable c implies that the function is evaluated on the continuous line, i.e. at κ and $\beta_c(\kappa)$ and

$$\triangle\beta \equiv \beta - \beta_c(\kappa).$$

At the tricritical point the line of continuous transitions, $g'(0) = 0$, which becomes the line of transitions from wide to narrow retrieval, is expanded as

$$\triangle\beta_c(\kappa) = a\triangle\kappa + b(\triangle\kappa)^2 \tag{3}$$

and the line of discontinuous transitions to no retrieval as

$$\triangle\beta_{dc}(\kappa) = a_1\triangle\kappa + c(\triangle\kappa)^2 \tag{4}$$

where $\triangle\kappa = \kappa - \kappa_{tr}$. On substituting the first expansion into (3.56) one finds:

$$a = -\frac{f^{(1)}_\kappa(tr)}{f^{(1)}_\beta(tr)} \tag{5}$$

$$b = -\frac{a^2 f^{(1)}_{\beta\beta}(tr) + 2a f^{(1)}_{\beta\kappa}(tr) + f^{(1)}_{\kappa\kappa}(tr)}{2f^{(1)}_\beta(tr)} \tag{6}$$

and (tr) indicates evaluation at the tricritical point. The equation for the discontinuous transition, near the tricritical point, reads

$$1 = f^{(1)}(\beta,\kappa) + 3f^{(3)}(\beta,\kappa)m^2 + 5f^{(5)}(\beta,\kappa)m^4. \tag{7}$$

Combined with the condition for the fixed point, it gives the relation:

$$(f^{(3)})^2 = -4f^{(5)}(1 - f^{(1)}), \tag{8}$$

which is an equation relating β and κ along the discontinuous transition line. After expanding this expression about the tricritical point at which $f^{(3)} = 0$, $f^{(3)}_\beta = 0$, $f^{(1)} = 1$, and substituting Eqs. 3 and 4, one finds

$$\begin{aligned} 0 &= \triangle\kappa \left[4f^{(5)}(tr)(f^{(1)}_\kappa(tr) + a_1 f^{(1)}_\beta(tr))\right] \\ &+ (\triangle\kappa)^2 \left[[f^{(3)}_\kappa(tr)]^2 + 4c\left(f^{(5)}(tr)f^{(1)}_\beta(tr) + f^{(5)}(tr)_\beta[f^{(1)}_\kappa(tr) + a_1 f^{(1)}_\beta(tr)]\right)\right. \\ &- \left. -8bf^{(5)}(tr)f^{(1)}_\beta(tr) + 4f^{(5)}_\kappa(tr)\left(f^{(1)}_\kappa(tr) + a_1 f^{(1)}_\beta(tr)\right)\right] \end{aligned} \tag{9}$$

This equation gives

$$a_1 = a \tag{10}$$

$$c = b - \frac{[f^{(3)}_\kappa(tr)]^2}{4f^{(1)}_\beta(tr)f^{(5)}(tr)}. \tag{11}$$

With a and b given by Eqs. 5 and 6. Hence the line of transition from retrieval to no retrieval is continuous and has a continuous slope at the tricritical point. The discontinuity is in the curvature.

Bibliography

[1] Hebb D O 1949 *The Organisation of Behavior* New York: Wiley

[2] Hodgkin A L and Huxley A F 1952 *J. Physiol. (London)* **117** 500

[3] Kandel E R 1976 *Cellular Basis of Behavior* San Fransisco: Freeman

[4] McCulloch W S and Pitts W A 1943 *Bull. Math. Biophys.* **5** 115

[5] Grossberg S 1986 *The Adaptive Brain* Amsterdam: North Holland

[6] Hopfield J J 1984 *Proc. Natl. Acad. Sci. USA* **81** 3088

[7] Little W A 1974 *Math. Biosci.* **19** 101

[8] Peretto P and Niez J-J 1986 in *Disordered Systems and Biological Organisation* eds. Bienenstock E, Fogelman Soulie F and Weisbuch G, Berlin: Springer-Verlag

[9] Rosenblatt F 1961 *Principles of Neurodynamics* New York: Spartan Books

[10] Willshaw D J, Buneman O P and Longuet-Higgins H C 1969 *Nature* **222** 960

[11] Kohonen T, Reuhkala E, Makisara K and Vainio L 1976 *Biol. Cyber.* **22** 159

[12] Minsky M and Papert S 1969 *Perceptrons* Cambridge MA: MIT press

[13] Amari S 1972 *IEEE Trans. Comput.* **21** 1197

[14] Hopfield J J 1982 *Proc. Natl. Acad. Sci. USA* **79** 2554

[15] Amit D J, Gutfreund H and Sompolinsky H 1985 *Phys. Rev. A* **32** 1007

[16] ——1985 *Phys. Rev. Lett* **55** 1530

[17] ——1987 *Ann. Phys. ,NY* **173** 30

[18] Sompolinsky H 1987 *Heidelberg Colloquium on Glassy Dynamics* eds. J L Van Hemmen and I Morgestern (Heidelberg: Springer-Verlag)

[19] Mas J and Ramos E 1989 *J. Phys. A: Math. Gen.* **22** 3379

[20] Kree R and Zippelius A 1988 *J. Phys. A: Math. Gen.* **21** L813

[21] von der Malsburg C and Bienenstock E 1987 *Europhys. Lett.* **3** 1243

[22] Bienenstock E and von der Malsburg C 1987 *Europhys. Lett.* **4** 121

[23] Dotsenko V 1988 *J. Phys. A: Math. Gen.* **21** L783

[24] Coolen A C C and Kuijk F W 1989 *Neural Networks* **2** 495

[25] Gardner E, Derrida B and Mottishaw P 1987 *J. Phys. (Paris)* **48** 741

[26] Sherrington D and Kirkpatrick S 1975 *Phys. Rev. Lett.* **35** 1792

[27] Kirkpatrick S and Sherrington D 1978 *Phys. Rev. B* **17** 4384

[28] Crisanti A, Amit D J and Gutfreund H 1986 *Europhys. Lett.* **2** 337

[29] Parisi G 1980 *J. Phys. A: Math. Gen.* **13** L115

[30] ——1980 *J. Phys. A: Math. Gen.* **13** 1101

[31] ——1980 *J. Phys. A: Math. Gen.* **13** 1887

[32] ——1983 *Phys. Rev. Lett.* **50** 1946

[33] Mézard M, Parisi G, Virasoro M A 1987 *Spin Glass Theory and Beyond* Singapore: World Scientific

[34] Edwards S F and Anderson P W 1975 *J. Phys. F: Metal Phys.* **5** 965

[35] Forrest B M 1988 *J. Phys. A: Math. Gen.* **21** 245

[36] Gardner E 1987 *Europhys. Lett.* **4** 481

[37] ——1988 *J. Phys. A: Math. Gen.* **21** 257

[38] Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271

[39] Derrida B and Weisbuch G 1987 *J. Phys. (Paris)* **47** 1297

[40] Derrida B, Gardner E and Zippelius A 1987 *Europhys. Lett.* **4** 167

[41] Gardner E 1989 *J. Phys. A: Math. Gen.* **22** 1969

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

[43] ——1990 *Optimally adapted attractor neural networks in the presence of noise J. Phys. A: Math. Gen.* in press

[44] Abbot L F and Kepler T B 1989 *J. Phys. A: Math. Gen.* **22** 2031

[45] Wong K Y M and Sherrington D 1990 to be published in *Lecture Notes in Physics: Proceedings of the XI Sitges Conference on Neural Networks* Berlin: Springer Verlag

[46] Abeles M 1982 *Local Cortical Circuits* Berlin: Springer

[47] Abeles M, Vaadia E and Bergman H 1990 *Network* **1** 13

[48] Bruce A D, Gardner E, Wallace D J 1987 *J. Phys. A: Math. Gen.* **20** 2909

[49] Amit D J, Gutfreund H and Sompolinsky H 1987 *Phys. Rev.* **A 35** 2293

[50] Sejnowski T J 1977 *J. Math. Biol* **4** 303

[51] Tsodyks M V and Feigel'man M V 1988 *Europhys. Lett.* **6** 101

[52] Buhmann J, Divko R and Schulten K 1989 *Phys. Rev.* **A 39** 2689

[53] Wilson H R and Cowan J D 1972 *Biophys. J.* **12** 1

[54] Fontanari J F and Köberle R 1988 *J. Phys. A: Math. Gen.* **21** 2477

[55] Mézard M, Nadal J-P, Toulouse G 1986 *J. Physique (Paris)* **47** 1457

[56] Fontanari J F, Köberle R 1988 *J. Phys. A: Math. Gen.* **21** L253

[57] Virasoro M A 1989 *J. Phys.* **22A** 2227

[58] Shinomoto S 1987 *Biol. Cyber.* **57** 197

[59] Tsodyks M V 1988 *Europhys. Lett.* **7** 203

[60] Eccles J C 1964 *Physiology of Synapses* Berlin: Spinger

[61] Dobson V G 1987 *J. Intelligent Systems* **1** 43

[62] Golomb D, Rubin N, Sompolinsky H 1990 *Phys. Rev.* **A 41** 1843

[63] Mayashita Y and Chang H S 1988 *Nature* **331** 68

[64] Amit D J and Treves A 1989 *Proc. Natl Acad. Sci. USA* **86** 7671

[65] Treves A and Amit D J 1989 *J. Phys. A: Math. Gen.* **22** 2205

[66] Rubin N and Sompolinsky H 1989 *Europhys. Lett.* 10 465

[67] Buhmann J 1989 *Phys. Rev.* **A 40** 4145

[68] Treves A 1989 *The onset of order in associative nets of neurons* PhD thesis, Hebrew University

[69] Gray C M and Singer W 1989 *Proc. Natl Acad. Sci. USA* **86** 1698

[70] Gray C M, Konig P, Engel A K and Singer W 1989 *Nature* **338** 334

[71] Abeles M and Gerstein G L 1988 *J. Neurophysiol.* **60** 909